Joint image and depth completion in shape-from-focus: Taking a cue from parallax

Rajiv R. Sahay* and A. N. Rajagopalan

Image Processing and Computer Vision Lab, Department of Electrical Engineering, Indian Institute of Technology Madras, Chennai 600 036, India
**Corresponding author: sahayiitm@gmail.com*

Received September 14, 2009; revised February 8, 2010; accepted February 21, 2010; posted March 4, 2010 (Doc. ID 116856); published April 30, 2010

Shape-from-focus (SFF) uses a sequence of space-variantly defocused observations captured with relative motion between camera and scene. It assumes that there is no motion parallax in the frames. This is a restriction and constrains the working environment. Moreover, SFF cannot recover the structure information when there are missing data in the frames due to CCD sensor damage or unavoidable occlusions. The capability of filling-in plausible information in regions devoid of data is of critical importance in many applications. Images of 3D scenes captured by off-the-shelf cameras with relative motion commonly exhibit parallax-induced pixel motion. We demonstrate the interesting possibility of exploiting motion parallax cue in the images captured in SFF with a practical camera to jointly inpaint the focused image and depth map. © 2010 Optical Society of America

OCIS codes: 150.6910, 150.5670.

1. INTRODUCTION

In shape-from-focus (SFF) [1], a sequence of images is captured by moving a real-aperture camera relative to a 3D object. By measuring the degree of focus in the stack of space-variantly blurred observations, SFF arrives at the structure of the 3D scene. SFF is particularly well-suited for applications such as endoscopy [2] where spatial constraints preclude access to camera controls [required in methods such as depth from defocus (DFD)] and camera motion can be treated as piece-wise axial. The small working distances used in such applications demand techniques that work with real-aperture off-the-shelf cameras (OTS), and SFF clearly has advantages over other shape-extraction methods such as stereo, structure-from-motion, and shape-from-shading, which assume a pinhole camera.

SFF makes the critical assumption of "no pixel motion" in order to be able to apply a local window around a pixel and compute the focus measure profile by traversing through the stack. Researchers acknowledge [3–5] the fact that magnification will cause errors in depth estimation in methods that use the blur cue and have devised methods to avoid it. In the related area of DFD, telecentricity on the image-side has been proposed [3] to avoid magnification effects. Object-side telecentricity is also prevalent in the literature [6], but in order to achieve this effect the aperture of the front lens needs to be as large as the object to be viewed. This necessitates large, heavy, and expensive lenses. The assumption of "no parallax" is valid only for telecentric optics and not for OTS digital cameras in which pixel motion is a naturally occurring phenomenon. In one of our recent works [7], we have proposed a method to obtain the super-resolved focused image of 3D specimens in the SFF scenario using space-variantly defocused observations from a stack captured using an optical microscope with object-side telecentric lenses (i.e., no parallax effect). In this paper, we discuss depth and image inpainting in SFF using an OTS camera in the presence of missing data in observations and occlusions. We demonstrate that by exploiting parallax-induced pixel motion it is possible to fill-in missing regions due to simulated scratches on the CCD sensor and thin occluders.

Image inpainting assumes that the area to be inpainted is known *a priori* and uses information in the neighborhood to recover missing details. Applications include restoration of scanned old photographs, astronomical images, removal of text/subtitles in images as well as in videos, and also for disocclusion/removal of objects. Some of the early works on image inpainting can be found in [8–13]. Recent works have focused on automatic detection and inpainting of corrupted regions [14,15], and on joint super-resolution of video and inpainting [16]. Inpainting has also been used for shadow removal in [17,18].

The problem of occlusion-handling has been addressed in several works [19–23]. The authors in [19] estimate both the radiance and the shape of the occluding object as well as the background scene that is occluded. The property of a real-aperture camera is exploited to "see through" the occluder. However, they assume both the foreground and the background objects to be equifocal planes. The problem of self-occlusion is tackled in [20] again with the equifocal plane assumption. Levoy *et al.* [24,25] extend the idea of confocal microscopy and use an array of cameras/mirrors/projectors to see beyond occluders. In [21], high-resolution (HR) images are captured using an SLR camera, and depth maps for challenging data sets such as hair (a wig) are estimated. However, the method requires hundreds of HR images and yet suffers from unreliable depth estimates at low-textured regions,

where inpainting is resorted to by propagating boundary information using smoothness constraints [9] as a post-processing step. In the work of [22], three synchronized video streams are captured that share the same point-of-view but differ in their plane of focus. By over-constraining the problem, an improved alpha-matte is estimated that reveals the occlusion effect. The work in [23] models occlusions at defocused layer boundaries using a layered image formation model. However, the scene is assumed to consist of multiple, overlapping, fronto-parallel layers, which is a simplification. Only the camera aperture is changed while capturing different observations and there is no relative motion between the sensor and the scene. Importantly, the work in [23] does not perform disocclusion *per se* but accounts for occlusion effects only at defocused boundaries. We remark that all the above works [19–23] have been proposed in the DFD framework, where camera parameters such as focal length, image-plane to lens distance, and aperture are changed to obtain defocused observations. It can be theoretically shown that the observations captured by relative motion between the camera and the 3D object in SFF and those obtained by changing camera parameters, as in DFD, are not identical.

We observe that in all the above works [8–23], there is no motion parallax in the observations. In the literature, image/video inpainting has primarily been attempted using focused/defocused images without any magnification or parallax. In this work, we model the formation of the stack of space-variantly blurred observations affected by motion parallax and missing data in the SFF scenario. We explicitly relate the parallax-induced pixel motion to the structure of the 3D object. The space-variant defocusing is also related to the depth of the points on the specimen from the real-aperture camera. We seek to exploit the intricate coupling of parallax and defocus with the shape of the object. We model the inpainted focused image and the completed depth profile using independent Markov random fields (MRF) and obtain their maximum *a posteriori* (MAP) estimates in an integrated manner. SFF methods typically avoid modeling the point spread function (PSF) of the camera. However, because of the complexity of the challenges tackled in this paper, we are constrained to use a model for the PSF. All approximations made are validated with real data.

The works in [16,26] are the closest in comparison to our approach. However, there are several differences, too. The work of [16] assumes a planar scene and performs image inpainting while obtaining a super-resolved video sequence from low-resolution frames. The captured low-resolution images are shifted with respect to one another globally, and there is no motion parallax in the frames. In a recent work [26], the authors perform inpainting in a stereo setup using only pinhole images. By virtue of using two viewpoints, they have access to regions that are missing/occluded in one of the views, which enables them to inpaint the focused image and the depth map. Unlike in [26], we do not need a pre-calculated depth map, nor are we constrained to work with pinhole images. To the best of our knowledge, ours is the first work of its kind in SFF to use *parallax and defocus cues jointly* for handling occlusions and sensor damage. We would like to emphasize that our aim here is not to account for motion parallax but rather to use it judiciously as a cue for inpainting image and the structure.

In Section 2, we discuss the issue of missing data in SFF. Pixel motion due to parallax effect is treated in Section 3. The degradation model for missing observations is explained in Section 4. The proposed method for joint image and depth inpainting is given in Section 5. Experimental results are described in Section 6. Concluding remarks are presented in Section 7.

