Shape-From-Focus Depth Reconstruction With a Spatial Consistency Model

Chen-Yu Tseng and Sheng-Jyh Wang, *Member, IEEE*

***Abstract*—This paper presents a maximum *a posteriori* (MAP) framework to incorporate a spatial consistency prior model for depth reconstruction in the shape-from-focus (SFF) process. Existing SFF techniques, which reconstruct a dense 3-D depth from multifocus image frames, usually have poor performance over low-contrast regions and usually need a large number of frames to achieve satisfactory results. To overcome these problems, a new depth reconstruction process is proposed to estimate the depth values by solving an MAP estimation problem with the inclusion of a spatial consistency model. This consistency model assumes that within a local region, the depth value of each pixel can be roughly predicted by an affine transformation of the image features at that pixel. A local learning process is proposed to construct the consistency model directly from the multifocus image sequence. By adopting this model, the depth values can be inferred in a more robust way, especially over low-contrast regions. In addition, to improve the computational efficiency, a cell-based version of the MAP framework is proposed. Experimental results demonstrate the effective improvement in accuracy and robustness as compared with existing approaches over real and synthesized image data. In addition, experimental results also demonstrate that the proposed method can achieve quite impressive performance, even with only the use of a few image frames.**

***Index Terms*—3-D reconstruction, depth estimation, depth map, shape-from-focus (SFF).**

I. Introduction

THE shape-from-focus (SFF) technique is a method to compute 3-D depth maps from image sequences acquired with varying focus settings. Since different focus settings correspond to different depths of field, we would expect that an object in the 3-D scene would be best focused by adopting one of the focus settings if there is a sufficient number of focus settings to cover the whole depth range of the 3-D scene. By searching for the best focus setting, we can roughly estimate the 3-D depth value of each object in the scene. Typically, the criterion to distinguish focused image regions from defocused regions is realized by a focus measure operator, whose output response is usually called focus measure value. To generate a 3-D depth image, the depth value of each pixel is inferred by searching for the maximal focus measure value over the acquired multifocus image data at that pixel.

To obtain the focus measure value, a variety of focus measure operators have been designed in the literature, such as the Laplacian-based operator in [1], the gradient-based operator in [2], the variation-based operator in [3], and the transform-domain-based operator in [9]. A common assumption of these operators is that a properly focused region usually contains sharper edges or stronger high-frequency components. Even though this assumption is basically true in most cases, the accuracy of focus measure values may get dramatically degraded by two factors. One factor is the small focus measure values over low-contrast or week-texture regions, while the other factor is the insufficient number of focus settings. Under these two situations, the performance of the SFF technique may get degraded and the inferred depth map would be noisy and spatially inconsistent with the image contents.

To deal with these two situations, two major approaches have been developed. One approach tries to improve the focus measures by including more information from neighboring regions, while the other approach suggests the use of a depth reconstruction process to reconstruct a more reasonable depth image from the originally noisy depth image. In the first approach, a common aspect is to expand the support of the local measurement to include more information from the neighborhood. However, expanding the local support may cause edge bleeding artifacts as the operator is applied across two surfaces of different depth values. To handle this edge bleeding problem, researchers have suggested several solutions. For example, Aydin and Akgul [4] present an adaptive focus measure operator with weighted support windows. The shape and weights of the support window are determined based on the local image characteristics of an additional all-in-focus image. Thelen *et al.* [5] suggest another adaptive method to select the size of neighborhood for the local operator based on a confidence criterion. In that method, the level of confidence is based on the difference of the focus measure values between the best focused image and the average image.

In the second approach, some researchers propose the use of a depth reconstruction process [6], [14], [15]. Mahmood and Choi [6] suggest the use of an iterative 3-D anisotropic nonlinear diffusion filter (ANDF) to enhance the estimated focus volume. Here, the focus volume refers to a stack of image planes consisting of the focus measure values of the multifocus image sequence. Gaganov and Ignateko [14] present a framework that uses a Markov random field (MRF) model.

Manuscript received June 2, 2013; revised March 25, 2014; accepted May 8, 2014. Date of publication September 18, 2014; date of current version December 3, 2014. This work was supported by the National Science Council of Taiwan under Grant 101-2221-E-009-142-MY2. This paper was recommended by Associate Editor P. Le Callet.

The authors are with the Department of Electronics Engineering, Institute of Electronics, National Chiao Tung University, Hsinchu 30010, Taiwan (e-mail: rocket367@gmail.com; shengjyh@faculty.nctu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2014.2358873

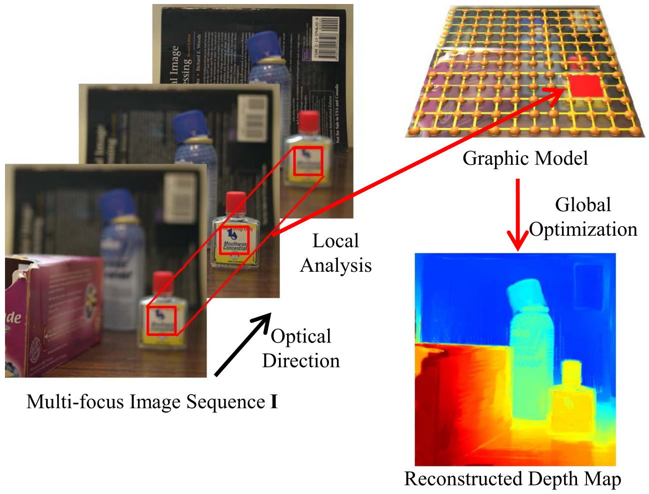


Fig. 1. Illustration of the proposed scheme.

Based on the MRF model, they propose an SFF method to yield a globally optimal solution based on some enforced smoothness priors. Ramnath and Rajagopalan [15] present a discontinuity-adaptive MRF framework with a nonconvex prior to capture sharp edges. For these methods, a major drawback is the required intensive computations in finding the optimal 3-D depth map.

Although the SFF technique has already been applied to many industrial applications, such as medical imaging systems, industrial inspection, 3-D object modeling, surveillance systems, and microelectronics [5], [7], [8], [10]–[13], it is still a challenge to deal with natural images in some real-time applications, like entertainment applications in consumer electronics. In such circumstances, there could be a lot of low-texture regions in the images and only a small number of image frames can be used to fit the real-time requirement. In this paper, we propose a global approach to deal with these two problems. The overview of the proposed scheme is illustrated in Fig. 1. Given a multifocus image sequence, a local analysis is first performed to explore both the focus measure values and the information about spatial consistency. The focus measures provide a cue for the depth inference along the optical axis of the camera. On the other hand, the spatial consistency constraint provides a useful key for depth inference by assuming that the depth values within the neighborhood of a pixel should be consistent with the image contents in the spatial domain. In the proposed framework, we build a likelihood model based on the spatially varying focus information and *a prior* model based on the spatial consistency learned from the image data. *A posteriori* model is deduced thereafter. By treating the depth reconstruction process as a maximum *a posteriori* (MAP) estimation problem, we derive a closed-form solution for the SFF problem.

The proposed framework is based on the spatial coherence recovery approach proposed in [16]. In [16], we had presented a MAP framework to recover the depth image using a matting Laplacian prior. In that framework, the matting Laplacian prior is constructed based on an additional all-in-focus image besides the multifocus image sequence. However, the need of an all-in-focus image is a barrier in practical applications. Hence, in this paper, we further propose a local learning scheme to derive the prior model directly from the multifocus image sequence, without the need of the all-in-focus image. Moreover, since this prior model is learned directly from the multifocus image sequence, the newly proposed scheme may also properly avoid the blurring of sharp edges that usually occurs in existing approaches.

The outline of this paper is organized as follows. In Section II, we present the proposed framework for depth reconstruction. In Section III, we introduce a cell-based framework to further reduce the required computations. In Sections IV and V, the experimental results and conclusion are given.

II. Proposed Depth Reconstruction

A. Overview of Proposed Scheme

Given an multifocus image set $\mathbf{I}^{\mathbf{set}} = \{\mathbf{I}^{\mathbf{1}}, \mathbf{I}^{\mathbf{2}}, \ldots, \mathbf{I}^{\mathbf{K}}\}$, where K is the number of frames and $\mathbf{I}^{\mathbf{j}}$ is the jth image frame, we aim to estimate the depth value at each image pixel. In this paper, we denote the depth image as $\mathbf{D}$ and treat the depth reconstruction as an MAP estimation problem, in which we search for the optimal depth image $\mathbf{D}^*$ that is

$$\mathbf{D}^* = \arg\max_{\mathbf{D}}\{p(\mathbf{D}|\mathbf{I}^{\mathbf{set}})\}. \tag{1}$$

Based on Bayes' formula, the posteriori probability function can be expressed as the product of the likelihood function $p(\mathbf{I}^{\mathbf{set}}|\mathbf{D})$ and the prior probability function $p(\mathbf{D})$ that is

$$p(\mathbf{D}|\mathbf{I}^{\mathbf{set}}) \propto p(\mathbf{I}^{\mathbf{set}}|\mathbf{D})p(\mathbf{D}). \tag{2}$$

In the following paragraphs, we will introduce the construction of the likelihood and the prior models in term of local analysis. The likelihood model is designed based on local depth prediction with spatial-varying precision, which can properly suppress inaccurate depth estimations over low-contrast regions. On the other hand, the prior model is designed based on the spatial consistency property among pixels. This spatial consistency property enables the propagation of high-confidence depth information to revise unreliable depth values. With the combination of the likelihood and prior models, we formulate an optimization problem to derive more reliable 3-D depth maps.

B. Local Analysis

For the sake of model simplification, we assume the global posteriori probability function $p(\mathbf{D}|\mathbf{I}^{\mathbf{set}})$ in (1) can be decomposed into a product of local posteriors. This decomposition is based on the assumption that a typical depth map can be approximated by a set of piece-wise smooth functions, with each function being an affine transformation of the image features within the corresponding local window. With this assumption, we independently solve the optimal parameters of the affine transformation for each window. On the other hand, we use overlapped windows to maintain spatial consistency.

