Matting and Depth Recovery of Thin Structures using a Focal Stack

Chao Liu
chaoliu1@cs.cmu.edu

Artur W. Dubrawski
awd@cs.cmu.edu

Srinivasa G. Narasimhan
srinivas@cs.cmu.edu

Abstract

Thin structures such as fences, grass and vessels are common in photography and scientific imaging. They contribute complexity to 3D scenes with sharp depth variations/discontinuities and mutual occlusions. In this paper, we develop a method to estimate the occlusion matte and depths of thin structures from a focal image stack, which is obtained either by varying the focus/aperture of the lens or computed from a one-shot light field image. We propose an image formation model that explicitly describes the spatially varying optical blur and mutual occlusions for structures located at different depths. Based on the model, we derive an efficient MCMC inference algorithm that enables direct and analytical computations of the iterative update for the model/images without re-rendering images in the sampling process. Then, the depths of the thin structures are recovered using gradient descent with the differential terms computed using the image formation model. We apply the proposed method to scenes at both macro and micro scales. For macro-scale, we evaluate our method on scenes with complex 3D thin structures such as tree branches and grass. For micro-scale, we apply our method to in-vivo microscopic images of micro-vessels with diameters less than 50 μm. To our knowledge, the proposed method is the first approach to reconstruct the 3D structures of micro-vessels from non-invasive in-vivo image measurements.

1. Introduction

Thin structures such as meshes, grass or tree branches are common in photography. Similarly, in medical and microscopic imaging, thin curvilinear structures such as vessels and neurons appear very often. Recovering the 3D information for such structures with non-invasive imaging modalities is useful for study of plants [5, 25], blood vessels [20, 21], and neurons [2, 12].

Segmenting thin structures from the background and recovering their depths is a challenging task for multiple reasons. First, thin structures located in close range might occlude more distant objects. So the ray corresponding to a pixel may encounter multiple occluders at different depths due to the partial occlusion. Second, the 3D structures of curvilinear objects in nature such as vessels and grass are often complex and non-planar, thus the methods based on planarity assumption [9, 7, 29] fail in those cases. Third, because of the small widths of the thin structures, the high spatial frequency depth discontinuities are likely to be recovered coarsely using patch-based depth-from-focus/defocus methods [22, 11, 10, 3].

Figure 1: Example scenes with thin structures: mesh, grass, tree branches, and micro-vessels. Such structures are often non-planar, located at multiple depths, and occluding one another. The goal of this paper is to matte and recover depths of these thin structures from a single-view focal image stack.

In this work, we present a method for matting and depth recovery of 3D thin structures with self-occlusions using single-view focal stack images. To this end, we first propose a general image formation model that explicitly describes the spatially varying blur and multiple partial occlusions along a line of sight. Jointly optimizing the occlusion mattes and depths in the model is computationally intractable. We derive a Markov Chain Monte Carlo (MCMC) inference algorithm for the thin structure matting where the image/model update is directly and analytically computed. The analytic computation enables efficient updates of the model without re-rendering new images during the MCMC process, which makes the algorithm practical. The depths of thin structures are then recovered using gradient descent with the differential terms calculated from the model.

We evaluate the performance of the proposed method using images of scenes at both macro and micro scales. For macro-scale, we evaluate our method on scenes with complex 3D thin structures such as meshes, tree branches and

grass. For micro-scale, we apply our method to in-vivo microscopic images of micro-vessels with diameters less than 50 μm. We reconstruct the 3D structure of the micro-vessels despite spatially varying blur and occlusions. To our knowledge, this is the first method to reconstruct the 3D structures of micro-vessels from a non-invasive in-vivo imaging system.

2. Related Work

Occlusion estimation and removal: Learning-based and physics-based methods have been used to remove occluders or recover the depths and patterns of the occlusions. In [4], a neural network was trained to detect and remove the dirt of rain drops. In [16] the translational symmetry pattern of the foreground has been exploited. Other methods estimate and remove the occlusion by using an image formation model that takes into account occlusions. [6, 9, 17]. In [6], an inverse projection model is used to recover the geometry and radiance of the scene following a variational framework. Gu *et.al* [9] model the captured radiance as a superposition of the foreground then recover the occlusion pattern and the occluded background from images captured with different focus settings by assuming that the foreground is fronto-parallel and dark. In [26], the occlusions are removed using large synthetic aperture images captured with an array of cameras.

Scene matting with obstructions: Xue *et.al* [29] exploit the difference between the edge flows of the obstruction surface and the background in a video to separate and recover the foreground and background radiances. In [7], light field matting is used to recover both the foreground and background layers. In [11, 10], the simplified multilayer scene model, where the radiance is assumed to come from an all-in focus scene layer, is solved in order to perform post-capture image refocus. The radiance for all layers are approximated by a single all-in-focus radiance map. For thin structure occlusions in [7], the multilayer model is simplified to consist of a single pair of fronto-parallel foreground and background layers. Rather than first simplify the multilayer model and then solve the more constrained problem like in [10] and [7], we will directly solve the full multilayer model with multiple non-fronto-parallel occlusion layers.

Reconstruction and depth estimation with occlusions: Due to lack of correspondences, traditional 3D reconstruction methods such depth from defocus and stereo matching fail to work well on scenes with occlusions. Yamazaki *et.al* in [30] use shadows cast from a point light source to reconstruct intricate objects that are difficult for traditional shape-from-silhouettes methods.

In [28], the occlusions have been modeled in the 4D light field and the occlusions are explicitly handled to get better depth estimation near depth disparities. Photo-consistency is extended to points at the depth disparity edges to handle occlusions more explicitly. The partial occlusion is modeled in the angular space of the input 4D light field. In our method, the occlusions are modeled using the multilayer matting function based on 2D spatially varying defocus kernels. In addition, we also demonstrate our approach in cases where occluders block each other.