2. MISSING DATA PROBLEM

In this section, we illustrate the effect of missing data in SFF. Let us consider that the 3D object is occluded by another object that is placed in the foreground. The occluder is assumed to be static and it is only the 3D object that moves.

Let us initially assume that there is no magnification (parallax) in the observations (assumption not valid for OTS cameras). We use the term "magnification" or "scaling" loosely to refer to parallax-induced pixel motion. Therefore, data are missing at some fixed spatial locations in the captured frames. A similar situation could arise if some portion on the CCD sensor was corrupted. We evoke a synthetic experiment, where there is no pixel motion, and show that depth and focused image (in the missing regions) cannot be recovered faithfully in SFF. We choose a simple ramp-shaped 3D specimen with "calf" [27] texture mapped onto its surface. We simulate the motion of the object along the optical axis to obtain a sequence of images with missing data in some parts of the images. Fifty frames are captured in steps of $\Delta d = 1$ mm. In Figs. 1(a) and 1(b), two of the frames are shown. The missing regions due to the damaged sensor or occlusions are shown with black pixels (whose locations are chosen arbitrarily). The reconstructed depth map obtained using the SFF technique [1] is shown as a grayscale image in Fig. 1(c). The estimated depth at regions where image data are missing in the observations is quite poor. We also reconstruct the focused image of the underlying specimen by selecting the particular frame in which a pixel comes into focus and by picking the grayscale intensity at that location. By repeating this procedure for all the pixels, we can construct an (approximate) all-in-focus image, which is shown in Fig. 1(d). We observe that with such a procedure, it is impossible to fill-in the missing regions in the focused image.

A. Motion-Cue for Data Completion

For OTS cameras, parallax results in apparent motion of image features. The center of projection of the rays coming from the scene changes every time the object is moved along the optical axis in finite steps. As an example, let us consider Fig. 2, where we have shown images from a stack captured by a *real* OTS camera. To simulate the effect of a damaged sensor we have marked some of the pixels black. The 3D specimen is a small wooden piece on which a face has been carved. The captured images are both space-variantly blurred and scaled. The location of image features shifts as we travel from one frame to another [shown by white arrows in Fig. 2(a)]. Note that the

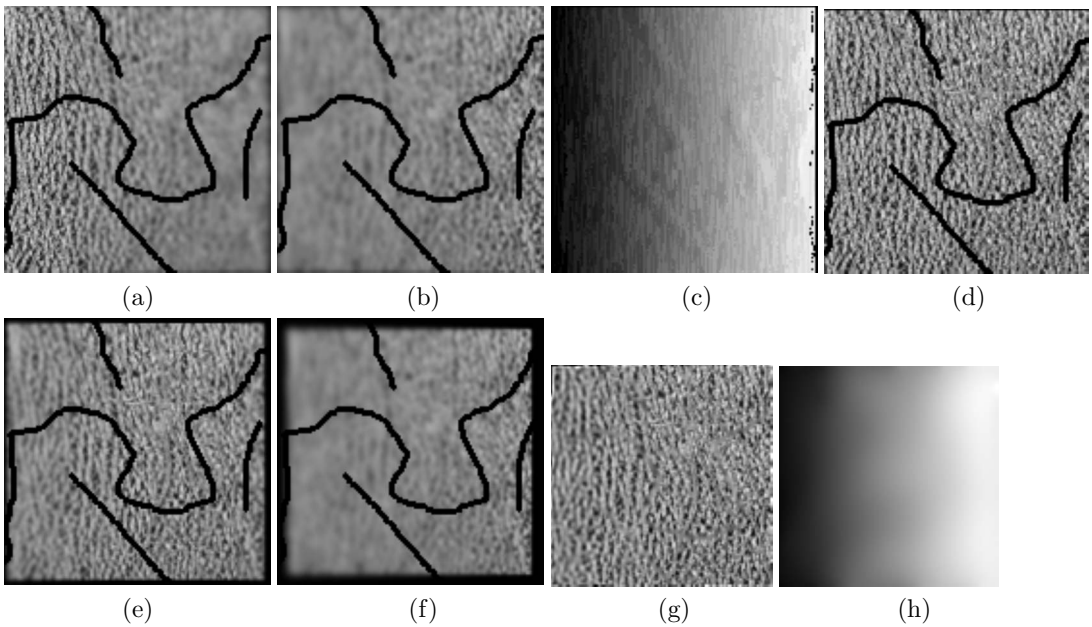

Fig. 1. (a,b) Frames 1 and 40 (unscaled stack). (c,d) Reconstructed depth profile and focused image using [1]. (e,f) Frames 20 and 40 from scaled stack. (g,h) Inpainted focused image and completed depth profile, respectively, using our method.

position of the damaged regions is fixed but different image features are covered by the black pixels in a different image [Fig. 2(b)]. We shall subsequently exploit this cue for inpainting both focused image and structure.

3. INPAINTING IN SFF

The basic principle of SFF is depicted in Fig. 3(a). The object is initially placed such that the translating stage is on the focused plane. The 3D object is translated downwards (along the optical axis in fixed finite steps of Δd) such that at every step a space-variantly blurred image is captured. As the point (k,l) approaches the focused plane it gradually comes into focus. The quantity $\bar{d}(k,l)$ is the amount by which the stage should be translated to bring the point (k,l) to the focused plane (at a distance of w_d from the lens plane), when it is in perfect focus (neglecting diffraction effects and aberrations). The basic idea in SFF is to estimate for every point the position of best focus, $\bar{d}(k,l)$. The variable $\bar{\mathbf{d}}=\bar{d}(k,l)$ for all the points (k,l) on the 3D specimen characterizes its shape. To detect the position of

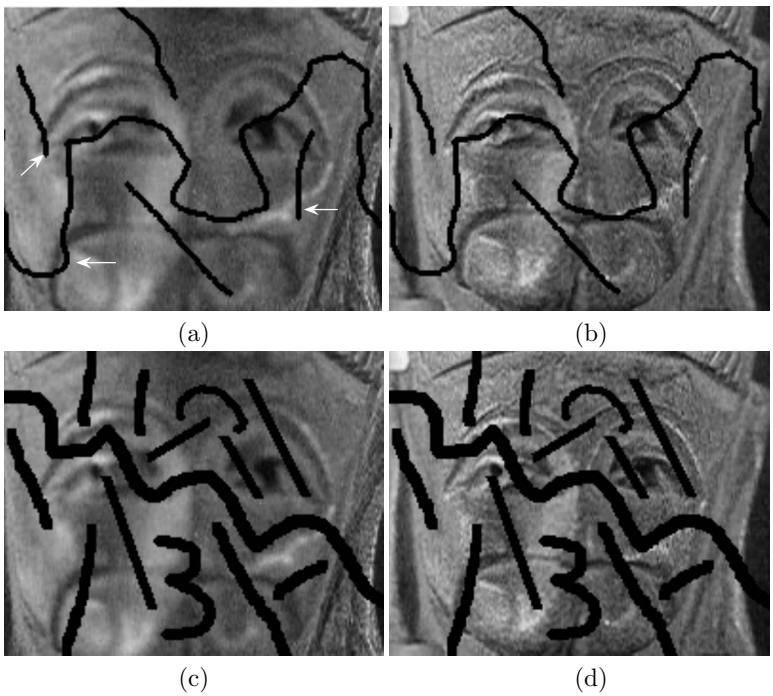

Fig. 2. Wooden face specimen. (a,b) Second and eighth frames, respectively. (c,d) Corresponding frames with thicker scratches.