To define the local posterior, we first denote I_i^k as the image data of the pixel i on the kth image frame and denote

d_i as the value of the depth map at pixel i. On the other hand, we define $\mathbf{I}_\mathbf{i}^{\mathbf{Set}} = [I_i^1\ I_i^2\ \ldots\ I_i^K]^T$ to represent the observed intensity values at pixel i in the multifocus image sequence. Moreover, we denote W_q as an $r \times r$ local window centered at pixel q and denote $N_q \equiv \{\tau_1, \tau_2, \ldots, \tau_{r^2}\}$ as the set of pixels within W_i. Based on the above notations, we define $\mathbf{d}_q \equiv [d_{\tau_1}, d_{\tau_2}, \ldots, d_{\tau_{r^2}}]^T$ as the vector made of the depth values of the pixels within W_q. On the other hand, the observed multifocus Red, Blue, and Green (RGB) data within the local window W_q around pixel q are represented as $\mathbf{I}_q = [(\mathbf{I}_{\tau_1}^{\mathbf{Set}})^T (\mathbf{I}_{\tau_2}^{\mathbf{Set}})^T \ldots (\mathbf{I}_{\tau_{r^2}}^{\mathbf{Set}})^T]^T$, which is formed by cascading the $\mathbf{I}_\mathbf{i}^{\mathbf{Set}}$ vectors within W_q. With the above notations, the global posterior probability function is decomposed into a product of local posteriors as

$$p(\mathbf{D}|\mathbf{I}^{\mathbf{set}}) \propto \left(\prod_{q\in\Omega} p(\mathbf{d}_q|\mathbf{I}_q)\right)^{\frac{1}{r^2}} \tag{3}$$

where Ω denotes the whole set of q's. In (3), the inclusion of the power term $1/r^2$ is due to the fact that the multifocus data $\mathbf{I}_\mathbf{i}^{\mathbf{Set}}$ at each pixel will be considered r^2 times as we scan the $r \times r$ local window pixel by pixel through the whole image domain. This power term can actually be ignored since it does not affect the MAP solution at all. Hence, based on the decomposition in (3) and the Bayes' formula, we further rewrite the original MAP formulation in (1) as

$$\begin{aligned}\mathbf{D}^* &= \arg\max_{\mathbf{D}}\{p(\mathbf{D}|\mathbf{I}^{\mathbf{set}})\}\\ &= \arg\max_{\mathbf{D}}\left\{\prod_{q\in\Omega} p(\mathbf{d}_q|\mathbf{I}_q)\right\}\\ &= \arg\max_{\mathbf{D}}\left\{\prod_{q\in\Omega} p(\mathbf{I}_q\,|\mathbf{d}_q)\,p(\mathbf{d}_q)\right\}.\end{aligned} \tag{4}$$

In the following paragraphs, we will explain in detail how we design the local likelihood model $p(\mathbf{I}_q|\mathbf{d}_q)$ and the local prior model $p(\mathbf{d}_q)$.

1) Local Likelihood Model: Since it is an ill-posed problem to directly model the relation between $\mathbf{I}_q$ and $\mathbf{d}_q$, we cannot explicitly formulate the likelihood model $p(\mathbf{I}_q\,|\mathbf{d}_q)$. Instead, we introduce the estimated depth $\tilde{\mathbf{d}}_q \equiv \{\tilde{d}_{\tau_1}, \tilde{d}_{\tau_2}, \ldots, \tilde{d}_{\tau_{r^2}}\}$ and use it as a bridge to relate the observation $\mathbf{I}_q$ and the hidden model $\mathbf{d}_q$. To relate $\mathbf{I}_q$ with $\tilde{\mathbf{d}}_q$, we employ a difference of Gaussians (DoG) operator over each image frame of the image sequences to estimate the depth value $\tilde{d}_i$ at pixel i. Here, for the pixel i at (x,y) on the kth image frame, we define the focus measure value as

$$F^k(x,y) = \left|(G_{\sigma_1}(x,y) - G_{\sigma_2}(x,y)) * I^k(x,y)\right| \tag{5}$$

where $G_{\sigma_1}(x,y)$ and $G_{\sigma_2}(x,y)$ are two zero-mean Gaussian kernels with the standard deviation $\sigma_1 = 0.5$ and $\sigma_2 = 0.8$, respectively. In general, this DoG operator generates stronger responses for sharper edges. Hence, at each image pixel, by finding the image frame on which the DoG operator outputs the strongest response, we can estimate the depth value of that pixel accordingly. In other words, the depth value at a pixel, say (x_0, y_0), is estimated to be

$$\tilde{d}(x_0, y_0) = \arg\max_k(F^k(x_0, y_0)). \tag{6}$$

With (5) and (6), we can estimate the local depth $\tilde{\mathbf{d}}_q$ at each image pixel q. Here, we simply denote the image frame index as the depth value. In practice, we need to roughly measure the 3-D depth value for each focus setting and then convert the image frame index to the 3-D depth value accordingly. Moreover, note that only K discrete depth values are provided at this stage, rather than continuous-valued depth values. Later, with the introduction of the proposed MAP framework, we will be able to generate a continuous-valued 3-D depth map.

On the other hand, to relate $\tilde{\mathbf{d}}_q$ with $\mathbf{d}_q$, we assume $\tilde{d}(x,y)$ is a random variable centered at the true depth $d(x,y)$ with a certain level of variations. With the bridging of the depth feature $\tilde{\mathbf{d}}_q$, we formulate the local likelihood model $p(\mathbf{I}_q\,|\mathbf{d}_q)$ as

$$p(\mathbf{I}_q\,|\mathbf{d}_q) \equiv p(\tilde{\mathbf{d}}_q\,|\mathbf{d}_q). \tag{7}$$

In our approach, we treat the predicted depth data $\tilde{\mathbf{d}}_q$ as being governed by the hidden depth data $\mathbf{d}_q$ and adopt an independent and identically distributed Gaussian model with a spatially varying precision matrix $\mathbf{\Lambda}_q$ to model $p(\tilde{\mathbf{d}}_q|\mathbf{d}_q)$

$$p(\tilde{\mathbf{d}}_q|\mathbf{d}_q) \equiv N\left(\tilde{\mathbf{d}}_q\middle|\mathbf{d}_q, \mathbf{\Lambda}_q^{-1}\right). \tag{8}$$

By taking the negative of the logarithm of $p(\tilde{\mathbf{d}}_q|\mathbf{d}_q)$, we have

$$\begin{aligned}-\log p(\tilde{\mathbf{d}}_q|\mathbf{d}_q) &\equiv (\tilde{\mathbf{d}}_q - \mathbf{d}_q)^T \mathbf{\Lambda}_q (\tilde{\mathbf{d}}_q - \mathbf{d}_q)\\ &= \sum_{i\in N_q} \lambda_i (\tilde{d}_i - d_i)^2.\end{aligned} \tag{9}$$

Here, $\mathbf{\Lambda}_q$ is an $M \times M$ diagonal matrix, in which the diagonal terms are made of the λ_i values within W_q and $M = r^2$ is the number of pixels within W_q. The definition of the precision term λ_i will be mentioned later. Basically, λ_i models the certainty about the estimation of depth value at pixel i. For a low-contrast case, we expect that the uncertainty would increase and the precision value drops. That is, the depth value $\tilde{d}_i$ could be more deviated from the hidden depth value d_i.

In our design, the definition of the precision term λ_i is based on local entropy, a measure of the uncertainty in the determination of the best focused frame for the pixel i. The entropy would increase as the contrast decreases. To measure the local entropy, we first denote $p_i^k = p(C_i = k|\mathbf{I}_\mathbf{i}^{\mathbf{Set}})$ as the probability that the kth frame is the best focused frame for pixel i. Here, C_i denotes the frame index of the best focused frame. By expecting that a larger focus measure value at an image pixel usually means that the image pixel is better focused, we assume the probability p_i^k of the pixel i at (x,y) is proportional to the focus measure value F^k at that pixel. That is, we define p_i^k as

$$p_i^k = \frac{F^k(x,y)}{\sum\limits_{j=1}^{K} F^j(x,y)} \tag{10}$$

where K is the number of image frames in the multifocus image sequence. Based on (10), the local entropy at pixel i is defined as

$$h_i = \sum_{j=1}^{K} \left(-p_i^j \log\left(p_i^j\right)\right). \tag{11}$$

With the definition in (11), a higher entropy value corresponds to a higher uncertainty in determining the best focused frame. This corresponds to a lower precision value in local depth inference. Hence, in our approach, we define the precision term at pixel i to be

$$\lambda_i = \begin{cases} 1 - \bar{h}_i & \text{for } \bar{h}_i < t_0 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where $\bar{h}_i = h_i / h_{\max}$ is the normalized entropy and $h_{\max}$ denotes the maximal entropy over all pixels. Here, t_0 is a preselected clipping threshold and we empirically set $t_0 = 0.95$ in our experiments. Once the entropy value exceeds t_0, we expect that the contrast is too low and the observed data are highly unreliable. In that case, the precision value is simply set to zero.

2) Local Prior Model: In our approach, the local prior model $p(\mathbf{d}_q)$ is based on the spatial consistency assumption that the depth values of adjacent pixels would be roughly the same if the image features, like intensity or colors, at these pixels are similar. Moreover, the depth values at adjacent pixels may change rapidly only when the image features at these pixels have apparent changes. As will be demonstrated later, this spatial consistency assumption provides a useful key for depth inference over low-contrast regions. In one aspect, most low-contrast regions contain smoothly changing image features and we expect that the depth values within these regions would be highly correlated. With the use of the spatial consistency assumption, we will be able to maintain the high correlation of depth values over these smoothly changing image regions. In another aspect, the employed spatial consistency assumption may also help in identifying regions of dramatic depth changes. This can help us to effectively suppress the previously mentioned edge bleeding artifacts.

In [16], we have presented a spatial coherence recovery framework with the use of matting Laplacian matrix. The matting Laplacian matrix is originally proposed in [17] to solve the supervised matting problem. The supervised matting is a process to extract foreground objects, along with the opacity of the foreground object, from an image with user's guidance. The foreground opacity is typically called alpha matte. In [17], by deriving the optimal matting values based on the matting Laplacian matrix, the authors obtain image matting results with very impressive quality. Inspired by their work, we adopt the matting Laplacian matrix as *a prior* model to provide the spatial coherence constraint for the SFF process and we have obtained greatly improved performance in 3-D depth estimation [16]. However, in that previous work, an additional all-in-focus image is required to generate the matting Laplacian matrix. The requirement of the all-in-focus image causes a big barrier in practical applications. Hence, in this paper, we will propose a new framework to learn the prior model directly from the multifocus image sequence, without the involvement of any all-in-focus image.

To construct the prior model, we present a local learning scheme by assuming that over a local neighborhood, the depth value of each pixel can be predicted by an affine transformation of its image features. The local prediction model is based the assumption that the distribution of depth values within a local region can be approximated by a regression model based on image features. Several existing learning-based depth estimation methods are based on similar assumptions. For example, Saxena *et al.* [20] present a supervised learning approach to estimate depth from local features based on a linear model. Saxena *et al.* [21] propose to decompose an image into a number of planar surfaces and then infer the orientation of each surface to reconstruct the 3-D models. The coefficients of the affine transformation are locally constant, but can be globally varying. With this assumption, if the image features within a local region are similar under constant illumination, the depth values will also be similar. On the other hand, if the image features within a local region are changing, the depth values may also (but not necessarily) be changing.