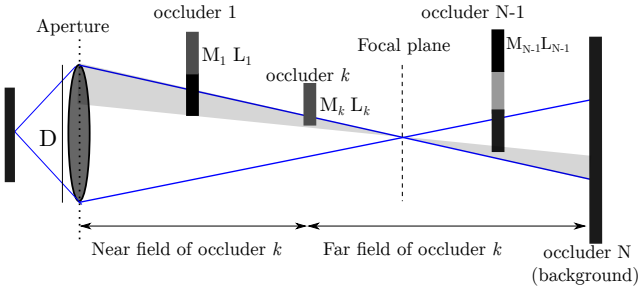


Figure 2: Viewing geometry of a single pixel in a camera with finite aperture. The camera is focused between occluder k and occluder $N-1$. The pixel receives radiance contributions from rays within the double-sided cone determined by the focal plane and aperture size. The occluders are represented with the occlusion map M and radiance map L. Occluder k is partially occluded by the occluders in its near field and occludes the occluders/background in its far field.

In [9], a single fronto-parallel layer of occlusions is removed using two or three images captured with different aperture sizes. The occlusions are assumed to be dark without contributing any radiance. In [7], the occlusion is also assumed to be in single fronto-parallel layer. In contrast, we address occlusions that are located in different depths and may occlude one another.

3. Image Formation Model

For a camera with finite aperture, one pixel at the image plane receives radiance contributions from *multiple* rays from the points within a double-sided cone determined by the focal plane and aperture size, as shown in Figure 2. With the image coordinate denoted as $\mathbf{v}$, we represent the occluder with occlusion matte $M(\mathbf{v}) \in \{0, 1\}$ and radiance $L(\mathbf{v}) \in \mathcal{R}^+$. If there is only one opaque occluder in the scene, the image intensity at $\mathbf{v}$ in the m-th image R^m in the focal stack is

$$R^m(\mathbf{v}) = \int_{\mathbf{u}} L(\mathbf{u}) M(\mathbf{u}) B^m\left(\mathbf{v} - \mathbf{u}; d(\mathbf{v})\right) \mathrm{d}\mathbf{u}$$

where $B^m\left(\mathbf{v} - \mathbf{u}; d(\mathbf{v})\right)$ is the spatially varying blur kernel dependent on the scene point depth $d(\mathbf{v})$.

For scenes with opaque occluders located at multiple depths, the image intensity for one pixel is contributed by

multiple points at different depths, with possible attenuations due to occlusions as shown in Figure 2. We denote the occlusion index $k \in \{1, 2, \cdots, N\}$ to be the order in which the double-sided cone from the camera encounters the scene points. The image R^m is the superposition of contributions from scene points across all occlusion indexes:

$$R^m(\mathbf{v}) = \sum_{k=1}^{N} \alpha_k^m(\mathbf{v}) \int_{\mathbf{u}} L_k(\mathbf{u}) M_k(\mathbf{u}) B_k^m(\mathbf{v}-\mathbf{u}; d_k(\mathbf{v}))\, \mathrm{d}\mathbf{u} \tag{1}$$

with $B_k^m(\mathbf{v}-\mathbf{u}; d_k(\mathbf{v}))$ denoting the spatially varying blur kernel for the scene point with occlusion index k. The attenuation term $\alpha_k^m(\mathbf{v})$ describes the attenuation of the radiance from occluder k due to occlusions. As shown in Figure 2, the occluder with occlusion index $k > 1$ is only obstructed by points in the near field with occlusion index smaller than k, thus the attenuation term can be written as:

$$\alpha_k^m(\mathbf{v}) = \begin{cases} 1, & \text{if } k = 1 \\ \prod_{j=1}^{k-1} 1 - \int_{\mathbf{u}} M_j(\mathbf{u}) B_j^m(\mathbf{v}-\mathbf{u}; d_j(\mathbf{u})), & \text{otherwise} \end{cases} \tag{2}$$

Eq. 1 and Eq. 2 describes the general case shown in Figure 2 where the defocus blur is spatially-variant and the occluders in the scene may partially occlude one another.

Because the blur kernels in Eq. 1 and Eq. 2 are compact in space, the range of $\mathbf{u}$ in the integral is within a local patch $\mathcal{N}(\mathbf{v})$. So we can write the discretized image formation model as:

$$R^m(\mathbf{v}) = \sum_{k=1}^{N} \alpha_k^m(\mathbf{v}) \sum_{\mathbf{u}\in\mathcal{N}(\mathbf{v})} L_k(\mathbf{u}) M_k(\mathbf{u}) B_k^m(d_k(\mathbf{v})) \tag{3}$$

with the discretized attenuation term:

$$\alpha_k^m(\mathbf{v}) = \begin{cases} 1, & \text{if } k = 1 \\ \prod_{j=1}^{k-1} 1 - \sum_{\mathbf{u}\in\mathcal{N}(\mathbf{v})} M_j(\mathbf{u}) B_j^m(d_j(\mathbf{u})), & \text{otherwise} \end{cases} \tag{4}$$

with $B(d(\mathbf{u})) = B(\mathbf{v}-\mathbf{u}; d(\mathbf{u}))$ for notation simplicity.

The image formation model in Eq. 3 and Eq. 4 generalizes the models used in previous works. When occluders are fronto-parallel, the blur kernel for each occluder is spatially-invariant. In this case the integrals in Eq. 3 and Eq. 4 become convolutions with blur kernels $B_j^m(\mathbf{v}-\mathbf{u})$. For $N = 1$, and the image formation model becomes:

$$R^m(\mathbf{v}) = L_1 M_1 * B_1(\mathbf{v})$$

which is the scene model used in [11, 10] for depth recovery and post-capture re-focusing. For $N = 2$, there is only one occlusion in front of the background, the image formation model becomes:

$$R^m(\mathbf{v}) = L_1 M_1 * B_1(\mathbf{v}) + (1 - M_1 * B_1(\mathbf{v}))(L_2 * B_2(\mathbf{v}))$$

which is the image formation model used in previous works on image matting [18, 15] and occlusion reasoning [9].

4. Efficient MCMC for Occlusion Matting

In this work, the goal is to estimate the occlusion matte $M_k(\mathbf{v})$, depth $d_k(\mathbf{v})$ and scene radiance $L_k(\mathbf{v})$ for occlusion index $k \in \{1, 2, \cdots, N\}$, given the measured focal stack images and calibrated defocus blur kernels B_k^m. In the following, we will first describe our method to estimate the occlusion mattes from a focal stack, followed by the depth recovery for the occluders explained in Section 5.