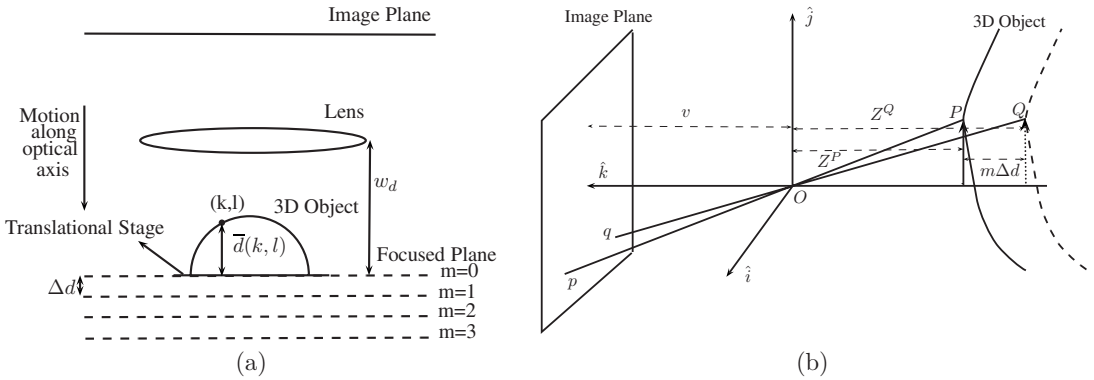

Fig. 3. (a) Working principle of SFF. (b) Schematic showing mechanism of structure-dependent pixel motion in SFF.

best focus for a point on the 3D object a focus measure operator is used to measure the quality of image focus in a local region. Specifically, the sum-modified Laplacian (SML) operator [1] is used to compute a focus measure profile for each pixel in the frame. The estimate of the shape of the object is obtained by Gaussian interpolation.

The above formulation assumes that the pixel locations do not change between frames. This is crucial since the window used for computing the SML is centered around the same point in all the frames. In real-world situations, data in portions of the observations may be missing due to several factors including damage to the CCD sensor or occlusions. Since the SFF analysis is local, and depth estimates for a particular point are computed using data in the observations in the immediate neighborhood of that pixel, this technique cannot recover structure in missing regions [as shown in Figs. 1(c) and 1(d)]. Interestingly, we show next that when there is magnification in the stack, structure-dependent pixel motion can be tapped to perform inpainting of both the depth profile as well as the focused image of the 3D specimen.

A. Motion Parallax in the Stack

When the relative motion between scene and camera is significant, there will be magnification (due to motion parallax), and image features will shift spatially from frame to frame. Let the 3D specimen move relative to the camera along the axial direction as a sequence of frames is captured. Initially, we consider a pinhole camera. As shown in Fig. 3(b), we examine a specific point on the specimen that is moved relative to the pinhole camera. A point on the 3D object with world coordinates $P(X^P, Y^P, Z^P)$ moves to $Q(X^Q, Y^Q, Z^Q)$ along the Z-axis by a distance of $m\Delta d$ and away from the pinhole denoted by O. The distances of the points P and Q from the pinhole are Z^P and Z^Q, respectively. The point P is imaged at p on the image plane and has coordinates (x,y). Let this image be the reference plane. When the 3D object is moved away from the pinhole by an amount $m\Delta d$, the point Q is imaged at q with coordinates (x',y') on the image plane.

According to basic perspective projection equations,

$$x = \frac{vX^P}{Z^P}, \quad x' = \frac{vX^Q}{Z^Q}, \quad \text{and} \quad y = \frac{vY^P}{Z^P}, \quad y' = \frac{vY^Q}{Z^Q}, \tag{1}$$

where v is the distance between the pinhole and the image plane. The motion of the object relative to the pinhole is along the Z-axis only, since the 3D specimen is translated away from or towards the camera along the optical axis. Hence, for the SFF scenario, we have $X^P = X^Q$, $Y^P = Y^Q$, $Z^Q = Z^P + m\Delta d = w_d - \bar{d} + m\Delta d$, and $-M/2 \leq x', y' \leq M/2$. Here, $M \times M$ is image size. Thus, it can be shown that

$$x' = \frac{x(w_d - \bar{d})}{(w_d - \bar{d}) + m\Delta d}, \quad y' = \frac{y(w_d - \bar{d})}{(w_d - \bar{d}) + m\Delta d}. \tag{2}$$

Note that the pixel motion is a function of $\bar{d}$, the 3D structure of the scene.

4. DEGRADATION MODEL

Consider N frames, $\{y_m(i,j)\}$, $m = 0, 1, \ldots, N-1$ from the stack. Assume that these are derived from a single focused image $\{x(i,j)\}$ of the 3D specimen. The scaled and defocused frames can be related to the focused image by the degradation model

$$\mathbf{y}_m^{\text{vis}} = \mathbf{O}_m[\mathbf{H}_m(\bar{\mathbf{d}})\mathbf{W}_m(\bar{\mathbf{d}})\mathbf{x} + \mathbf{n}_m], \qquad m = 0, \ldots, N-1, \tag{3}$$

where $\mathbf{y}_m^{\text{vis}}$ is the lexicographically arranged vector of size $M^2 \times 1$ derived from the visible regions in the mth defocused and scaled frame, $\mathbf{W}_m$ is the matrix describing the motion of the pixels in the mth frame, $\mathbf{H}_m$ is the blurring matrix for the mth frame, $\mathbf{n}_m$ is the $M^2 \times 1$ Gaussian noise vector in the mth observation, and $\mathbf{O}_m$ is the operator that removes the missing/damaged regions and crops out the visible portions of the observations. Note that for both the cases of sensor damage or occlusions (due to a static occluder), the cropping operator is unchanged across the stack, i.e., $\mathbf{O}_m = \mathbf{O}$. Hence, the spatial locations of the inpainting region remain the same in all the frames.

Observe the interesting relationship between the shape of the object, pixel-motion, and space-variant defocusing in the frames. The degree of space-variant defocus blur induced at each point in the image of a 3D scene is dependent upon the depth of the object ($\bar{\mathbf{d}}$) from the lens plane. Also, the pixel motion is a function of the 3D structure of the object. In fact, the twin cues of defocus and motion parallax are intertwined with the 3D structure and must be exploited.