In our algorithm, we choose the image feature at a pixel as the R, G, and B values at that pixel. Here, we use the notation $\mathbf{v}_i^k = [r_i^k, g_i^k, b_i^k]^T$ to represent the feature vector at pixel i in the kth image frame, with r_i^k, g_i^k, and b_i^k being the R, G, and B values at that pixel. Based on the affine transformation assumption, the depth value d_i^k at pixel i on frame k can be expressed as

$$d_i^k = \left[\mathbf{v}_i^k\right]^T \boldsymbol{\beta} + \beta_0 \tag{13}$$

where $\boldsymbol{\beta} = [\beta_r, \beta_g, \beta_b]^T$ and β_0 is a constant. As mentioned above, $\beta_r, \beta_g, \beta_b$, and β_0 are locally constant, but may be different for different image regions.

Moreover, to combine the values of d_i^k at different image frames, we adopt

$$d_i = \sum_{k=1}^{K} p_i^k \cdot d_i^k. \tag{14}$$

In (14), $p_i^k = p(C_i = k | \mathbf{I}_\mathbf{i}^{\mathbf{Set}})$ has been defined in (10) to represent the probability that the kth frame is the best focused frame for the image pixel i. By combining (13) with (14), we have the representation

$$d_i = \mathbf{p}_i \begin{bmatrix} \mathbf{V}_i^T & \mathbf{1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \beta_0 \end{bmatrix} \tag{15}$$

where $\mathbf{p}_i = [p_i^1, \ldots, p_i^K]$ is a $1 \times K$ matrix, $\mathbf{V}_i = [\mathbf{v}_i^1, \ldots, \mathbf{v}_i^K]$ is an $3 \times K$ matrix, and $\mathbf{1}$ is a $K \times 1$ vector with all elements being 1. By defining $\mathbf{f}_i = \mathbf{p}_i \mathbf{V}_i^T$, (15) can be further rewritten as

$$d_i = \begin{bmatrix} \mathbf{f}_i & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \beta_0. \end{bmatrix}. \tag{16}$$

In (16), we represent the depth value at a single pixel as an affine transformation of image features at that pixel. Since we have assumed that the affine transformation coefficients $\{\boldsymbol{\beta}, \beta_0\}$ are locally constant, we can further derive an affine model for the depth values within a local neighborhood.

Same as before, we define W_q as an $r \times r$ window, $\mathbf{d}_q \equiv [d_{\tau_1}, d_{\tau_2}, \ldots, d_{\tau_M}]^T$ as the vector of depth values of all pixels within W_q, and $M = r^2$ as the number of pixels in W_q. In addition, we define $\mathbf{F}_q = [\tilde{\mathbf{f}}_{\tau_1}^T, \ldots, \tilde{\mathbf{f}}_{\tau_j}^T, \ldots, \tilde{\mathbf{f}}_{\tau_M}^T]^T$ to denote an $M \times 4$ matrix stacked by the corresponding feature vectors $\tilde{\mathbf{f}}_{\tau_j} = [\mathbf{f}_{\tau_j}\ \mathbf{1}]$. Based on the above notations, the depth prediction for all pixels within W_q can be expressed as

$$\mathbf{d}_q = \mathbf{F}_q \begin{bmatrix} \boldsymbol{\beta} \\ \beta_0. \end{bmatrix}. \tag{17}$$

Equation (17) relates the depth values within W_q with the corresponding image features within W_q. When crossing different surfaces, the entries of the depth vector $\mathbf{d}_q$ can be rapidly changing with respect to the feature vectors.

If both $\mathbf{d}_q$ and $\mathbf{F}_q$ are given, then the optimal $\boldsymbol{\beta}$ and β_0 can be derived by minimizing the quadratic cost function as

$$E(\boldsymbol{\beta}, \beta_0) = \left\| \mathbf{d}_q - \mathbf{F}_q \begin{bmatrix} \boldsymbol{\beta} \\ \beta_0 \end{bmatrix} \right\|^2 + c_\beta \boldsymbol{\beta}^T \boldsymbol{\beta} \tag{18}$$

where c_β is a parameter for regularization. For the cost function in (18), the optimal solution of $\boldsymbol{\beta}$ and β_0 can be easily derived to be

$$\begin{bmatrix} \boldsymbol{\beta}^* \\ \beta_0^* \end{bmatrix} = \left(\mathbf{F}_q^T \mathbf{F}_q + c_\beta \mathbf{D}_\beta\right)^{-1} \mathbf{F}_q^T \mathbf{d}_q. \tag{19}$$

In (19), we denote $\boldsymbol{\beta}^*$ and β_0^* to be the optimal coefficients for W_q and define $\mathbf{D}_\beta = \begin{bmatrix} \mathbf{I}_3 & 0 \\ 0 & 0 \end{bmatrix}$ as a 4×4 matrix, where $\mathbf{I}_3$ is the 3×3 identity matrix. By substituting (19) back to (17), we can express the optimal depth value $\mathbf{d}_q^*$ as

$$\mathbf{d}_q^* = \mathbf{Z}_q^T \mathbf{d}_q \tag{20}$$

where $\mathbf{Z}_q = \mathbf{F}_q(\mathbf{F}_q^T \mathbf{F}_q + c_\beta \mathbf{D}_\beta)^{-1} \mathbf{F}_q^T$.

In (20), $\mathbf{Z}_q$ is an $M \times M$ transformation matrix. In this equation, each entry in the left-hand side $\mathbf{d}_q^*$ is expressed as a linear combination of the entries in the right-hand side $\mathbf{d}_q$. This means that, with the spatial consistency assumption, the depth value of each pixel in W_q can actually be expressed as a linear combination of the depth values themselves within W_q. With this property, we will be able to eliminate outliers in $\mathbf{d}_q$ over low-contrast regions.

Based on the above deduction, we design the local prior model based on the following square error function with respect to $\mathbf{d}_q$:

$$\begin{aligned} -\log(p(\mathbf{d}_q)) &= \left\| \mathbf{d}_q - \mathbf{d}_q^* \right\|^2 \\ &= \left\| \mathbf{d}_q - \mathbf{Z}_q^T \mathbf{d}_q \right\|^2 \\ &= \mathbf{d}_q^T (\mathbf{I}_M - \mathbf{Z}_q)^T (\mathbf{I}_M - \mathbf{Z}_q) \mathbf{d}_q \\ &= \mathbf{d}_q^T \mathbf{L}_q \mathbf{d}_q \end{aligned} \tag{21}$$

where $\mathbf{I}_M$ is the $M \times M$ identity matrix and $\mathbf{L}_q = (\mathbf{I}_M - \mathbf{Z}_q)^T (\mathbf{I}_M - \mathbf{Z}_q)$ is the graph Laplacian matrix. In [16], we need an additional all-in-focus image to calculate the Laplacian matrix. Now, based on the local learning scheme, we derive the Laplacian matrix directly from the multifocus image sequence.

To interpret the graph Laplacian matrix, we may refer to the spectral graph theory in [19] and [20]. Assume we define a graph $\boldsymbol{\Gamma}_q$ in which the vertices represent the image pixels in W_q and the edge between a pair of vertices represents the affinity between the corresponding image pixels. For $\boldsymbol{\Gamma}_q$, its corresponding graph Laplacian matrix is defined as

$$\mathbf{L}_q = \mathbf{D}_q - \mathbf{A}_q \tag{22}$$

where $\mathbf{D}_q$ is the degree matrix and $\mathbf{A}_q$ is the affinity matrix. The entry $\mathbf{A}_q(ij)$ represents the affinity value between pixels i and j, while the degree matrix $\mathbf{D}_q$ is a diagonal matrix with its diagonal term being defined as

$$\mathbf{D}_q(i, i) = \sum_{j=1}^{N} \mathbf{A}_q(i, j). \tag{23}$$

In our approach, we do not explicitly define the affinity matrix. Instead, the affinity matrix is implicitly embedded in the graph Laplacian, which is the result of the optimization process expressed in (21). Furthermore, the local prior model in (21) can also be interpreted as

$$\begin{aligned} -\log(p(\mathbf{d}_q)) &= \mathbf{d}_q{}^T \mathbf{L}_q \mathbf{d}_q \\ &= \sum_{i \in N_q} \sum_{j \in N_q} \frac{1}{2} \mathbf{A}_q(i, j) \| d_i - d_j \|^2. \end{aligned} \tag{24}$$

Strictly speaking, the model in (24) is not a typical prior model since it actually depends on the image features of the given image data. However, we can treat it as a generalized prior model and use it to obtain spatially consistent depth maps. This generalized prior model prefers smoothly changing depth values for pixel pairs with larger affinity values and may allow depth values to fluctuate more for pixel pairs with smaller affinity values.

C. Global Optimization

With the local prior model in (21) and the local likelihood model in (9), we have the local MAP model as

$$\begin{aligned} &-\log(p(\mathbf{I}_q | \mathbf{d}_q) p(\mathbf{d}_q)) \\ &\qquad = (\tilde{\mathbf{d}}_q - \mathbf{d}_q)^T \boldsymbol{\Lambda}_q (\tilde{\mathbf{d}}_q - \mathbf{d}_q) + \mathbf{d}_q^T \mathbf{L}_q \mathbf{d}_q. \end{aligned} \tag{25}$$

As mentioned before, by assuming that the local observations are mutually independent, the global posterior probability can be represented as a product of local posterior probabilities. That is

$$\begin{aligned} &-\log(p(\mathbf{I}^{\mathbf{Set}} | \mathbf{D}) p(\mathbf{D})) \\ &\qquad = \sum_{q \in \Omega} -\log(p(\mathbf{I}_q | \mathbf{d}_q) p(\mathbf{d}_q)) \\ &\qquad = \sum_{q \in \Omega} \left\{ (\tilde{\mathbf{d}}_q - \mathbf{d}_q)^T \boldsymbol{\Lambda}_q (\tilde{\mathbf{d}}_q - \mathbf{d}_q) + \mathbf{d}_q^T \mathbf{L}_q \mathbf{d}_q \right\} \end{aligned} \tag{26}$$

where $\boldsymbol{\Omega}$ denotes the whole set of q's. We can further deduce a matrix format of (26) as

$$-\log(p(\mathbf{I}^{\mathbf{Set}} | \mathbf{D}) p(\mathbf{D})) = (\tilde{\mathbf{d}} - \mathbf{d})^T \boldsymbol{\Lambda} (\tilde{\mathbf{d}} - \mathbf{d}) + \mathbf{d}^T \mathbf{L} \mathbf{d}. \tag{27}$$

In (27), we define $\tilde{\mathbf{d}} = [\tilde{d}_1, \ldots, \tilde{d}_N]^T$ as an $N \times 1$ vector that denotes the predicted depth values of all image pixels.