Given the measured focal stack images $\{I^m(\mathbf{v})\}$ captured with different focal plane distances, the estimated occlusion mattes $M(\mathbf{v})$ are determined by minimizing the energy function:

$$E(M(\mathbf{v})) = E_{\text{data}}(M(\mathbf{v})) + \lambda E_{\text{smooth}}(M(\mathbf{v}))$$

with

$$E_{\text{data}} = \sum_{m,\mathbf{v}} (I^m(\mathbf{v}) - R^m(\mathbf{v}))^2$$

$$E_{\text{smooth}} = \sum_{(\mathbf{u},\mathbf{v})\in\mathcal{N}_8} 1 - \delta(M(\mathbf{u}) - M(\mathbf{v}))$$

where the smoothness term E_{smooth} enforces the local spatial consistency for occlusion matting. $R^m(\mathbf{v})$ is the forward rendered image using the image formation model in Eq. 3 and Eq. 4. We can see from Eq. 4 that changing the occlusion matte value $M_j(\mathbf{v})$ will effect the attenuation terms α_k for all $k > j$. The range of the influence is the size of the blur kernel, which could be large when the occluder is highly defocused. This influence is propagated to the other occlusion mattings M_k through Eq. 3. Therefore, there are high-order relationships among the occlusion mattings values. So the data term E_{data} is of high-order w.r.t. $M_k(\mathbf{v})$ for $k \in \{1, 2, \cdots, N\}$.

Because of these high-order relationships, traditional graph-based methods dealing with relatively low-order potentials will not apply. Methods that include high-order potentials [23, 8, 13, 14] either require the graph to be in specific structure [23] or the relationship can be analytically modeled [8, 13, 14]. Instead, we derive an efficient MCMC inference method where the image/model updates are *directly* and *analytically* computed based on the image formation model without re-rendering the images. This makes an otherwise intractable problem practical to solve.

We will assume: 1) the radiances of the thin structures are different from the radiance of background; 2) the maximal number of occlusions along a line of sight is known or

pre-set. The first assumption enables us to detect and separate the occluders from background using the focal stack; the second assumption simplifies the derivation.

MCMC inference:

Consider a point $\mathbf{x}$ on the k-th occluder on the line of sight, as shown in Figure 2. During the MCMC inference process, the occlusion matte value $M_k(\mathbf{x}) \in \{0, 1\}$ is sampled from the probability distribution:

$$p\left(M_k(\mathbf{x}) = 1\right) = \frac{\mathrm{e}^{-E(M_k(\mathbf{x})=1)/T}}{\mathrm{e}^{-E(M_k(\mathbf{x})=1)/T} + \mathrm{e}^{-E(-M_k(\mathbf{x})=0)/T}} = \frac{\mathrm{e}^{-\Delta E(\mathbf{x})/T}}{1 + \mathrm{e}^{-\Delta E(\mathbf{x})/T}} \tag{5}$$

where $E(M_k(\mathbf{x}) = b)$ for $b = \{0, 1\}$ are the energy functions for the binary assignments for $M_k(\mathbf{x})$; $\Delta E(\mathbf{x}) = E(M_k(\mathbf{x}) = 1) - E(M_k(\mathbf{x}) = 0) = \Delta E_{\text{data}} + \lambda \Delta E_{\text{smooth}}$ represents the increase of the energy function when the sampling in the MCMC process changes the occlusion matte value at $\mathbf{x}$ $M_k(\mathbf{x})$ from 0 to 1. T is the temperature parameter controlling the acceptance rate for an update and the convergence of the MCMC process.

Estimating ΔE_{data} for MCMC Inference:

By denoting $R^m(\mathbf{v}; b)$ to be the forward rendered image when $M_k(\mathbf{v}) = b$ for $b = \{0, 1\}$, the data term of ΔE can be written as:

$$\begin{aligned}\Delta E_{\text{data}} &= \sum_{m,\mathbf{v}} (I^m(\mathbf{v}) - R^m(\mathbf{v}; 1))^2 - (I^m(\mathbf{v}) - R^m(\mathbf{v}; 0))^2 \\ &= \sum_{m,\mathbf{v}} \Delta R^m(\mathbf{v}) \left(\Delta R^m(\mathbf{v}) + 2\left(R^m(\mathbf{v}; 0) - I^m(\mathbf{v})\right)\right)\end{aligned} \tag{6}$$

where $\Delta R^m(\mathbf{v}) = R^m(\mathbf{v}; 1) - R^m(\mathbf{v}; 0)$ is the change of the rendered image by changing $M_k(\mathbf{v})$ from 0 to 1. Similarly, we can write the change of the data term for switching $M_k(\mathbf{v})$ from 1 to 0 as:

$$\Delta E_{\text{data}} = \sum_{m,\mathbf{v}} \Delta R^m(\mathbf{v}) \left(-\Delta R^m(\mathbf{v}) + 2\left(R^m(\mathbf{v}; 1) - I_i(\mathbf{v})\right)\right) \tag{7}$$

Analytically Computing $\Delta R^m(\mathbf{v})$ for ΔE_{data}:

The naive approach is to render images $R^m(\mathbf{v}; 0)$ and $R^m(\mathbf{v}; 1)$ directly and estimate $\Delta R^m(\mathbf{v})$ for all pixels and occlusion indexes. In addition, we need several iterations since the results from the burn-in period of the MCMC process is not reliable. So the computational complexity for the naive approach is too high for any real world application.