In the previous section, we elucidated the motion parallax effect that is modeled by $\mathbf{W}_m(\bar{\mathbf{d}})$ in Eq. (3). We now describe the space-variant blurring mechanism as specified by $\mathbf{H}_m(\bar{\mathbf{d}})$ in Eq.(3). In its most general form, the image formation process [as described by Eq. (3)] is difficult to comprehend, and some simplifying assumptions must be made to gain a handle on the problem. To this end, we assume a parametric model for the PSF of the camera. Because of diffraction and lens aberrations, the PSF is best described by a circularly symmetric 2D Gaussian function [28] with standard deviation $\sigma=\rho r_b$, where ρ is a camera constant and r_b is the blur circle radius. There exist several works that have validated this approximation [19,29]; hence, we too are motivated to use this model. Since the blur parameter σ is a function of depth [29], the blurring induced is space-variant.

When the stage is moved downwards in steps of Δd [Fig. 3(a)], for the mth frame the blur parameter for a 3D point whose image coordinates are (k,l) is given as

$$\sigma_m(k,l) = \rho R v\left(\frac{1}{w_d} - \frac{1}{w_d + m\Delta d - \bar{d}(k,l)}\right), \tag{4}$$

where $1/w_d=(1/f)-(1/v)$, f is the focal length, R is the radius of the aperture of the lens, and v is the distance between the lens and the image plane.

The value of the camera intrinsic parameter ρRv in Eq. (4) must be deduced by camera calibration but needs to be determined only once for a given camera. To this end, we conducted the following experiment. A focused image I_f of a planar object (with an appreciable amount of texture) was first acquired by positioning it parallel to the lens plane. Next, we displaced this sample by a known distance along the optical axis to obtain a space-invariantly blurred image I_b. The originally captured focused image I_f was blurred using a Gaussian kernel for different values of ρRv incremented in small steps. That value of ρRv that minimized the mean-squared error between the "blurred" focused image and the captured observation was taken to be the estimate of ρRv. Note that since the object chosen in the calibration experiment was planar, there are no parallax effects when the object is translated.

We next analyze the degree of occlusion that can be handled using parallax and defocus cues. Let h_x and h_y denote the maximum widths of the scratch along X and Y directions, respectively. Let $x_{\min}=\min_{x\in S}(|x|)$ and $y_{\min}=\min_{y\in S}(|y|)$, where S is the set of pixels to be inpainted. From Eq. (2), it can be shown that if

$$h_x \leq \max_{0\leq m\leq N-1}\left|\frac{x_{\min} m\Delta d}{w_d - \bar{d}_{\min} + m\Delta d}\right|, \quad \text{or}$$

$$h_y \leq \max_{0\leq m\leq N-1}\left|\frac{y_{\min} m\Delta d}{w_d - \bar{d}_{\min} + m\Delta d}\right|, \tag{5}$$

then every pixel in S will be rendered visible in at least one of the N observations due to the parallax effect. Here, $\bar{d}_{\min}=\min_{x,y\in S}\bar{\mathbf{d}}(x,y)$. The bounds in Eq. (5) are derived for the worst-case situation, when the corresponding pixel in S is assumed to be farthest from the lens among all the points in S. While inpainting due to motion cue is direct, the defocus cue contributes indirectly by affecting observation values outside of S. Since the Gaussian blur is symmetric, this yields the condition for inpainting as

$$h_x \text{ or } h_y \leq \min\Big(\max_{x,y\in S}\ \max_{0\leq m\leq N-1}(6\sigma_m(x,y)+1)\Big). \tag{6}$$

Note that 99% of the area under a Gaussian is within $\pm 3\sigma$ from its mean value. We remark that near the image center, pixel motion is absent and the method relies only on the defocus cue for inpainting these locations.

5. PROPOSED FRAMEWORK FOR INPAINTING/DISOCCLUSION

The problem that we attempt to solve is the following: Given a stack of scaled and defocused observations, where there are regions of missing data due to possible occlusions in the scene or due to faulty camera sensor, how do we simultaneously inpaint the depth profile and the focused image of the object?

We propose to inpaint both the depth profile as well as the focused image of the object within a unified framework. Note that for the problem scenario considered in this work, the spatial locations of the missing regions do not change from frame to frame in the stack. It is highly probable that whatever region of the 3D specimen was occluded/missing in one observation of the stack will come into view in another frame, since there is magnification in the stack due to parallax effect.

Simultaneous reconstruction of depth profile $\bar{\mathbf{d}}$ and focused image $\mathbf{x}$ is an ill-posed inverse problem, and hence, the solution must be regularized using *a priori* constraints. Real-world objects usually have depth profiles that are locally smooth. The same argument holds good for the focused image also. MRFs have the capability to model spatial dependencies so as to incorporate prior information [30]. We model the structure and the focused image of the 3D specimen using separate Gauss MRF (GMRF) with a first-order neighborhood. We use the GMRF model for its simplicity. The Hammerseley–Clifford theorem [31] provides the all-important equivalence between Gibbs random field and MRF, enabling the specification of the prior joint probability density fuction for the depth map and the focused image. For details on MRF, see [30].

We seek to solve for the MAP estimate of $\bar{\mathbf{d}}$ and $\mathbf{x}$. Let us consider a set of p frames chosen from the stack of N observations. Assuming noise processes $\mathbf{n}_m's$ to be independent, the MAP estimates of $\bar{\mathbf{d}}$ and $\mathbf{x}$ can be obtained by minimizing

$$U^p(\bar{\mathbf{d}},\mathbf{x}) = \sum_{m\in O}\frac{\|\mathbf{y}_m^{\text{vis}} - \mathbf{O}_m[\mathbf{H}_m(\bar{\mathbf{d}})\mathbf{W}_m(\bar{\mathbf{d}})\mathbf{x}]\|^2}{2\sigma_\eta^2} + \lambda_{\bar{\mathbf{d}}}\sum_{c\in C_{\bar{\mathbf{d}}}} V_c^{\bar{\mathbf{d}}}(\bar{\mathbf{d}}) + \lambda_{\mathbf{x}}\sum_{c\in C_{\mathbf{x}}} V_c^{\mathbf{x}}(\mathbf{x}), \tag{7}$$

where $O=\{u_1,u_2,\ldots,u_p\}$, u_i is the frame number, and σ_η^2 is the variance of the Gaussian noise. The clique potential function for $\bar{\mathbf{d}}$ is

Table 1. Errors for Synthetic and Real Specimens

Specimen	RMSE for Image Inpainting (in Gray Levels)	RMSE for Depth Completion (as % of Maximum $\bar{d}$)
	$k_s=0,1,2,3$	$k_s=0,1,2,3$
$Ramp_1$(calf)	14, 15, 15.5, 14.5	2.8%, 2.96%, 3%, 3.2%
$Ramp_2$(random-dot)	4.5, 5, 5.53, 4.8	1.0678%, 1.074%, 1.07%, 1.069%
$Ramp_3$(bark)	5.66, 7.64, 7.55, 7.52	2.1%, 2.1%, 2.113%, 2.112%
$Ramp_4$(straw)	10.37, 13.13, 13.01, 13.88	2.25%, 2.27%, 2.28%, 2.29%
$Ramp_5$(brick wall)	3.02, 4.4, 4.25, 4.75	3.055%, 3.061%, 3.056%, 3.06%
	RMSE for Depth Completion (as % of Maximum $\bar{d}$) $k_s=0,1,2,3$	
	$Cylinder_1$(1 cm) 7%, 7.5%, 7.479%, 7.207%	
	$Cylinder_2$(1.3 cm) 8.22%, 8.25%, 8.28%, 8.26%	
	$Cylinder_3$(1.6 cm) 5.98%, 6.019%, 6.077%, 6.022%	