Here, N denotes the total number of pixels in an image frame. Moreover, we define $\mathbf{d} = [d_1, \ldots, d_N]^T$ as an $N \times 1$ vector that denotes the corresponding target depth values. In addition, $\mathbf{\Lambda}$ is an $N \times N$ diagonal matrix, whose diagonal term $\mathbf{\Lambda}(i, i)$ equals to λ_i, the precision value at the pixel i. $\mathbf{L}$ is an $N \times N$ graph Laplacian matrix defined as

$$\mathbf{L} = \sum_{q \in \Omega} \mathbf{L'}_q \tag{28}$$

where $\mathbf{L'}_q$ is an $N \times N$ matrix expanded from the $M \times M$ matrix $\boldsymbol{L}_q$ in (26). Here, M denotes the total number of pixels within the local window W_q around the pixel q. For those pixels in W_q, the related entries in $\mathbf{L'}_q$ are equal to the corresponding entries in $\mathbf{L}_q$; while for those pixels outside W_q, the corresponding entries in $\mathbf{L'}_q$ are simply set to zero.

Finally, the global minimum of (27) can be obtained by solving system of linear equations as

$$(\mathbf{L} + \mathbf{\Lambda})\mathbf{d} = \mathbf{\Lambda}\tilde{\mathbf{d}}. \tag{29}$$

In summary, with the inclusion of the prior model in the proposed MAP estimation framework, we can reconstruct a spatially consistent depth image based on the spatial affinity information embedded in the image intensity data. This will improve the performance of depth estimation over low-contrast regions and also suppress edge-bleeding artifacts over boundary regions.

III. Efficient Cell-Based Framework

A. Overview

In Section II, we have presented an MAP approach for generating a 3-D depth map from a multifocus image sequence. The final global optimal solution can be obtained by solving the linear equations described in (29). However, it will be very time consuming to deal with large-scale images that require solving a huge system of linear equations. Hence, in this paper, we further propose a cell-based framework to improve the computational efficiency.

The proposed scheme is motivated by the observation that depth values at adjacent pixels are usually highly correlated. If we can properly utilize this property, we would be able to eliminate a considerable amount of redundant computations without sacrificing the quality of the output results.

The idea of the proposed cell-based framework is shown in Fig. 2. Unlike the pixel-based approach that obtains the optimal 3-D depth image (green points) for all pixels based on the observed image data (red points), the cell-based approach proposes the use of an intermediate grid cells [orange points in Fig. 2(b)] by grouping pixels into cells. In the cell-based scheme, we estimate the depth value of each cell first and then estimate the cell-wise 3-D depth map based on the depth values at cells. Since the number of cells could be much less than the number of pixels, we can greatly reduce the computational load by performing cell-wise MAP estimation. After that, the pixel-wise 3-D depth map can be reconstructed based on the cell-wise depth estimation results.

In [16], we have proposed a cell-based framework to reduce the computations by coarsening the matting Laplacian matrix. To achieve that, we express the image pixels as vectors scattering in a 5-D space, with each vector containing the spatial coordinates and the RGB values at a pixel. The reason that performing down sampling in that space rather than in the spatial domain is because the RGB values can be very helpful in avoiding blending conflicting depth values from different surfaces. In that 5-D space, we apply a grid data structure for down sampling and then estimate the depth data for grid cells. After that, we reconstruct the final 3-D depth map based on a nonlinear interpolation over the grid data. In this paper, we adopt a similar approach. However, a major difference is that we do not directly apply a coarsening process to reduce the scale of the linear equation system in (29). Instead, we revisit the construction of the local posterior model in the previous section and derive a more efficient scheme.

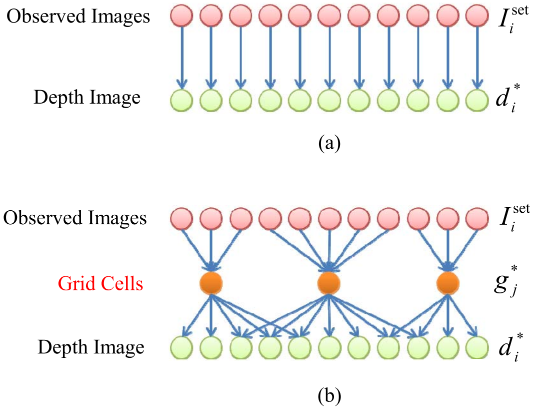


Fig. 2. Illustration of depth inference with (a) pixel- and (b) cell-based scheme.

B. Cell-Based MAP Estimation

To reduce the computations for depth inference, we modify the pixel-based MAP estimation into a cell-based one, where a pixel-to-cell mapping function is derived using a grid data structure in a high-dimensional space. Given a pixel i, we have its spatial coordinates $\mathbf{s}_i$ and its multifocus feature vector $\mathbf{f}_i = \mathbf{p}_i \mathbf{V}_i{}^T$, which is used to predict the depth value in the prior model in (16). For the pixel i, we define its index vector $\mathbf{h}_i = [\mathbf{s}_i \;\; \mathbf{f}_i]^T$ in the high-dimensional space, as shown in Fig. 3(a). We then apply a grid structure in that space for the grouping of the index vectors. This grid structure is constructed by uniformly down sampling the spatial coordinates into b_s bins and uniformly down sampling the multifocus feature coordinates into b_f bins. After scanning through the entire image, we record the pixel-to-cell mapping in terms of an $N \times R$ binary matrix $\mathbf{m}$, where N is the number of image pixels and R is the number of grid cells. If the pixel i is classified into the cell j, we define $\mathbf{m}(i, j) = 1$ and $\mathbf{m}(i, k) = 0$ for all $k \neq j$.

Assume we denote g_j as the 3-D depth value of the jth cell in the grid structure and denote $\boldsymbol{g}$ as the collection of g_j's. Based on the grid structure, we treat the cell-based depth reconstruction process as an MAP estimation problem, in which we derive the optimal cell-wise depth vector $\boldsymbol{g}^*$ as

$$\mathbf{g}^* = \arg\max_{\mathbf{g}} \left\{ p(\mathbf{g} \,|\mathbf{I}^{\mathbf{Set}}) \right\}. \tag{30}$$

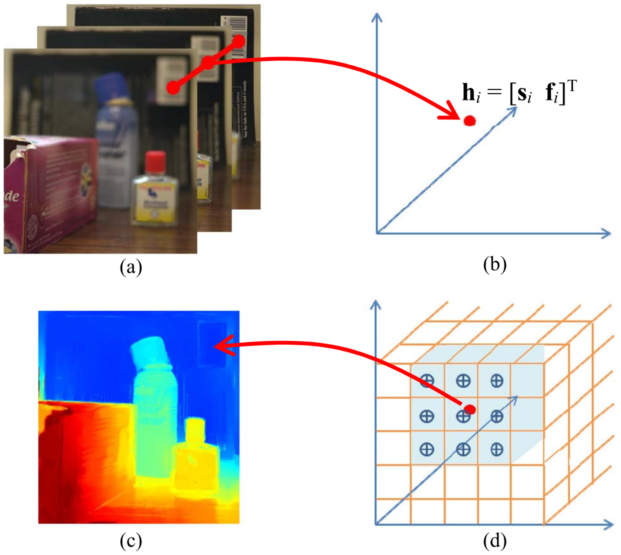


Fig. 3. Illustration of cell-based depth inference with a high-dimensional grid. (a) Multifocus image sequence I. (b) High-dimensional space. (c) Reconstructed depth image. (d) Grid structure in high-dimensional space.

To formulate the MAP estimation, we rewrite the posteriori probability in (30) as

$$p(\mathbf{g}\,|\mathbf{I}^{\mathbf{Set}}) \propto p(\mathbf{I}^{\mathbf{Set}}\,|\mathbf{g})\,p(\mathbf{g}). \tag{31}$$

In the following paragraphs, we proceed to present the formulation of the cell-based likelihood model $p(\mathbf{I}^{\mathbf{Set}}|\mathbf{g})$ and the prior model $p(\boldsymbol{g})$ based on the derivations in (26) and (27).

1) Cell-Based Likelihood Model: To model the cell-based likelihood function $p(\mathbf{I}^{\mathbf{Set}}|\mathbf{g})$, we first find an $R \times 1$ predicted depth vector $\tilde{\mathbf{g}}$, where R denotes the total number of cells in the grid structure. Similar to (9), we assume $\tilde{\mathbf{g}}$ is governed by the hidden depth $\mathbf{g}$ with a cell-based precision matrix $\boldsymbol{\Lambda}_g$. That is

$$-\log(p(\mathbf{I}^{\mathbf{Set}}|\mathbf{g})) = (\tilde{\mathbf{g}} - \mathbf{g})^T \boldsymbol{\Lambda}_g (\tilde{\mathbf{g}} - \mathbf{g}). \tag{32}$$

For the cell j, its predicted depth value $\tilde{g}_j$ is computed as the mean of the predicted depth values of all the pixels mapped into the cell j. That is

$$\tilde{g}_j = \frac{1}{w_j} \sum_{i=1}^{N} \mathbf{m}(i, j)\tilde{d}_i \tag{33}$$

where $w_j = \sum_{i=1}^{N} \mathbf{m}(i, j)$ and $\mathbf{m}(i, j)$'s are the entries of the previously mentioned pixel-to-cell mapping matrix $\mathbf{m}$.

Similarly, the cell-based precision matrix $\boldsymbol{\Lambda}_g$ is computed by

$$\boldsymbol{\Lambda}_g(j, j) = \frac{1}{w_j} \sum_{i=1}^{N} \mathbf{m}(i, j)\boldsymbol{\Lambda}(i, i). \tag{34}$$

2) Cell-Based Prior Model: In Section II, the establishment of the pixel-based prior model is based on the local learning of multifocus feature vectors. For the cell-based prior model, we adopt a similar approach. Here, we first define the expected feature vector $\boldsymbol{\varphi}_j$ for each cell. For the cell j, its feature vector $\boldsymbol{\varphi}_j$ is derived by the accumulation of the pixel-wise feature vectors $\mathbf{f}_i = \mathbf{p}_i \mathbf{V}_i^T$ based on the pixel-to-cell mapping function $\mathbf{m}$. That is

$$\boldsymbol{\varphi}_j = \frac{1}{w_j} \sum_{i=1}^{N} \mathbf{m}(i, j)\mathbf{f}_i. \tag{35}$$

Here, we assume the depth value g_j of the cell j can be predicted by an affine transformation of the cell-wise feature vector $\boldsymbol{\varphi}_j$. That is

$$g_j = [\boldsymbol{\varphi}_j \quad 1] \begin{bmatrix} \boldsymbol{\beta} \\ \beta_0. \end{bmatrix}. \tag{36}$$

Similar to the derivation of the pixel-based prior model, the cell-based prior model is derived from an integration of local models. To compute the local models, we place the same $r \times r$ local window in the pixel domain. Within each window, we inspect the pixels and their corresponding cells. Here, we denote $\boldsymbol{\Omega}_\rho$ as the set of referred cells for the window around ρ and denote N_ρ as the number of cells in $\boldsymbol{\Omega}_\rho$. Since some pixels in the window may map to the same cell, N_ρ would be a value between 1 and r^2.