Rather than perform the full forward render process for each pixel, we propose to directly and analytically compute the $\Delta R^m(\mathbf{v})$ and its corresponding energy difference ΔE_{data} by using the image formation model in Eq. 3 and Eq. 4. The image intensity change $\Delta R^m(\mathbf{v})$ induced by switching $M_k(\mathbf{x})$ from 0 to 1 is contributed by radiance change from occluder k and occluders with occlusion index $i > k$ on the line of sight:

$$\Delta R^m(\mathbf{v}) = \alpha_k(\mathbf{v}) B_k^m(d_k(\mathbf{x})) L_k(\mathbf{x}) + \sum_{i=k+1}^{N} \Delta\alpha_i(\mathbf{v}) \tilde{L}_i(\mathbf{v}) \tag{8}$$

with the defocused image

$$\tilde{L}_i(\mathbf{v}) = \sum_{\mathbf{u} \in \mathcal{N}(\mathbf{v})} L_i(\mathbf{u}) M_i(\mathbf{u}) B_i^m(d_i(\mathbf{u})), \tag{9}$$

where $B_k(d_k(\mathbf{x}))$ and $B_i(d_i(\mathbf{u}))$ are spatially varying blur kernels; $\Delta\alpha_i(\mathbf{v})$ is the change of attenuation by switching the occlusion matte value $M_k(\mathbf{x})$ from 0 to 1. The first term in Eq. 8 is the radiance change contribution from the k-th occluder. The second term is the radiance change contributions from the occluders/background in the far field of occluder k.

Analytically Computing $\Delta\alpha_i(\mathbf{v})$ for $\Delta R^m(\mathbf{v})$:

For notation simplicity, we denote the blurred occlusion matte in Eq. 4 with:

$$\tilde{M}_j(\mathbf{v}) = \sum_{\mathbf{u} \in \mathcal{N}(\mathbf{v})} M_j(\mathbf{u}) B_j^m(d_j(\mathbf{u})) \tag{10}$$

From Eq.4, the attenuation change $\Delta\alpha_i(\mathbf{v})$ can be written as:

$$\begin{aligned}\Delta\alpha_i(\mathbf{v}) &= \left(1 - \tilde{M}_k(\mathbf{v}; 1)\right) \prod_{j=1; j\neq k}^{i-1} \left(1 - \tilde{M}_j(\mathbf{v})\right) \\ &\quad - \left(1 - \tilde{M}_k(\mathbf{v}; 0)\right) \prod_{j=1; j\neq k}^{i-1} \left(1 - \tilde{M}_j(\mathbf{v})\right) \\ &= -B_k(d_k(\mathbf{x})) \prod_{j=1; j\neq k}^{i-1} \left(1 - \tilde{M}_j(\mathbf{v})\right)\end{aligned} \tag{11}$$

By combining Eq. 8 and Eq. 11, we see that the image intensity change $\Delta R^m(\mathbf{v})$ induced by switching $M_k(\mathbf{x})$ from 0 to 1 is *independent* from $M_k(\mathbf{v})$ $\forall$ $\mathbf{v}$. Before the MCMC process for points at occluders with occlusion index k, we can pre-compute $\tilde{L}_i$ in Eq. 9 and blurred occlusion mattes $\tilde{M}_j$ in Eq. 10. Then during the MCMC process, the image update $\Delta R^m(\mathbf{v})$ can be *directly* and *analytically* estimated from Eq. 8 and Eq. 11. If the occlusion matte at pixel $\mathbf{x}$ changes after sampling from Eq. 5, the updated image can be easily computed with $R^m(\mathbf{v}) \leftarrow R^m(\mathbf{v}) + \Delta R^m(\mathbf{v})$ without re-rendering the images. In addition, due to the limited size of the blur kernel, the spatial range of $\Delta R(\mathbf{v})$ is limited within a small patch $\mathcal{N}(\mathbf{x})$ rather than over the whole image. In our implementation, we choose the size of

the patch $\mathcal{N}(\mathbf{x})$ to be 31-by-31. Therefore, the high-order data term change ΔE_{data} can be computed efficiently.

Estimating ΔE_{smooth}:

For the smoothness term change ΔE_{smooth}, since it does not include the forward rendering, it can be simply computed as:

$$\Delta E_{\text{smooth}} = \sum_{\mathbf{u}\in\mathcal{N}_8(\mathbf{x})} \delta(M(\mathbf{u})) - \delta(M(\mathbf{u})-1) \qquad (12)$$

when $M(\mathbf{x})$ changes from 0 to 1 and

$$\Delta E_{\text{smooth}} = \sum_{\mathbf{u}\in\mathcal{N}_8(\mathbf{x})} \delta(M(\mathbf{u})-1) - \delta(M(\mathbf{u})) \qquad (13)$$

for $M(\mathbf{x})$ changes from 1 to 0. $\mathcal{N}_8(\mathbf{x})$ is the 8-connectivity neighborhood of $\mathbf{x}$. As we can see from Eq. 12 and Eq. 13 the change of the smoothness term is simply the difference of numbers of occupant and empty pixels around $\mathbf{x}$.

Initialization:

A good initialization of the variables is important given the huge search space for the occlusion matte. To initialize the occlusion matte, for each pixel $\mathbf{v}$ in the measured image $I^m(\mathbf{v})$, we first compute the variance of Laplacian in the Lab color space of a local 9-by-9 patch around $\mathbf{v}$. For the occlusion matting M_k with occlusion index $k < N$, we set $M_k(\mathbf{v}) = 1$ if the maximal local variance happens in a focal depth is smaller than a pre-defined threshold and 0 otherwise. The matting $M_N(\mathbf{v}) = 1$ for all pixels for the background since any line of sight will intersect with the background. For depth initialization, the initial depth for the thin structures at one pixel is estimated as the depth index in the focal stack with the largest variance of Laplacian of a local patch around that pixel. The radiance for the points on the thin structures is the measured image intensity in the corresponding image in the focal stack. During the optimization, the radiance values are updated based on the current depth estimation, which is explained in the next section. Given the initialization, the steps for the MCMC inference for $M_k(\mathbf{v})$ are described in Alg. 1.