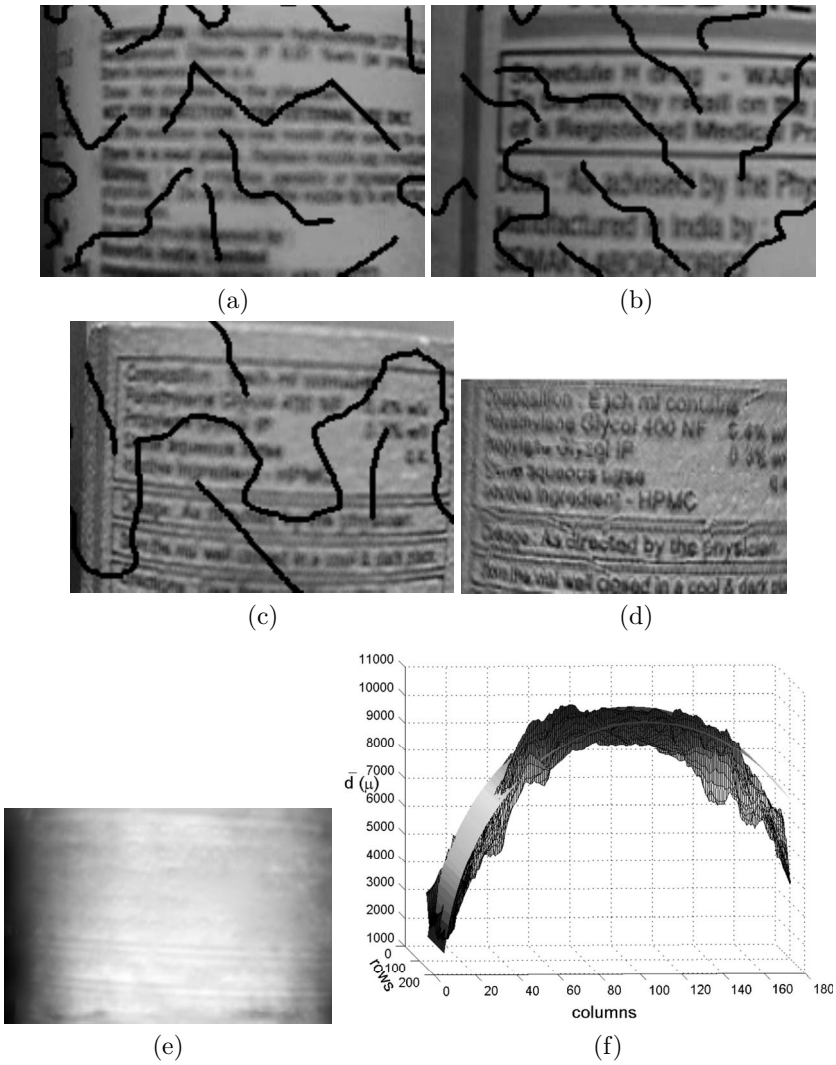

Fig. 4. (a,b,c) Observations corresponding to the second frame for three different specimens. (d) Inpainted focused image corresponding to specimen (c). (e) Completed depth map. (f) Cylindrical fit to the estimated depth profile.

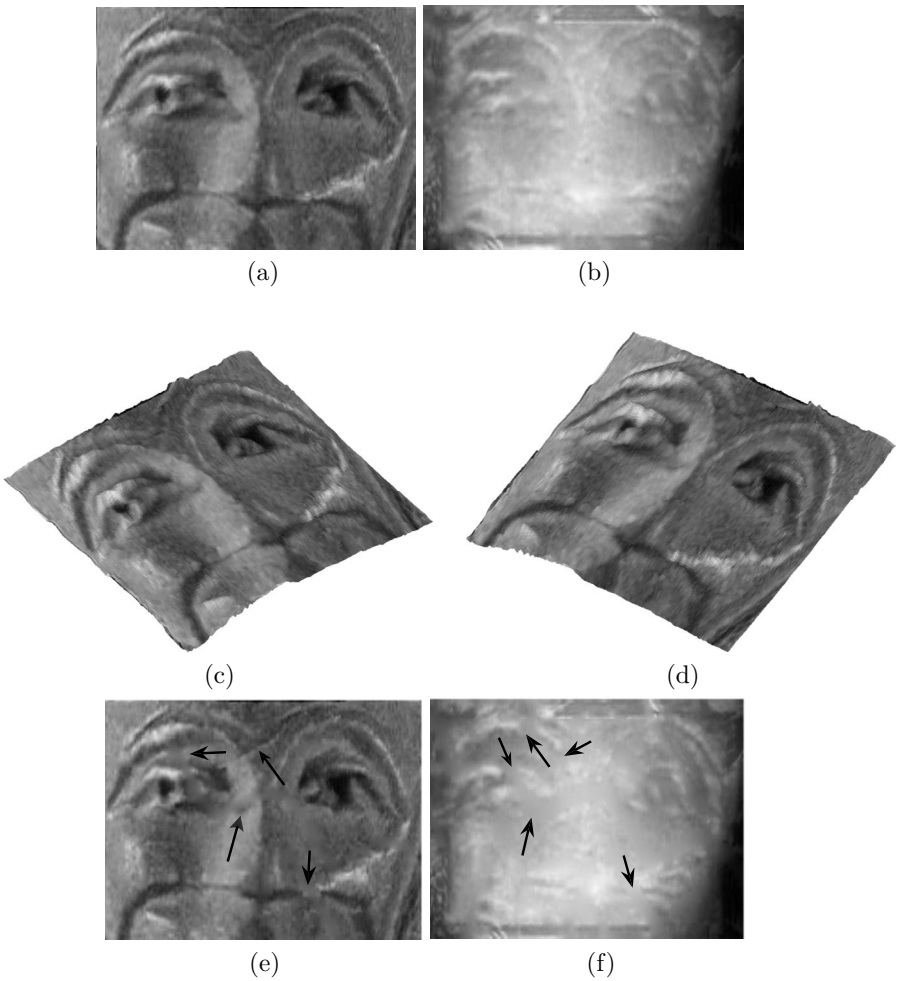

Fig. 5. (a) Inpainted focused image. (b) Inpainted depth map. (c,d) Novel view depiction. (e,f) Inpainted focused image and depth map, respectively, obtained using the observations in Figs. 2(c) and 2(d).

$$\sum_{c\in C_{\bar{\mathbf{d}}}} V_c^{\bar{\mathbf{d}}}(\bar{\mathbf{d}}) = \sum_{i=1}^{M}\sum_{j=1}^{M}[(\bar{d}(i,j)-\bar{d}(i,j-1))^2+(\bar{d}(i,j+1)-\bar{d}(i,j))^2+(\bar{d}(i+1,j)-\bar{d}(i,j))^2+(\bar{d}(i,j)-\bar{d}(i-1,j))^2], \quad (8)$$

where c is a clique, C is the set of all cliques, and $V_c(\cdot)$ is the potential associated with clique c. We choose $V_c^{\mathbf{x}}(\mathbf{x})$ to be of the same form as $V_c^{\bar{\mathbf{d}}}(\bar{\mathbf{d}})$.