Similar to the derivations of (17) from (16), we define the cell-based local prediction model as

$$\mathbf{g}_\rho = \boldsymbol{\Phi}_\rho \begin{bmatrix} \beta \\ \beta_0. \end{bmatrix}. \tag{37}$$

In (37), we use a $N_\rho \times 1$ vector $\mathbf{g}_\rho = \lfloor g_{\rho 1}, \ldots, g_{\rho N_\rho} \rfloor^T$ to denote the vector of depth values of all the cells in $\boldsymbol{\Omega}_\rho$ and denote $\boldsymbol{\Phi}_\rho = [\tilde{\boldsymbol{\varphi}}_{\rho 1}^T, \ldots, \tilde{\boldsymbol{\varphi}}_{\rho N_\rho}^T]^T$ as a matrix stacked by $\tilde{\boldsymbol{\varphi}}_i = [\boldsymbol{\varphi}_i \ \ 1]$. Note that in the local window, several pixels may map to the same cell. These many-to-one mappings are condensed into one-to-one mappings when constructing the cell-based local prediction model. The simplification of the many-to-one mappings will be compensated later in the construction of the cell-based global Laplacian matrix.

Similar to the derivation of pixel-based local model from (16) to (20), the cell-based local prediction is modeled as

$$\mathbf{g}_\rho^* = \mathbf{H}_\rho^T \mathbf{g}_\rho \tag{38}$$

where $\mathbf{H}_\rho = \boldsymbol{\Phi}_\rho(\boldsymbol{\Phi}_\rho^T \boldsymbol{\Phi}_\rho + c_\beta \mathrm{I}_\beta)^{-1} \boldsymbol{\Phi}_\rho{}^T$.
Based on (38), we define the cell-based local prior model as

$$\begin{aligned} -\log(p(\mathbf{g}_\rho)) &= \left\| \mathbf{g}_\rho - \mathbf{g}_\rho^* \right\|^2 \\ &= \left\| \mathbf{g}_\rho - \mathbf{H}_\rho^T \mathbf{g}_\rho \right\|^2 \\ &= \mathbf{g}_\rho^T \mathbf{Q}_\rho \mathbf{g}_\rho. \end{aligned} \tag{39}$$

In (39), $\mathbf{Q}_\rho = (\mathbf{I}_\rho - \mathbf{H}_\rho)^T (\mathbf{I}_\rho - \mathbf{H}_q)$ is the cell-based local Laplacian matrix. $\mathbf{I}_\rho$ is the $N_\rho \times N_\rho$ identity matrix.

After the construction of local Laplacian matrices, the cell-based global Laplacian matrix is derived via the summation of local Laplacian matrices. As mentioned, within a local window, there could be many pixels that map to the same cell. Hence, during the summation, the entry $\mathbf{Q}_\rho(i, \ j)$ has to be multiplied by a scalar that reflects both the duplicated mappings onto the cell i and the duplicated mappings onto the cell j. If we denote $\mathbf{Q}$ as the $R \times R$ cell-based global Laplacian matrix. Its entry $\mathbf{Q}(i, \ j)$ is calculated by

$$\mathbf{Q}(i, j) = \sum_{\rho \in \Omega} \eta_\rho(i)\eta_\rho(j)\mathbf{Q}_\rho(i, j) \tag{40}$$

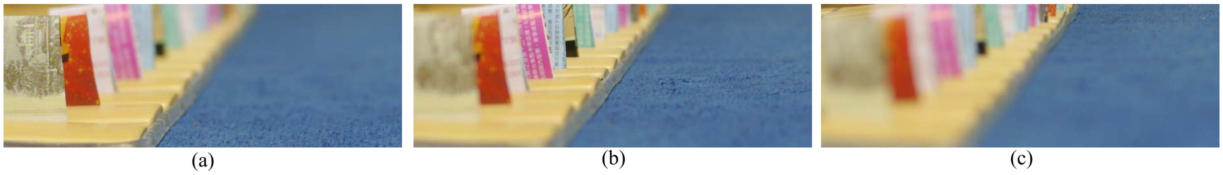

(a) (b) (c)

Fig. 4. Sample frames of the 13-frame image sequence for spatial consistency evaluation. The frame size is 1145 × 411. (a) Focused at the near end. (b) Focused in the middle. (c) Focused at the far end.

where $\eta_p(i)$ and $\eta_p(j)$ denote the number of duplicated pixels mapped into cells i and j, respectively. After the formation of $\mathbf{Q}$, the cell-based prior is modeled as

$$-\log(p(\mathbf{g})) = \mathbf{g}^T\mathbf{Q}\mathbf{g}. \quad (41)$$

C. Cell-Based MAP Estimation

With the cell-based likelihood model in (32) and the cell-based prior model in (41), the posterior probability is given by

$$-\log(p(\mathbf{I}^{\mathbf{Set}}|\mathbf{g})p(\mathbf{g})) = (\tilde{\mathbf{g}}-\mathbf{g})^T\mathbf{\Lambda}_g(\tilde{\mathbf{g}}-\mathbf{g}) + \mathbf{g}^T\mathbf{Q}\mathbf{g}. \quad (42)$$

The global minimum of (42) can be obtained by solving a system of following linear equations:

$$(\mathbf{Q}+\mathbf{\Lambda}_g)\mathbf{g} = \mathbf{\Lambda}_g\tilde{\mathbf{g}}. \quad (43)$$

Since the number of cells is typically much smaller than the number of image pixels, the dimension of the system in (43) is much smaller than the pixel-based system in (29). Hence, we can greatly reduce the computational load and efficiently estimate a cell-wise depth map.

D. Cell-Based Iterative Refinement

During the MAP estimation, the existence of some inaccurate data in the likelihood model may degrade the accuracy of depth inference. To solve this problem, a refinement process is proposed to iteratively eliminate inaccurate data. The refinement process is similar to the expectation-maximization algorithm and consists of two steps. The first step refers to the previously mentioned global optimization process in (43), while the second step refers to the update of the likelihood model. The update process aims to minimizing the influence of the inaccurate data in the likelihood model. To achieve that, after having derived the globally optimized depth values, we compare the optimized depth vector $\mathbf{g}^*$ with the previously predicted depth vector $\tilde{\mathbf{g}}$. For any cell i, if the square difference between $\mathbf{g}_i^*$ and $\tilde{\mathbf{g}}_i$ exceeds a predefined threshold t_r, we treat the previously predicted depth value as unreliable and we set the corresponding precision term $\mathbf{\Lambda}_g(i,i)$ to be zero. That is

$$\mathbf{\Lambda}_g(i,i) = \begin{cases} 0 & \text{for } \left\|\mathbf{g}_i^* - \tilde{\mathbf{g}}_i\right\|^2 > t_r \\ \mathbf{\Lambda}_g(i,i) & \text{otherwise} \end{cases} \quad (44)$$

after updating the likelihood model, the global minimum of (43) will be recomputed. In our system, we repeat this two-step refinement process five times. This iteration number five is chosen empirically.

E. Depth Map Reconstruction From Grid Cells

After obtaining the optimal cell-wise depth map by solving (43), we proceed to reconstruct a pixel-wise depth map. The construction of pixel-wise depth map is illustrated in Fig. 3(c) and (d). For any pixel i, we use $N(i)$ to denote a set of neighboring cells. In Fig. 3(d), the red dot represents a pixel in the high-dimensional space, with its neighboring cells $j \in N(i)$ colored in blue and the center of the neighboring cells marked by $\oplus$. The depth value of the pixel i can be interpolated from the depth values of its neighboring cells based on the conditional probability $p_{j|i}$

$$p_{j|i} = \frac{1}{F_i}\exp\left(-\frac{\|f_i - f_j\|^2}{\sigma_f}\right) \quad (45)$$

where $F_i = \sum_{j\in N(i)} \exp(-\|f_i - f_j\|^2/\sigma_f)$.

Here, we use f_i and f_j to denote the position of pixel i and the averaged position of the pixels inside cell j, respectively, in the high-dimensional space. The conditional probability in (45) models the probability that pixel i belongs to cell j, based on the distance between the pixel i and the averaged position of cell j in the high-dimensional space. A shorter distance between them refers to a higher probability with a Gaussian kernel and σ_f controls the bandwidth of the kernel. Finally, the interpolated depth value of pixel i, denoted as d_i^*, can be computed by

$$d_i^* = \sum_{j\in N(i)} g_j^* \cdot p_{j|i}. \quad (46)$$

IV. Experimental Results

A. Evaluation of Spatial Consistency

To evaluate the performance of our system, we first conduct an experiment to reconstruct a 3-D planar surface. In our experiment, a blue carpet is laid on the ground and a few pasteboards markers are evenly placed on the left side of the carpet to help camera focusing and the measurement of physical distance. Since this carpet is horizontally placed in the scene, a planar 3-D depth map is expected. In this experiment, a 13-frame image sequence is acquired with 13 different focus settings. Three frames of the image sequence are shown in Fig. 4. By keeping one frame for every two frames of the 13-frame sequence, we obtain a 7-frame sequence. Similarly, by keeping one frame for every four frames of the 13-frame sequence, we obtain a 4-frame sequence. Based on these three image sequences, we compare the proposed method with some related approaches, including the Laplacian-based approach [1], the ANDF approach [6], and the adaptive focus measure operator [4]. When implementing the Laplacian-based