Algorithm 1 Efficient MCMC inference for occlusion matte $M_k(\mathbf{v})$

Given initialization of $M_k^{(0)}$, d_k and L_k render $R^{(0)}$
for each iteration t **do**
  for each occlusion index $k \in \{1, 2, \cdots, N-1\}$ **do**
    compute $\alpha_k(\mathbf{v})$ and R using Eq.3 and Eq.4
    update $\tilde{M}$ and $\tilde{L}$ using Eq.10 and Eq.9
    for each pixel $\mathbf{x}$ with occlusion indx k **do**
      compute ΔR using Eq. 8 and Eq. 11
      compute ΔE_{data} using Eq. 6 or Eq. 7;
      compute ΔE_{smooth} using Eq. 12 or Eq. 13;
      sample $M_k(\mathbf{x})$ using Eq. 5;
      $R \leftarrow R + \Delta R$ if $M_k(\mathbf{x})$ changes.
    end for
  end for
end for

5. Estimating Depths of Thin Structures

In order to compute the depth, we assume that the objects are locally planar within a small area. Given the matting estimation, we first over-segment the matted thin structures into super-pixels using SLIC [1] implemented in [27]. To get small and thin super-pixels, we set the area of the super-pixel to be 10 and the regularization factor to be 0.1. Each super-pixel will be treated as one tiny planar segment in space. The depth of the occluder is recovered by optimizing the parameters of all the foreground planar segments such that the synthetic images given the depth are as close as possible to the measured focal stack.

Given a planar segment i with plane parameters $\mathbf{s}_i$, the depth of the point on the segment with pixel coordinate (x, y) is $d = \mathbf{s}_i^T(x, y, 1)$. By concatenating all the plane parameters for N_s segments into a $3N_s$-dimensional vector $\mathbf{s}$, the optimal parameters for segment planes are found by:

$$\min_{\mathbf{s}} \quad \sum_{n,m,\mathbf{v}} (I_n^m(\mathbf{v}) - R_n^m(\mathbf{v};\mathbf{s}))^2 + \lambda_d E_s(\mathbf{s}), \qquad (14)$$

where $I_n^m(\mathbf{v})$ is the measured image intensity of segment n at pixel $\mathbf{v}$. The first term of the energy is the data term measuring the difference between the synthesized images and the measured focal stack. The second term $E_s(\mathbf{s})$ is the smoothness term enforcing the depth smoothness for adjacent segments in 3D space. For two adjacent segments representing by their plan parameters $\mathbf{s_i}$ and $\mathbf{s_j}$, the depth smoothness energy is defined as the depth difference for the pixels on their shared boundary:

$$\begin{aligned} E_s^{(i,j)} &= (\mathbf{d_i} - \mathbf{d_j})^T(\mathbf{d_i} - \mathbf{d_j}) \\ &= (\mathbf{s_i} - \mathbf{s_j})P_b^T P_b(\mathbf{s_i} - \mathbf{s_j}) \\ &= (\mathbf{s_i} - \mathbf{s_j})A^{(i,j)}(\mathbf{s_i} - \mathbf{s_j}), \end{aligned}$$

where the three-column matrix P_b consists of the homogeneous coordinates of the pixels on the boundary of segment i and segment j; $A^{(i,j)} = P_b^T P_b$. The smoothness energy for all pairs of adjacent segments can be written in a way such that it is quadratic in terms of the concatenated plan parameters $\mathbf{s}$:

$$E_s(\mathbf{s}) = \mathbf{s}^T\Lambda\mathbf{s} = \mathbf{s}^T \sum_{(i,j)\in\mathcal{N}} \Lambda^{(i,j)}\,\mathbf{s}, \qquad (15)$$

where $\Lambda^{(i,j)}$ is a $3N$-by-$3N$ sparse matrix for the neighborhood segments $\mathbf{s_i}$ and $\mathbf{s_j}$ with non-zero block entries

$\Lambda_{i,i}^{(i,j)} = \Lambda_{j,j}^{(i,j)} = A^{(i,j)}$ and $\Lambda_{i,j}^{(i,j)} = \Lambda_{j,i}^{(i,j)} = -A^{(i,j)}$.

To optimize the objective function defined in Eq. 14 using gradient-based method, we also need to calculate the gradient of the data term with respect to the plane parameters $\mathbf{s}$, for which we need to estimate:

$$\frac{\partial R}{\partial \mathbf{s}} = \frac{\partial R}{\partial \mathbf{d}} \frac{\partial \mathbf{d}}{\partial \mathbf{s}} = \frac{\partial R}{\partial \mathbf{d}} P \tag{16}$$

with P the N-by-3 location matrix with each row as the homogeneous coordinate $(x, y, 1)$ for one pixel.

For a point corresponding to pixel $\mathbf{v}$ on the k-th occluder, the gradient of the rendered image w.r.t. its depth can be written as :

$$\frac{\partial R}{\partial d_k}(\mathbf{v}) = \alpha_k(\mathbf{v}) L_k(\mathbf{v}) \frac{\partial B_k}{\partial d_k} + \sum_{i=k+1}^{N} \frac{\partial \alpha_i}{\partial d_i}(\mathbf{v}) \tilde{L}_i(\mathbf{v}) \tag{17}$$

with

$$\frac{\partial \alpha_i}{\partial d_i}(\mathbf{v}) = -\frac{\partial B_i}{\partial d_i} \prod_{\substack{j=1 \\ j \neq k}}^{i-1} 1 - \sum_{\mathbf{u} \in \mathcal{N}(\mathbf{v})} M_j(\mathbf{u}) B_j^m(d_j(\mathbf{u})) \tag{18}$$

The derivation is similar as in Section 4. The differential blur kernel $\frac{\partial B}{\partial d}$ is pre-computed during the calibration process detailed in the supplementary material. The gradient of Eq.14 can be evaluated by combing Eq.15, Eq.16 and Eq.17. We use the conjugate-gradient method for optimizing the plane parameters $\mathbf{s}$. Given the the optimal $\mathbf{s}$, the depth of the segments is calculated as $\mathbf{d} = P\mathbf{s}$.

6. Experiments

6.1. Implementation Details

For all experiments, we choose the size of the local patch for MCMC update to be 31-by-31. We set the maximal occlusion index $N = 3$. The temperature parameter T in Eq. 5 is set to 5 and the smoothness parameter λ in Eq. 5 is set to 0.8. For depth estimation, we set the depth smoothness factor λ_d in Eq. 14 to be 0.5 and the step size of the gradient descent to be 0.1. The MCMC process converges within 10 iterations and the gradient descent for depth recovery converges within 50 iterations. The running time on a 620x480 focal stack with 26 focal planes is about 20 min using MATLAB implementation on a desktop with Intel Core-i7 5940 CPU and 64 GB RAM memory size.