In recent literature, graph cuts have been used as a fast optimization technique for submodular energy functions [32,33]. In applications that involve blur [34–36] the energy functions are non-submodular and graph cuts have not performed well. Roof duality works in cases where the number of non-submodular terms is small [35]. However, for a 5×5 sized kernel, it is shown in [35] that the number of unassigned labels for the quadratic pseudo boolean optimization (QPBO) and the "probing" QPBO (QPBOP) methods is a whopping 80%. In contrast, simulated annealing (SA) is shown to outperform all the methods, including QPBO and QPBOP [35]. The energy at convergence is zero and all the nodes are labeled. Hence, we have used the SA algorithm to minimize $U^p(\bar{\mathbf{d}},\mathbf{x})$. Parameters $\lambda_{\bar{\mathbf{d}}}$ and $\lambda_{\mathbf{x}}$ must be carefully chosen to obtain a good estimate of both $\bar{\mathbf{d}}$ and $\mathbf{x}$.

6. EXPERIMENTAL RESULTS

In all our experiments, we feed the entire SFF stack as input to the method in [1] to obtain an initial estimate of the inpainted depth map. During the inpainting process, following literature [9], we assume knowledge of the locations of the regions that have to be filled-in. For the inpainted focused image, we choose an arbitrary initial estimate (cropped portion of the Lena image). We use four frames with good relative blur among them from the stack to reconstruct the completed focused image and the depth map in all the experiments. To avoid boundary problems, we operate over an inner region of the observations. Hence, the inpainted depth profile and the focused image obtained from our method are slightly smaller in size compared to the observations.

Initially, we present results for a synthetic experiment corresponding to a ramp-shaped depth map. We consider different heights and textures for the depth map. The heights of the ramps $\{\text{ramp}_i, i=1,2,3,4,5\}$ are 3.0 cm,

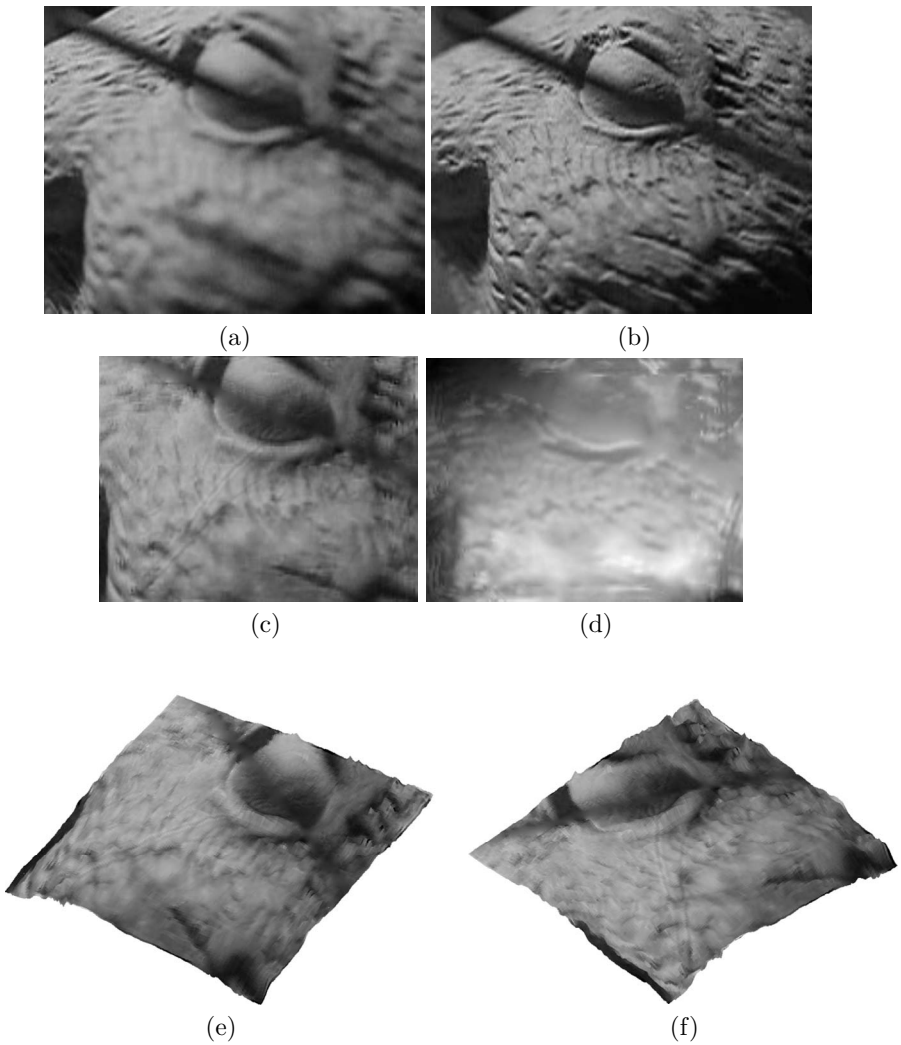

Fig. 6. (a,b) Observations of a clay model of a bunny occluded by a pin. (c) Inpainted focused image. (d) Estimated shape profile. (e, f) Novel views of the bunny.

3.5 cm, 4.0 cm, 4.5 cm, and 5.0 cm, respectively. The corresponding textures were chosen as "calf" [27], a random-dot pattern, "bark," "straw," "brick wall" (last three from [27]), respectively. We synthetically generate scaled and defocused stacks of fifty frames corresponding to each of these cases by simulating the motion of the object away from the camera along the optical axis in finite steps of $\Delta d = 1$ mm with the effect of parallax factored in. To simulate CCD sensor damage, we scratched the observations by three different scratch patterns (denoted by $k_s = 1, 2, 3$ in Table 1, with $k_s = 0$ representing the no-scratch case).

Four observations from each scaled stack were given as input to our method. For each of the synthetic experiments, we choose MRF parameters as $\lambda_{\bar{\mathbf{d}}} = 10^9$ and $\lambda_{\mathbf{x}} = 0.005$. As a representative example, the 20th and the 40th observations of the first ramp-shaped specimen (with scratch pattern corresponding to $k_s = 1$) are shown in Figs. 1(e) and 1(f), respectively. Our method fills up both the image and the depth map completely, as shown in Figs. 1(g) and 1(h) {unlike the unscaled case discussed earlier in Section 2 [Figs. 1(c) and 1(d)]}. The structure information is recovered even at locations where the data were missing. Our algorithm was tested on all the fifteen different cases (corresponding to five different ramp heights each scratched by three different patterns), and the rms errors in image and depth inpainting are reported in Table 1. The average rms error for the inpainted structure over all the fifteen samples is less than 3.3% of the maximum height of the ramp in each case. The average rms error incurred in image inpainting is less than 15 gray levels. In Table 1, we also report the rms error for the case when there are no scratches (i.e., for $k_s = 0$ and for each specimen). The average rms error incurred for the "no scratch" case (for both structure and image inpainting) is only marginally lower compared to the case when scratches are present. This amply demonstrates the ability of our method to effectively fill-in missing information.