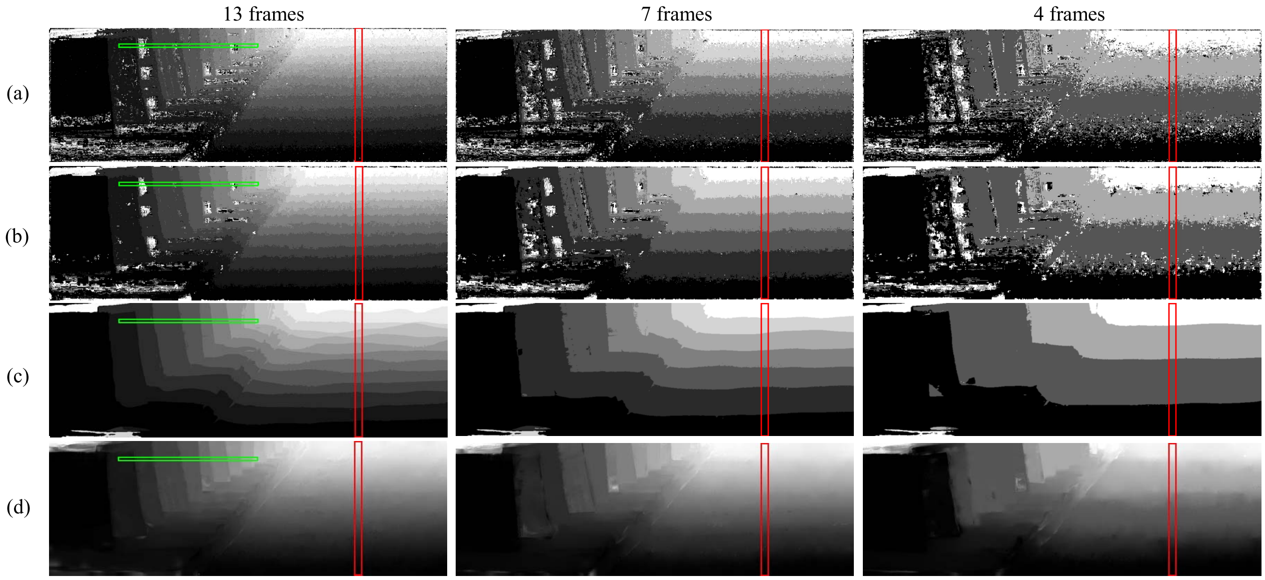


Fig. 5. Depth reconstruction results. (a) Results by [1], (b) [6], (c) [4], and (d) our results. For these depth maps, the black color indicates the closest while the white color indicates the farthest. From left to right, the depth maps are reconstructed based on the 13-, 7-, and 4-frame sequences, respectively.

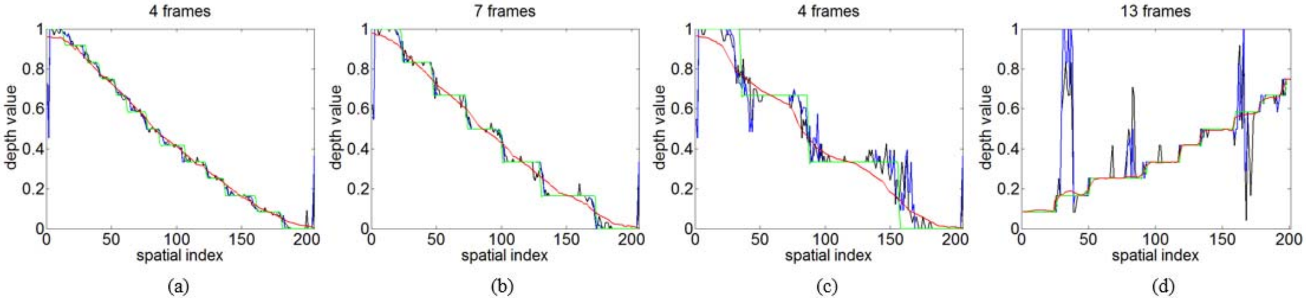
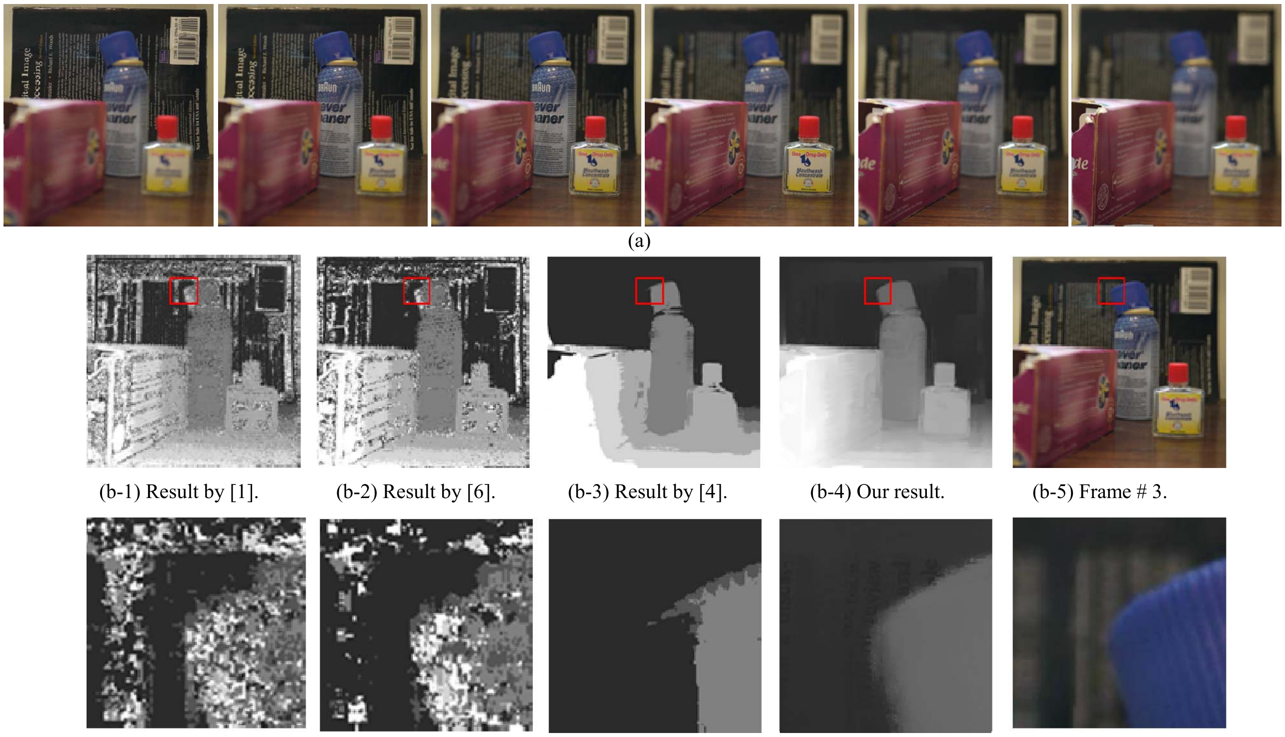


Fig. 6. (a)–(c) Depth profiles with respect to the vertical coordinate. Each curve refers to the horizontal average of the depth values from the 440th column to the 450th column of the depth maps. For these plots, black is the result by [1], blue is the result by [6], green is the result by [4], and red is the result of ours. (d) Depth profiles with respect to the horizontal coordinate based on the 13-frame sequences (bounded by the green rectangles in Fig. 5).

approach, we employ the DoG operator as described in (5). As shown in Fig. 5, the result obtained by the Laplacian-based approach is quite noisy over the low-contrast regions. On the other hand, Mahmood and Choi [6] suggest; the use of a 3-D ANDF to enhance the estimated focus volume, which refers to a stack of image planes consisting of the focus measure values of the multifocus image sequence. Unfortunately, to obtain satisfactory results, the ANDF approach usually requires a large number of image frames to form a dense focus volume. As shown in Fig. 5, the performance of the ANDF approach deteriorates quickly for the 7- and 4-frame sequences. In comparison, Aydin and Akgul [4] present an adaptive focus measure operator, which includes more information from neighboring pixels using adaptive weightings based on both the spatial distance and the color distance. Even though this approach can obtain less noisy results, the estimated depth values are restricted to discrete levels. In addition, the lack of smooth transitions between two adjacent depth values may cause inconsistent boundaries in the estimated depth map. Compared with these three approaches, our approach infers the depth values by maximizing the posterior probability. This MAP approach provides continuous depth values. In addition, the inclusion of the spatial consistency model may effectively recover the depth values for low-contrast regions. As a result, the proposed method can generate more consistent results even for the image sequences that contain only four or seven image frames.

To assess the performance of the depth maps in Fig. 5, we select a few columns bounded by the red rectangles and average the depth values along the horizontal direction. The profiles of the averaged depth values with respect to the vertical axis are plotted in Fig. 6(a)–(c). The depth profile is expected to be a decreasing function from the top to the bottom according to the planar surface shape in the 3-D scene. As shown in Fig. 6, both the ANDF approach [6] and the adaptive focus measure operator [4] may properly suppress the fluctuation of depth value. However, we can observe unstable jitters in the results of the ANDF approach and some discontinuous depth values in the results of the adaptive focus measure operator, especially for the 4-frame image sequence. In addition, since the prior model guides the depth inference through a graph, we would expect our


(a)

(b-1) Result by [1]. (b-2) Result by [6]. (b-3) Result by [4]. (b-4) Our result. (b-5) Frame # 3.

(c-1) Result by [1]. (c-2) Result by [6]. (c-3) Result by [4]. (c-4) Our result. (c-5) Frame # 3.

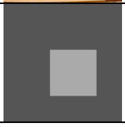
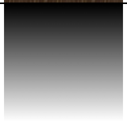
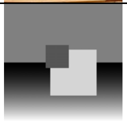
Fig. 7. (a) Frames of multifocus image sequence. The image size is 680 × 720 pixels. (b) Depth reconstruction results. For these depth maps, white indicates the closest and black indicates the farthest. (c) Zoomed depth reconstruction results.

approach can generate sharp depth edges along with sharp image structures based on the image features. Fig. 6(d) shows the depth profiles of the averaged depth values with respect to the horizontal axis bounded by the green rectangles in Fig. 5. This experiment shows that our approach can properly preserve sharp edges in the depth image, with only a slight level of blurring.

We further analyze the spatial consistency of the reconstructed depth maps around edges and low-contrast regions. Fig. 7(a) and (b) illustrate a multifocus image sequence and the reconstructed depth maps, respectively. Again, we compare our approach with the Laplacian-based approach [1], the ANDF approach [6], and the adaptive focus measure operator [4]. As shown in Fig. 7, these existing approaches usually have problems over low-contrast regions. Among these approaches, the adaptive focus measure operator [4] obtains the smoothest depth map, but it fails in avoiding the edge bleeding problem due to the lack of continuous-valued depth inference. In Fig. 7(c), we show a zoomed portion of the reconstructed depth map. In this example, both the blue cap and the background contain smooth surfaces, while the boundary between the blue cap and the background has an apparent depth change. It can be easily observed that our approach may not only properly handle the low-contrast problem but also avoid the occurrence of edge bleeding artifacts. However, our model is based the assumption of constant illumination. This assumption may fail when dealing with transparent objects or objects with specular reflection. In these cases, we may infer incorrect depth information due to the interference of varying illumination. As shown in Fig. 7(b-4), the specular reflection does induce some error in depth estimation.