6.2. Calibrating Blur Kernels

For macro-scale scenes, we use a Lytro ILLUM light field camera to generate the focal stack with 26 focal planes. Using a light field camera avoids the magnification variation due to focal changes and the need for post-processing to compensate the magnification. The refocused images are estimated from the 4D light field images by shearing the light field and projecting it into 2D slices as described in [19].

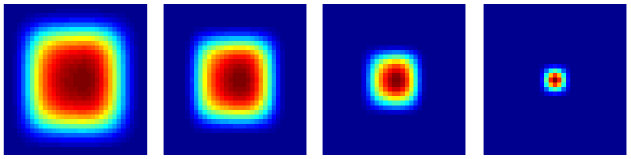


Figure 3: The calibrated blur kernels of refocused image for a plane placed 520 mm from the light field camera. The shapes of the blur kernels are not circularly symmetrical since the blur kernel for a refocused image from light field camera is related to both the main lens shape and the spatial arrangement of the secondary lenslets.

We calibrate the blur kernels for a set of 21 reference depths from 200 mm to 1000 mm equally spaced with 40 mm. In the calibration process, we use a planar reference plane with checkerboard textures and place the plane parallel to the image plane. The optical blur kernel is assumed to be a separable filter kernel such that it can be written as a convolution of two 1D functions. Then the 1D functions are optimized. The details are in the supplementary material.

Examples of the calibrated blur kernels for the focal stack images generated using the light field are shown in Figure 3. Note that the shape of the blur kernel is not circularly symmetrical since the blur kernel for a refocused image from light field camera is related to both the main lens shape and the arrangement of the secondary lenslets array. For the microscopic camera, we model the blur kernels as Gaussian functions with σ related to the focal plane distance and scene depth.

6.3. Aperture Size vs. Depth vs. Occluder Size

We first analyze the performance of our method under varying camera and scene configurations to evaluate the influence of aperture size, the depths and widths of the occluders. We synthesize the focal stack images with different camera and scene settings. With larger aperture size, we are able to collect rays from more angles coming from a point thus more rays can be imaged from the occluded regions [26]. The benefit of having a finite aperture decreases as the foreground occlusions are further from the camera. The synthetic scene includes two foreground occlusion layers with parts of the second layer being occluded.

The performance is evaluated in terms of the averaged error ratio of the rendered focal stacks. As shown in Figure 4, the reconstruction error decreases as the aperture size becomes larger since for larger aperture size, more rays from the partially occluded regions are collected. On the other hand, more reconstruction error of the background is introduced when the occlusion is closer to the background as re-

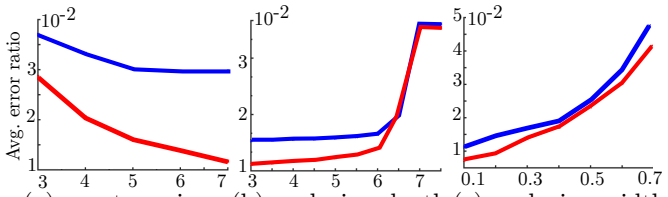


Figure 4: The reconstruction error varies with camera aperture size, the depth and size of occluder. The blue and red curves in (a) are errors with occlusion distance set to 5 and 6 respectively. The blue and red curves in (b)(c) correspond to aperture size 5 and 11 respectively.

Table 1: RMSEs of the recovered depth for the slanted plane placed at different depths.

	Distance from the camera (mm)			
	250	380	510	680
DFF [24]	94.22	61.50	129.1	161.2
Proposed	30.49	34.35	36.13	60.18

gions in the background are completely occluded. Similar results can be observed for occlusions with different sizes.

6.4. Performance on Real Data

To quantitatively assess the performance of the proposed matting and depth recovery method, we place slanted planar mesh at measured distances and evaluate the matting and depth estimations. The selected set distances are listed in Table 1. We compare our method with the baseline depth-from-focus (DFF) method used in [24]. The occlusion matte is estimated by thresholding the recovered depth map based on the fact that the mesh plane is located in the near field. Because the degree of in-focus is measured from the image intensities within a local patch, the DFF method tends to generate an over-smoothed depth map where the depth estimations near the occlusion boundaries are inaccurate. In our method, since the defocus and occlusion are modeled explicitly for each pixel, we are able to recover the high frequency depth discontinuities for thin structures. Therefore, as shown in Table 1, the RMSEs of estimated depths for our method are lower than the ones for the DFF approach.

We also compare with the approach [28] using light field inputs with the occlusion boundaries explicitly modeled. As shown in Figure 5, the DFF method in [24] fails to recover high frequency depth changes in regions such as the edges of the grass where multiple depth discontinuities are close. This is because the patch-based estimation of the degree of in-focus will include the edges of the depth boundary even if the center of the patch is not aligned on the boundary. As a result, the degree of in-focus is inaccurate near the depth boundary. The method in [28] is able to estimate the depths at places where the occlusion boundaries are close because in this method the occlusion boundaries are modeled and processed explicitly. However, the approach in [28] is unreliable for textures regions in the background. In addition, some sharp intensity edges in the background, such as the shadow boundaries, are estimated as occlusion boundaries and the recovered depths around those edges are inaccurate. In comparison, our method estimate the occlusion matting and depths pixelwisely , so it can handle sharp edges in the background and high spatial frequency depth changes for thin structures like mesh, grass and bush branches.

We apply our method to in-vivo micro-scale images of capillaries with diameter less than 50 μm. We use the Braedius CytoCam Camera to capture focal stacks of micro-vessels on the tongue of pigs. The focal planes distance range from 20 μm to 240 μm with step size of 20μm. As shown in Figure 6, the occlusion matting and depths of micro-vessels are estimated in the presence of spatially varying defocus blur and occlusions. Then we reconstruct the 3D structure of the micro-vessels based on the depth map. To our knowledge, our method is the first approach to reconstruct the 3D structures of capillaries using non-invasive image measurements.