Next, we captured real-world objects using an OTS Olympus C5050Z digital camera operating in the "super-macro" mode. To enable quantitative evaluation of accuracy, we tested with different bottles of ophthalmic medicine $\{\text{cylinder}_i, i = 1, 2, 3\}$. The corresponding radii were 1 cm, 1.3 cm, and 1.6 cm, respectively. Each bottle had its own wrapper containing written text. Synthetically generated cylindrical surfaces were fit and compared with the reconstructed depth maps. For general 3D objects

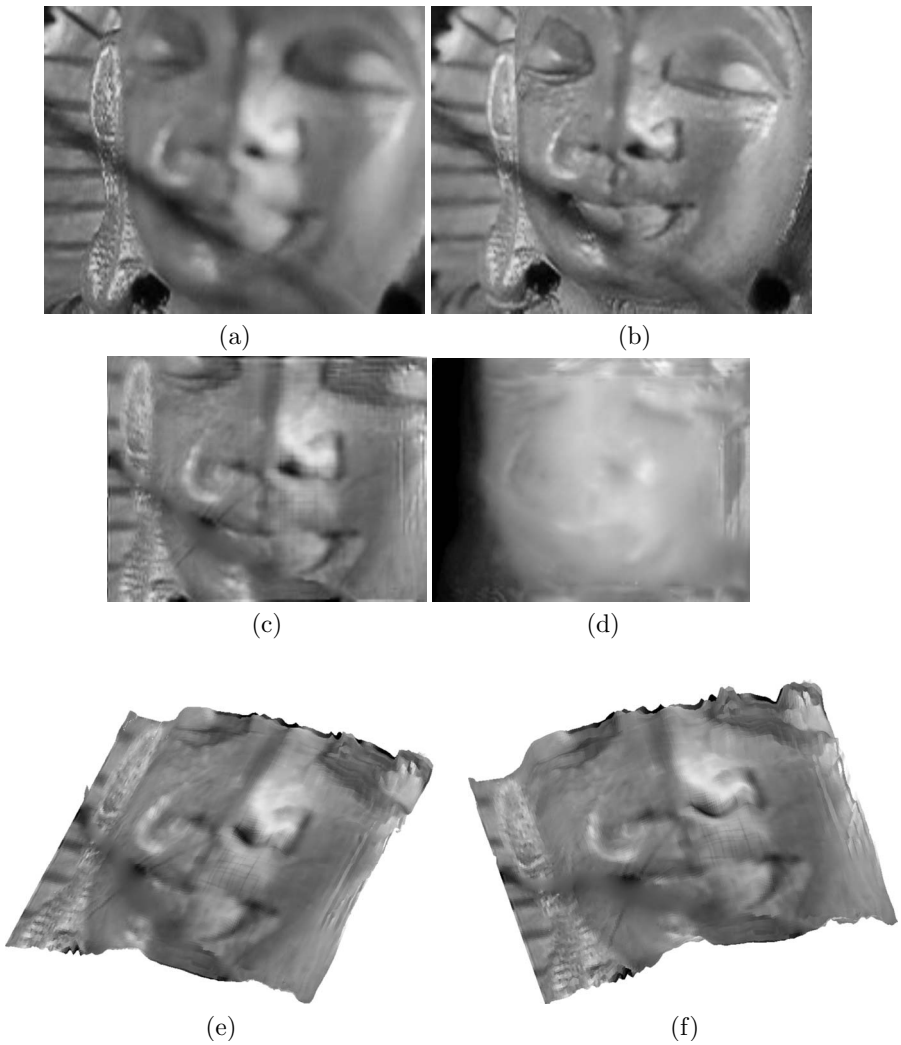

Fig. 7. Wooden Buddha statue occluded by a pin. (a, b) Frames 2 and 8. (c,d) Inpainted focused image and shape profile. (e,f) Novel views after texture mapping.

with arbitrary shapes, we provide novel views for qualitative evaluation.

For capturing defocused and scaled stacks, the digital camera was kept static, and the 3D specimen was placed on a translating stage that was moved away from the camera in fixed finite steps of $\Delta d = 1$ mm. A video was captured for the entire motion. The frames were then extracted from this video and processed. To simulate the effect of missing data due to possible damage to the CCD sensor, we randomly scratched the frames post-capture with the scratch patterns corresponding to $k_s = 1, 2, 3$ (discussed earlier). Sample observations corresponding to different cylindrical specimens are shown in Figs. 4(a)–4(c). The proposed algorithm was used to obtain the inpainted focused image and the completed depth profile in each case. For the real-world specimen, the values of the MRF parameters were chosen as $\lambda_{\bar{\mathbf{d}}} = 10^8$ and $\lambda_{\mathbf{x}} = 0.05$, respectively. Because of space constraints, we show the results only for the first object degraded by the first scratch pattern [Fig. 4(c)]. The inpainted focused image is shown in Fig. 4(d). Note that, after inpainting with our method, the text below the scratched regions in the observations becomes visible. The completed depth profile represented as a grayscale image is shown in Fig. 4(e). Note that the depth profile in the scratched region has been completely filled-in. The actual radius of the 3D object is 1 cm. The maximum height of the estimated depth profile is also approximately 1 cm which is in accordance with the physically measured dimension of the bottle. In Fig. 4(f), we depict together the plot of the estimated structure and the synthetically generated cylindrical surface. Note that the estimated depth profile closely follows the cylindrical surface, as expected. The size of the estimated depth profile and the fitted cylindrical surface is 112×169 pixels.

In Table 1, we give a summary of the rms errors for nine different cases (corresponding to three different bottles each scratched by three different patterns). The average rms error in depth completion is less than 8.3%. The rms errors for the "no scratch" case ($k_s = 0$) in Table 1 are again only marginally lower, thus validating the effectiveness of inpainting using our method. For real specimens, we do not provide rms errors for image inpainting since the original texture is not available.

Next, we return to the face example discussed in Subsection 2.A. Two of the four observations used by the proposed method are as given in Figs. 2(a) and 2(b). The in-

painted image and the reconstructed depth map using the proposed algorithm are given in Figs. 5(a) and 5(b), respectively. It is interesting to note that various features of the face, such as the eyes, the eyebrows, the nose, and the straight edge below it are faithfully reconstructed both in the inpainted image and the inpainted depth map. We also show two novel views [Figs. 5(c) and 5(d)] of the estimated inpainted structure after texture mapping. Note that there are no artifacts.

We also investigated a failure case corresponding to severe sensor damage. Two representative observations are shown in Figs. 2(c) and 2(d). The quality of the estimated focused image [Fig. 5(e)] and the depth profile [Fig. 5(f)] is degraded (see points indicated by arrows). Because of the thick nature of the scratches, the motion of pixels is insufficient to render visible (in other observations) the portions hidden behind the damaged regions.