TABLE I

SYNTHESIZED IMAGE SETS FOR QUANTITATIVE EVALUATION

	S1	S2	S3	S4
Image set				
Ground truth of depth				
Number of frames	4	6	6	7
Frame size	512×512	512×512	512×512	512×512

B. Quantitative Evaluation

Another experiment is conducted to quantitatively evaluate the performance of the proposed approach as compared with three related approaches [1], [4], [6]. In this experiment, depth reconstructions are performed over a set of synthesized image sequences. The synthesized scenes include disjointed planar surfaces, a tilted planar surface, a curved surface, and cluttered surfaces, as shown in Table I. The reconstructed depth images are reported in Table II. On the other hand, the corresponding mean square error (mse) measure and the bad pixel ratio are reported in Table III. Here, by setting a threshold over the

TABLE II

RECONSTRUCTED DEPTH MAPS. (a) LAPLACIAN-BASED APPROACH [1]. (b) ANDF APPROACH [6]. (c) ADAPTIVE FOCUS MEASURE OPERATOR [4]. (d) OUR APPROACH

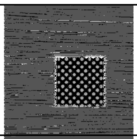
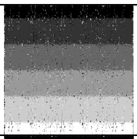
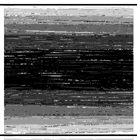
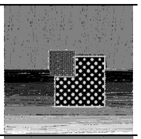
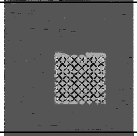
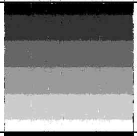
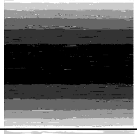
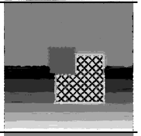
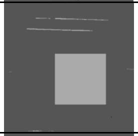
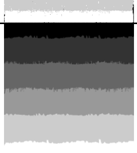
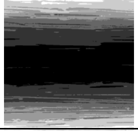
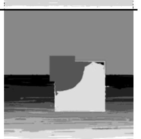
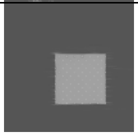
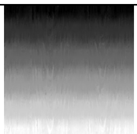
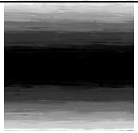
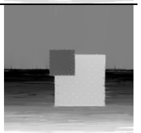
Set	S1	S2	S3	S4
(a)				
(b)				
(c)				
(d)				

TABLE III

MSE MEASURE OF RECONSTRUCTION RESULTS. RATIO OF BAD PIXEL IN RECONSTRUCTION RESULTS. ERROR Threshold $= 5e - 3$. ERROR Threshold $= 3e - 3$

Image set	S1	S2	S3	S4
(a) Laplacian [1]	5.54e-2	1.62e-2	2.70e-2	6.80e-2
(b) ANDF [6]	2.39e-2	9.20e-3	7.53e-3	4.35e-2
(c) Adaptive [4]	6.94e-4	3.40e-3	6.03e-3	8.87e-3
(d) Our method	**4.72e-4**	**5.19e-4**	**1.95e-3**	**2.08e-3**

Image set	S1	S2	S3	S4
(a) Laplacian [1]	20.2%	59.5%	49.9%	26.3%
(b) ANDF [6]	7.2%	57.7%	41.2%	14.9%
(c) Adaptive [4]	**0.62**%	57.2%	43.0%	14.5%
(d) Our method	1.07%	**0.211**%	**8.25**%	**6.12**%

Image set	S1	S2	S3	S4
(a) Laplacian [1]	20.2%	69.5%	59.1%	32.2%
(b) ANDF [6]	7.2%	68.1%	51.7%	22.0%
(c) Adaptive [4]	**0.62**%	67.3%	53.2%	19.8%
(d) Our method	1.94%	**1.24**%	**17.5**%	**14.3**%

TABLE IV

QUANTITATIVE EVALUATION FOR OUR APPROACH WITH DIFFERENT PARAMETER SETTINGS. (a) SETTING A: $b_s = 15$ AND $b_f = 7$. (b) SETTING B: $b_s = 20$ AND $b_f = 10$. (c) SETTING C: $b_s = 25$ AND $b_f = 15$. (d) WITHOUT USING GRID STRUCTURE

Mean square error measurement.

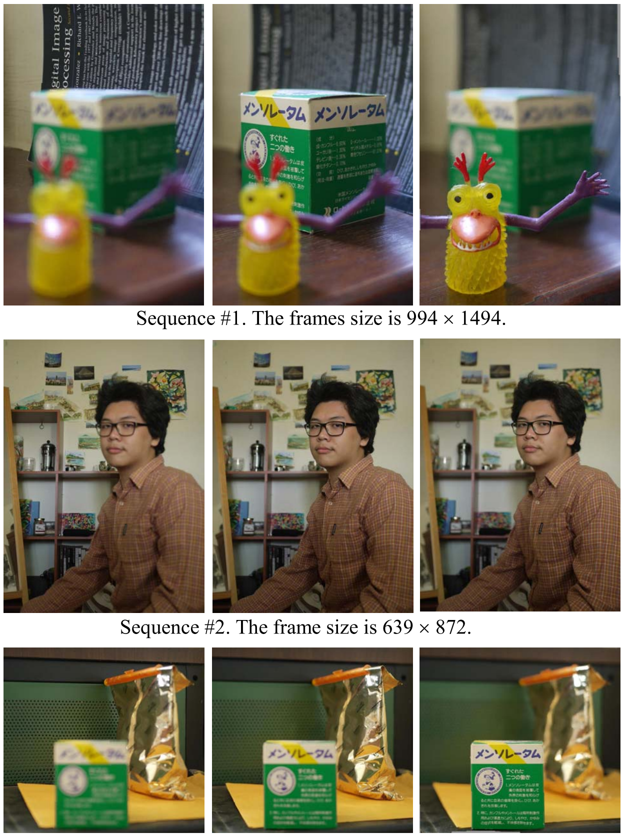
Image set	S1	S2	S3	S4
(a) Setting A	5.31e-4	1.27e-3	2.43e-3	4.40e-3
(b) Setting B	4.72e-4	5.19e-4	1.95e-3	2.08e-3
(c) Setting C	7.45e-4	1.20e-3	2.16e-3	1.82e-3
(d) Setting D	1.19e-3	2.94e-3	2.61e-3	2.26e-3

Number of cells R.

(a) Setting A	5.83k	5.56k	6.40k	3.36k
(b) Setting B	13.1k	12.5k	14.0k	7.01k
(c) Setting C	30.0k	28.5k	28.2k	13.7k
(d) Setting D	-	-	-	-

Computational time (sec).

(a) Setting A	3.68	2.92	3.96	3.27
(b) Setting B	6.26	6.05	6.74	4.28
(c) Setting C	13.1	12.6	12.1	6.70
(d) Setting D	235	214	307	271



Sequence #1. The frames size is 994 × 1494.

Sequence #2. The frame size is 639 × 872.

Sequence #3. The frame size is 1487 × 1375.

Fig. 8. Three test sequences, with each sequence containing three image frames only.

value of square error, we identify bad pixels that have large square errors and we measure the percentage of bad pixels in the depth image. In the S1 sequence, one low-contrast surface is placed at the center. The simulation result shows that our approach can effectively deal with the low-contrast problem. In S2 and S3, the challenge is to recover the continuously varying depth values. These simulation results show that our approach can provide more consistent continuous-valued depth maps. In comparison, the other three approaches can only generate discontinuous depth values.

In Table IV, the performance of the cell-based framework is evaluated in terms of mse, the number of cells, and the computation time, with respect to different parameter settings. Here, the first three types of parameter setting are obtained by adjusting the values of b_s (the number of spatial bins) and b_f (the number of feature bins) from small to large. The comparison shows that Setting B provides more balanced performance in accuracy and efficiency. In comparison, Setting A overly merges pixels into cells and the blending of conflicting data

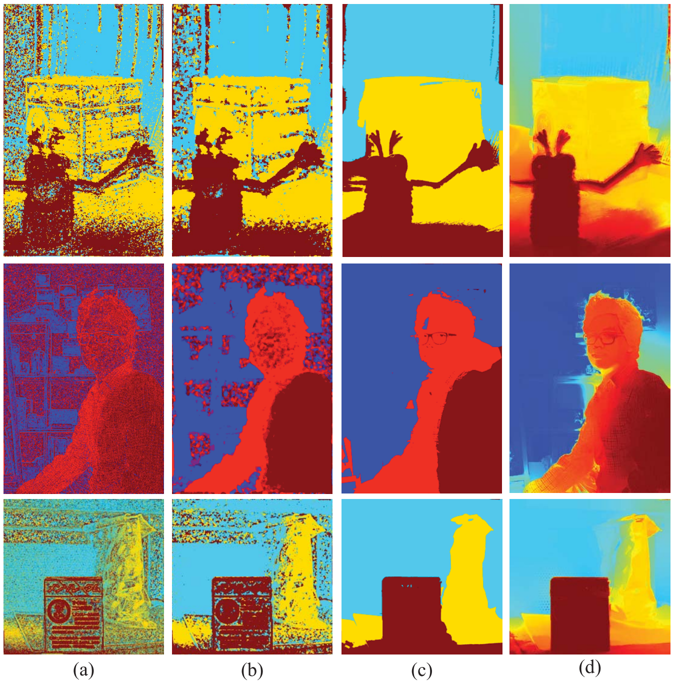


Fig. 9. Depth reconstruction results. For these depth maps, the red color indicates the closest objects, followed by the green color, and then the blue color. (a) Results by [1], (b) [6], (c) [4], and (d) our results.

may cause the degradation of accuracy. On the other hand, for Setting C, which corresponds to an oversampling situation, many cells may contain too few pixels so that the inference process may get easily biased by local observations. This also causes the degradation of accuracy. On the other hand, the fourth setting, Setting D, represents the case that does not use the grid structure at all. In this case, the accuracy is degraded and the computational time is much longer. One major factor for the degradation of accuracy in Setting D is that outlier data may easily bias the inference results. As a comparison, Settings A–C adopt the grid structure and they can effectively suppress the influence of outliers by averaging the data within each grid cell.

C. More Experiments Over Real Images

In Figs. 8 and 9, we present more experiments over real image sets for depth reconstruction. Here, we test three multi-focus image sequences, with each test sequence containing only three image frames acquired by a digital single-lens camera (Panasonic DMC-GX1 with a 20-mm f1.7 lens) using varying focus settings. In these experiments, we manually choose the camera setting to focus on objects of different depths in the scene. The goal of depth reconstruction is to infer the relative depths among the objects, rather than the physical distance of the objects away from the camera. In Fig. 9, we show the reconstructed depth maps by our approach and by the three previously mentioned approaches [1], [4], [6]. Since there are only three image frames in each sequence and there are several smooth surfaces in the scene, it is quite difficult for these existing approaches to obtain satisfactory depth maps. In comparison, our approach can generate much cleaner continuous-valued depth maps for all three cases.

TABLE V

PARAMETER SETTING

Parameter	c_β	t_0	b_s	b_f	σ_f	t_r	r
Value	10^{-3}	0.95	20	10	0.05	0.01	3

In Table V, we list the empirical parameter setting of our experiments. The meaning and the influence of these parameters are also briefly mentioned below.