7. Conclusions

We presented a method for matting and depth recovery for thin structures from a focal stack. We proposed a general image formation model with the spatially varying blur and mutual occlusions explicitly accounted for. Based on the model, for matting, we design an efficient MCMC inference method where the image/model update is computed analytically without explicitly rendering new images. The depth of thin structures is then recovered using gradient descent with the differential terms calculated from the image formation model. We evaluated the proposed method on images of scenes at both macro and micro scales.

We assume that the sizes/widths of objects are small compared to the aperture. In addition, if the foreground objects are far away from the camera, the camera model degrades to a pinhole camera model and the image formation model in Section 3 is invalid. To handle larger/distant occlusions, we can extend the method to include multiple cameras such that a large synthetic aperture [26] can be obtained. Another future direction is to extend the approach to scenes with transparent or semi-transparent occlusions, such as smoke, glass, and water droplets.

8. Acknowledgements

We thank the Disruptive Healthcare Technology Institute (DHTI) supported by Highmark Inc. and Allegheny Health Network, and NSF (award 1320347) for supporting this work.

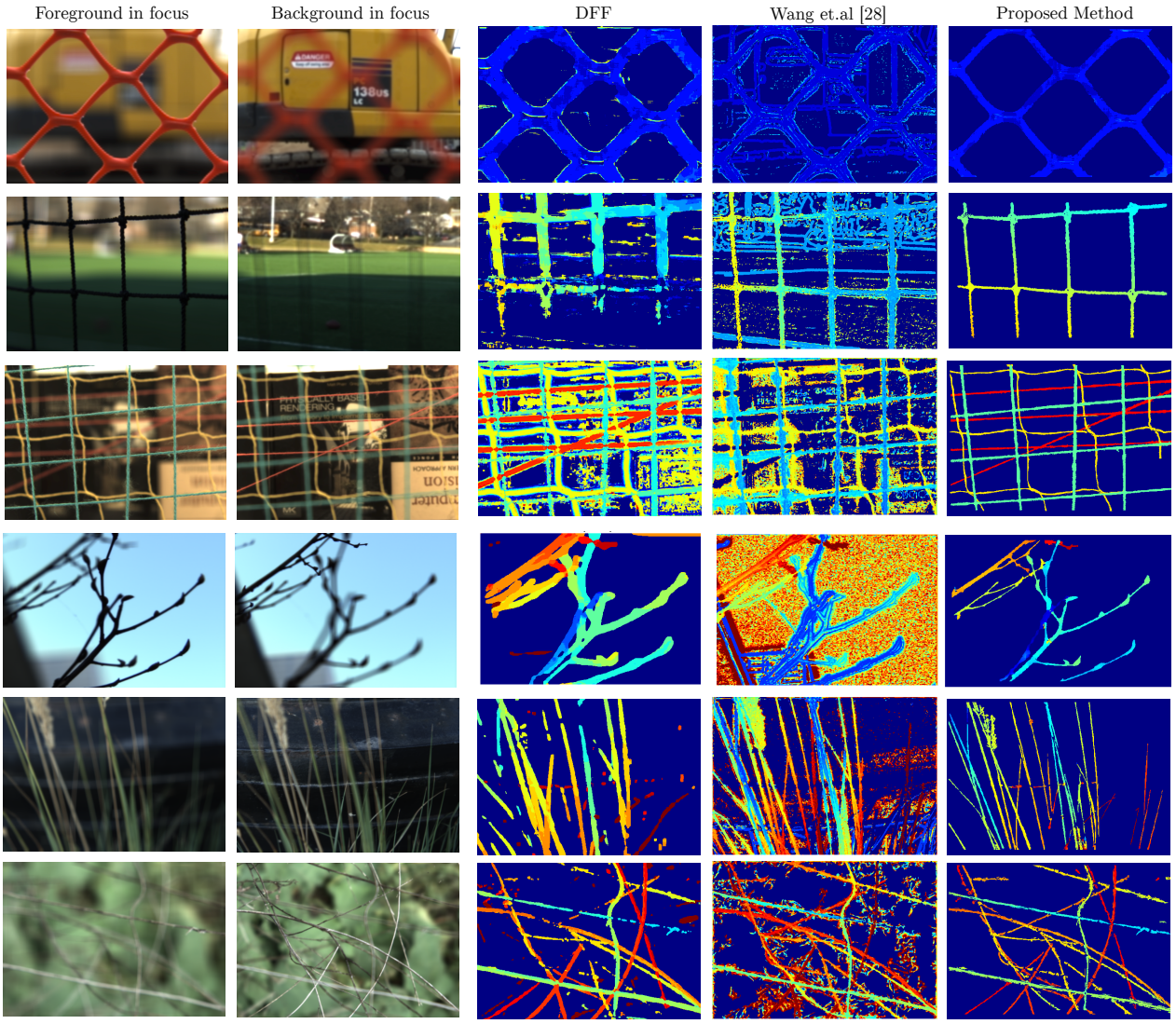
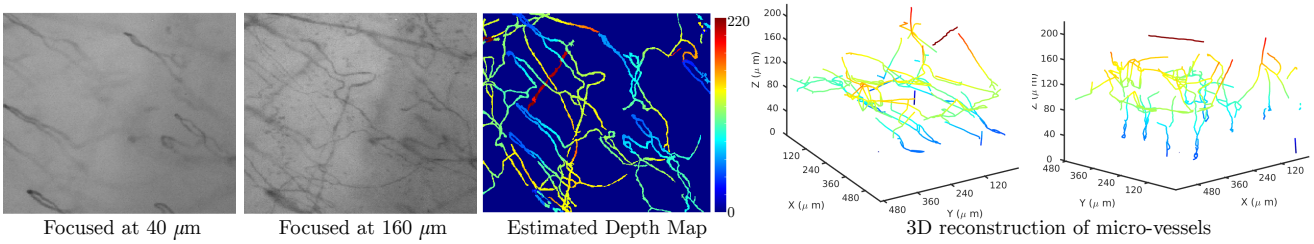


Figure 5: The depth recovery for the thin structures. Note that the depth estimations using the DFF method for points close to the occlusion boundaries are inaccurate due to high frequency depth discontinuity. The light field method in [28] does not perform well on the textureless regions and sharp edges in the background. Our method recovers the sharp depth discontinuity on the boundaries of the thin structures such as the grass and bush in the presence of spatially varying defocus blur.



Focused at 40 μm | Focused at 160 μm | Estimated Depth Map | 3D reconstruction of micro-vessels

Figure 6: The depth map and 3D reconstruction of micro-vessels. From left to right: 2 of 12 images in the focal stack; the estimated depth map, and two views of the reconstructed 3D structure. The 3D reconstruction is color coded to visualize the depth variations. To our knowledge, our method is the first approach to reconstruct the 3D structures of micro-vessels using non-invasive in-vivo image measurements.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–2282, Nov. 2012.

[2] B. Cubelos, A. Sebastin-Serrano, L. Beccari, M. E. Calcagnotto, E. Cisneros, S. Kim, A. Dopazo, M. Alvarez-Dolado, J. M. Redondo, P. Bovolenta, C. A. Walsh, and M. Nieto. Cux1 and cux2 regulate dendritic branching, spine morphology, and synapses of the upper layer neurons of the cortex. *Neuron*, 66(4):523 – 535, 2010.

[3] S. K. S. G. E. Alexander, Q. Guo and T. Zickler. Depth from focus with your mobile phone. In *ECCV*, October 2016.

[4] D. Eigen, D. Krishnan, and R. Fergus. Restoring an image taken through a window covered with dirt or rain. In *ICCV*, pages 633–640, 2013.

[5] J. B. Evers, A. R. van der Krol, J. Vos, and P. C. Struik. Understanding shoot branching by modelling form and function. *Trends in Plant Science*, 16(9):464 – 467, 2011.

[6] P. Favaro, S. Louis, and L. Angeles. Seeing Beyond Occlusions (and other marvels of a finite lens aperture). In *CVPR*, pages 1–8, 2003.

[7] J. Fiss, B. Curless, and R. Szeliski. Light Field Layer Matting. In *CVPR*, 2015.

[8] A. Fix, A. Gruber, E. Boros, and R. Zabih. A graph cut algorithm for higher-order markov random fields. In *ICCV*, pages 1020–1027, Nov 2011.

[9] J. Gu, R. Ramamoorthi, P. Belhumeur, and S. Nayar. Removing image artifacts due to dirty camera lenses and thin occluders. *ACM Transactions on Graphics (TOG)*, 2009.

[10] S. Hasinoff and K. Kutulakos. Multiple-Aperture Photography for High Dynamic Range and Post-Capture Refocusing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[11] S. W. Hasinoff and K. N. Kutulakos. A Layer-Based Restoration Framework for Variable-Aperture Photography. In *ICCV*, 2007.

[12] B. Huang, W. Wang, M. Bates, and X. Zhuang. Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. *Science*, 319(5864):810–813, 2008.

[13] H. Ishikawa. Transformation of general binary mrf minimization to the first-order case. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1234–1249, June 2011.

[14] V. Kolmogorov. A new look at reweighted message passing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(5):919–930, May 2015.

[15] A. Levin, D. Lischinski, and Y. Weiss. A closed form solution to natural image matting. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1*, CVPR '06, pages 61–68, Washington, DC, USA, 2006. IEEE Computer Society.

[16] Y. Liu, T. Belkina, J. H. Hays, and R. Lublinerman. Image De-fencing. In *CVPR*, pages 1–8, Jun 2008.

[17] S. McCloskey. Masking Light Fields to Remove Partial Occlusion. In *ICPR*, pages 2053–2058, 2014.

[18] M. McGuire, W. Matusik, H. Pfister, J. F. Hughes, and F. Durand. Defocus video matting. *ACM Trans. Graph.*, 24(3):567–576, 2005.

[19] R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan. Light Field Photography with a Hand-held Plenoptic Camera. pages 1–11, Apr 2005.

[20] D. Nordsletten, S. Blackett, M. Bentley, E. Ritman, and N. Smith. Structural morphology of renal vasculature. *American Journal of Physiology - Endocrinology and Metabolism*, 291(1), 2006.

[21] G. A. Strasser, J. S. Kaminker, and M. Tessier-Lavigne. Microarray analysis of retinal endothelial tip cells identifies cxcr4 as a mediator of tip cell morphology and branching. *Blood*, 115(24):5102–5110, 2010.

[22] S. Suwajanakorn, C. Hernandez, and S. M. Seitz. Depth from focus with your mobile phone. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[23] D. Tarlow, R. S. Zemel, and I. E. Givoni. HOP-MAP : Efficient Message Passing with High Order Potentials. *Journal of Machine Learning Research*, pages 812–819, 2010.

[24] M. Thomas, E. Laude, M. Moeller, J. Lellmann, and D. Cremers. SublabelAccurate Relaxation of Nonconvex Energies. In *CVPR*, number 1, pages 3948–3956, 2016.

[25] S. Trachsel, S. M. Kaeppler, K. M. Brown, and J. P. Lynch. Shovelomics: high throughput phenotyping of maize (zea mays l.) root architecture in the field. *Plant and Soil*, 341(1):75–87, 2011.

[26] V. Vaish, M. Levoy, R. Szeliski, C. L. Zitnick, and S. B. Kang. Reconstructing Occluded Surfaces Using Synthetic Apertures: Stereo, Focus and Robust Measures. In *CVPR*, pages 2331–2338, 2006.

[27] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.

[28] T.-c. Wang, A. A. Efros, and R. Ramamoorthi. Occlusion-aware Depth Estimation Using Light-field Cameras. In *ICCV*, 2015.

[29] T. Xue, M. Rubinstein, C. Liu, and W. T. Freeman. A computational approach for obstruction-free photography. *ACM Trans. Graph.*, 34(4):79:1–79:11, July 2015.

[30] S. Yamazaki, S. G. Narasimhan, S. Baker, and T. Kanade. The theory and practice of coplanar shadowgram imaging for acquiring visual hulls of intricate objects. *IJCV*, 81(3):259–280, 2009.