1) c_β is the parameter of regularization to avoid the overfitting problem in the learning process. If c_β is too small, results would be sensitive to image noise. In contrast, a large value of c_β will cause the suppression of edge sharpness.
2) t_0 is the threshold to remove uncertain data. If t_0 is too small, it may overly remove significant data and cause the decrease of accuracy. In contrast, using a large value of t_0 may generate noisy results.
3) b_s and b_f are down-sampling parameters. Detailed descriptions and experiments about these two parameters can be found in Section IV-B.
4) σ_f controls the level of smoothness for the depth reconstruction from grid cells. In general, with a larger value of σ_f, we generate smoother depth images. With a smaller value of σ_f, we generate sharper results but may also generate inconsistent artifacts between grid cells.
5) t_r controls the amount of data to be refined. If t_r is too large, some significant data may get removed and cause the degradation of accuracy. If t_r is too small, there would be almost no refinement.
6) r corresponds to the window size. For a larger value of r, more pixels are involved in the local prediction process and the derived result would be more spatially consistent. One drawback of using a large value of r is the increase of computational load. In addition, the local affine transformation assumption would not be suitable for a large window.

D. Limitations

Even though the proposed method can improve the performance of depth inference over low-contrast regions, it still cannot effectively deal with surfaces with no texture. In such circumstance, we can only obtain information from the boundaries between the surface and its neighboring surfaces. However, the neighboring surfaces may not be at the same depth with the smooth surface so that some conflictions may occur in the depth inference process. Fig. 10 shows an example of this problem. In this example, the white wall is a texture-less background. To infer the depth value of the white wall, we can only rely on the focus measure values on the surrounding boundaries of the white wall. Unfortunately, without any clue to identify whether the white wall should share the same depth value with the foreground human body or the painting on the wall, our MAP inference process chooses a neutral solution that infers a depth value in between. Fortunately, although this inferred depth value of the nontexture surface is not correct, it may still help in distinguishing the foreground object from the surrounding background. In the future, to deal with this kind of texture-less surfaces, we may need to further discuss

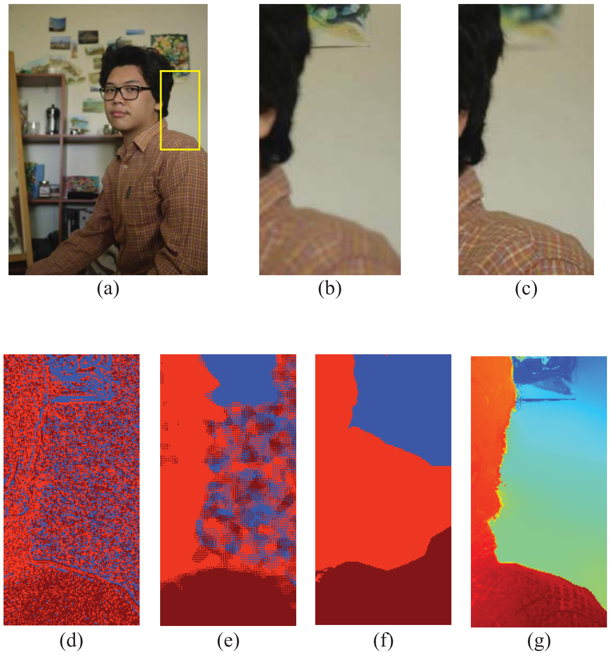


Fig. 10. Illustration of depth reconstruction for a nontexture surface. (a) Sequence #2. (b) A close look when focusing at the far end. (c) A close look when focusing at the near end. (d) Results by [1], (e) [6], (f) [4], and (g) ours.

how to learn a more robust foreground/background model for the whole image.

E. Computational Complexity

To analyze the computational complexity of the proposed approach, we divide the whole process into three major steps and analyze the complexity of each step individually. First, the computational complexity of the construction of the local prior model is dominated by the calculation of (38) as the window is scanning over the entire image pixels. For each $r \times r$ local window, the dominated complexity to calculate $\mathbf{H}_\rho$ in (38) is $O(N_\rho^2)$, where N_ρ denotes the number of corresponding cells in the window and N_ρ would be a value between 1 and r^2. Hence, the complexity of the entire local learning process would be $O(r^4 N)$, where N denotes the total number of image pixels in an image frame. Second, the complexity in solving the MAP optimization is dominated by the system of linear equations in (43). The complexity would be about $O(R^{3/2})$ by applying the conjugate gradient method, where R denotes the total number of grid cells. Empirically, the value of R is about 5–15 K. Finally, the complexity of the pixel-wise depth map reconstruction is dominated by the computation of the conditional probability $p_{j|i}$ in (45). Empirically, we choose two neighboring cells along each dimension in the 5-D space and we include 2^5 neighboring cells in the computation of (46). Hence, the complexity for the computation of (46) would be O($2^5 N$). Since typically R is much smaller than N, the computational complexity of the whole process would be $O((r^4 + 2^5)N)$.

Our algorithm has been implemented in MATLAB on an AMD FX6100 3.3-GHz CPU with 4 GB of memory. Currently, the proposed framework takes about 10 s to reconstruct an 800 × 640 depth image from a 3-frame multifocus image sequence. In comparison, for the approaches in [4] and [6], they may need several minutes to generate a depth image of similar size. This is because they need to locally refine the results of focus estimation by performing smoothing over the entire image sequence.

V. Conclusion

In this paper, we propose an MAP framework for the depth reconstruction from a multifocus image sequence. In the proposed MAP framework, a spatial-consistency prior model learned directly from the multifocus image sequence is proposed to deal with the low-SNR problem. With the inclusion of the prior model in the MAP framework, we can obtain spatially more consistent depth maps and prevent the occurrence of edge bleeding artifacts. Even for a multifocus image sequence that contains only a few image frames, the proposed method may still effectively suppress the noise and infer a reasonable depth map. The experimental results demonstrate that the proposed method can generate more convincing results as compared with some state-of-the art approaches. In addition, the proposed cell-based framework can effectively improve the computational efficiency so that the proposed SFF process can actually be applied to some practical applications.

References

[1] S. K. Nayar and Y. Nakagawa, “Shape from focus,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 8, pp. 824–830, Aug. 1994.

[2] J. M. Tenenbaum, “Accommodation in computer vision,” Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, USA, 1971.

[3] N. Yokoya, T. Shakunaga, and M. Kanbara, “Passive range sensing techniques: Depth from images,” *IEICE Trans. Inf. Syst.*, vol. E82-D, no. 3, pp. 523–533, 1999.

[4] T. Aydin and Y. S. Akgul, “A new adaptive focus measure for shape from focus,” in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2008, pp. 1–10.

[5] A. Thelen, S. Frey, S. Hirsch, and P. Hering, “Improvements in shape-from-focus for holographic reconstructions with regard to focus operators, neighborhood-size, and height value interpolation,” *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 151–157, Jan. 2009.

[6] M. T. Mahmood and T.-S. Choi, “Nonlinear approach for enhancement of image focus volume in shape from focus,” *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2866–2873, May 2012.

[7] M. B. Ahmad and T. S. Choi, “Application of three dimensional shape from image focus in LCD/TFT displays manufacturing,” *IEEE Trans. Consum. Electron.*, vol. 53, no. 1, pp. 1–4, Feb. 2007.

[8] A. S. Malik and T.-S. Choi, “Comparison of polymers: A new application of shape from focus,” *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 39, no. 2, pp. 246–250, Mar. 2009.

[9] M. Mahmood and T.-S. Choi, “Focus measure based on the energy of high-frequency components in the *S* transform,” *Opt. Lett.*, vol. 35, no. 8, pp. 1272–1274, Apr. 2010.

[10] M. Boissenin, J. Wedekind, A. N. Selvan, B. P. Amavasai, F. Caparrelli, and J. R. Travis, “Computer vision methods for optical microscopes,” *Image Vis. Comput.*, vol. 25, no. 7, pp. 1107–1116, 2007.

[11] M. Niederost, J. Niederöst, and J. Ščučka, “Automatic 3D reconstruction and visualization of microscopic objects from a monoscopic multifocus image sequence,” in *Proc. Int. Archives Photogram., Remote Sens. Spatial Inf. Sci.*, 2003.

[12] S. O. Shim, A. S. Malik, and T. S. Choi, “Accurate shape from focus based on focus adjustment in optical microscopy,” *Microsc. Res. Techn.*, vol. 72, no. 5, pp. 362–370, 2009.

[13] M. Niederoest, J. Niederoest, and J. Scucky, “Automatic 3D reconstruction and visualization of microscopic objects from a monoscopic multifocus image sequence,” in *Proc. Int. Workshop Vis. Animation Reality Based 3D Models*, 2002.

[14] V. Gaganov and A. Ignateko, “Robust shape from focus via Markov random fields,” in *Proc. Int. Conf. Comput. Graph. Vis.*, 2009, pp. 74–80.

[15] K. Ramnath and A. N. Rajagopalan, “Discontinuity-adaptive shape from focus using a non-convex prior,” in *Proc. 31st DAGM Symp. Pattern Recognit.*, 2009, pp. 181–190.

[16] C.-Y. Tseng and S.-J. Wang, “Maximum-a-posteriori estimation for global spatial coherence recovery based on matting Laplacian,” in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep./Oct. 2012, pp. 293–296.

[17] A. Levin, D. Lischinski, and Y. Weiss, “A closed-form solution to natural image matting,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 61–68.

[18] Y. Zheng and C. Kambhamettu, “Learning based digital matting,” in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 889–896.

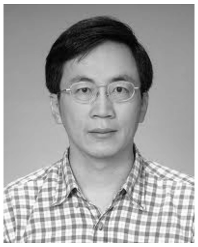
[19] B. Bollobás, *Modern Graph Theory*. New York, NY, USA: Springer-Verlag, 1998.

[20] A. Saxena, S. H. Chung, and A. Y. Ng, “Learning depth from single monocular images,” in *Neural Information Processing Systems*, vol. 18. Cambridge, MA, USA: MIT Press, 2005.

[21] A. Saxena, M. Sun, and A. Y. Ng, “Make3D: Learning 3D scene structure from a single still image,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2009.

Chen-Yu Tseng received the B.S. and M.S. degrees in electrical engineering from National Chiao Tung University, Hsinchu, Taiwan, in 2005 and 2007, respectively, where he is working toward the Ph.D. degree in electrical engineering.

His research interests include image processing and image analysis.



Sheng-Jyh Wang (M’95) received the B.S. degree in electronics engineering from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 1984, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1990 and 1995, respectively.

He is a Professor with the Department of Electronics Engineering, NCTU. His research interests include image processing, video processing, and image analysis.