To demonstrate the capability of our method in performing disocclusion, we present results for yet another real experiment. We attempt to demonstrate the usefulness of the proposed method in potential applications such as endoscopy, where fine body structures may occlude the object of interest. A small pin is kept across the field-of-view of the camera such that it occludes the specimen, which is chosen as a clay model of a bunny. Two frames from the stack are shown in Figs. 6(a) and 6(b). Note that the position of the occluded regions is fixed in all the observations and is assumed to be known. The proposed method yields the inpainted focused image shown in Fig. 6(c). The features of the eye portion at the locations of the occluder have been recovered well. The edges at the junction of the eyelids and the eyeball which were occluded are now clearly visible. Also, the texture on the surface of the specimen has been deblurred well and can be discerned easily. The estimated depth profile is shown in Fig. 6(d). Even in occluded regions, the algorithm is able to recover the depth information. Since the occlusion removal is accurate, the novel views [shown in Figs. 6(e) and 6(f)] do not exhibit any artifacts.

We present one more set of real results for disocclusion. In this case, a portion of a statue of Buddha was used for imaging. The object was occluded again by a pin and a stack of scaled and defocused frames was captured. Among the four chosen observations, the second and the eighth frames are shown in Figs. 7(a) and 7(b), respectively. Using these frames, the focused image of the specimen was estimated and this is shown in Fig. 7(c). Observe that the portion of the lower lip that was occluded in the observations has been recovered well. The occluder has almost completely been removed in the estimated focused image. The depth profile was also reconstructed and is depicted in Fig. 7(d). The proposed algorithm has estimated the depth map correctly despite the presence of the occluder. In Figs. 7(e) and 7(f) we depict two novel views of the object for qualitative assessment.

7. CONCLUSIONS

In this paper, we demonstrated the interesting possibility of using motion parallax as a cue for inpainting within the SFF scenario. When a real-world camera is moved relative to a 3D object to capture a stack, the frames are subject to parallax effects. In addition, there can be data loss in certain regions due to sensor damage and/or occlusions. Despite missing regions in the observations, we have shown that it is possible to judiciously exploit the parallax cue to obtain the inpainted focused image and 3D structure of the underlying specimen using only a few frames from the stack. In effect, we were able to integrate image deblurring, and image and structure inpainting within a single unified framework. Some interesting extensions include generalizing the proposed method to handle arbitrary camera motion and automatic detection of missing regions by utilizing the frames in the SFF stack.

REFERENCES

1. S. K. Nayar and Y. Nakagawa, "Shape from focus," IEEE Trans. Pattern Anal. Mach. Intell. **16**, 824–831 (1994).
2. A. Nedzved, V. Bucha, and S. Ablameyko, "Augmented 3D endoscopy video," in *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video*, 28–30 May 2008, Istanbul, Turkey (2008), pp. 349–352.
3. M. Watanabe and S. K. Nayar, "Telecentric optics for focus analysis," IEEE Trans. Pattern Anal. Mach. Intell. **19**, 1360–1365 (1997).
4. R. G. Willson and S. A. Shafer, "What is the center of the image?" J. Opt. Soc. Am. A **11**, 2946–2955 (1994).
5. T. Darrell and K. Wohn, "Pyramid based depth from focus," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 1988), pp. 504–509.
6. R. Kingslake, *Optical System Design* (Academic, 1983).
7. R. R. Sahay and A. N. Rajagopalan, "Extension of the shape from focus method for reconstruction of high-resolution images," J. Opt. Soc. Am. A **24**, 3649–3657 (2007).
8. M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous structure and image inpainting," IEEE Trans. Image Process. **12**, 882–889 (2003).
9. M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques* (ACM Press, 2000), pp. 417–424.
10. J. Verdera, V. Caselles, M. Bertalmio, and G. Sapiro, "Inpainting surface holes," in *Proceedings of the IEEE International Conference on Image Processing* (IEEE, 2003), pp. 903–906.
11. A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," IEEE Trans. Image Process. **13**, 1200–1212 (2004).
12. K. A. Patwardhan, G. Sapiro, and M. Bertalmio, "Video inpainting under constrained camera motion," IEEE Trans. Image Process. **16**, 545–553 (2007).
13. S. Esedoglu and J. Shen, "Digital inpainting based on the Mumford-Shah-Euler image model," Eur. J. Appl. Math. **13**, 353–370 (2002).
14. C. A. Z. Barcelos and M. A. Batista, "Image restoration using digital inpainting and noise removal," Image Vis. Comput. **25**, 61–69 (2007).
15. A. C. Kokaram, "On missing data treatment for degraded video and film archives: a survey and a new Bayesian approach," IEEE Trans. Image Process. **13**, 397–415 (2004).
16. M. K. Ng, H. Shen, E. Y. Lam, and L. Zhang, "A total variation regularization based super-resolution reconstruction algorithm for digital video," EURASIP J. Advances Signal Process. **2007**, Article ID, 74585 16 pages (2007).
17. G. D. Finlayson, M. S. Drew, and C. Lu, "Entropy minimization for shadow removal," Int. J. Comput. Vis. **85**, 35–57 (2009).
18. Y. Shor and D. Lischinski, "The shadow meets the mask: Pyramid-based shadow removal," Comput. Graph. Forum **27**, 577–586 (2008).
19. P. Favaro and S. Soatto, "Seeing beyond occlusions (and other marvels of a finite lens aperture," in *Proceedings of*

the International Conference on Computer Vision and Pattern Recognition (IEEE, 2003), pp. 579–586.
20. S. S. Bhasin and S. Chaudhuri, "Depth from defocus in presence of partial self occlusion," in *Proceedings of the International Conference on Computer Vision* (IEEE, 2001), pp. 488–493.
21. S. W. Hasinoff and K. N. Kutulakos, "Confocal stereo," Int. J. Comput. Vis. **81**, 82–104 (2009).
22. M. Mcguire, W. Matusik, H. Pfister, J. F. Hughes, and F. Durand, "Defocus video matting," in ACM Trans. Graph. (TOG) **24**, 567–576 (2005).
23. S. W. Hasinoff and K. N. Kutulakos, "A layer-based restoration framework for variable-aperture photography," in *Proceedings of the 11th International Conference on Computer Vision* (IEEE, 2007), pp. 1–8.
24. M. Levoy, B. Chen, V. Vaish, M. Horowitz, I. McDowall, and M. Bolas, "Synthetic aperture confocal imaging," ACM Trans. Graph (TOG) **23**, 825–834 (2004).
25. 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 *Proceedings of the International Conference on Computer Vision and Pattern Recognition* (IEEE, 2006), pp. 2331–2338.
26. L. Wang, H. Jin, R. Yang, and M. Gong, "Stereoscopic inpainting: joint color and depth completion from stereo images," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition* (IEEE, 2008), pp. 1–8.
27. P. Brodatz, *Textures: A Photographic Album for Artists and Designers*, (Dover, 1966).
28. A. P. Pentland, "A new sense for depth of field," IEEE Trans. Pattern Anal. Mach. Intell. **9**, 523–531 (1987).
29. S. Chaudhuri and A. N. Rajagopalan, *Depth from Defocus: A Real Aperture Imaging Approach* (Springer-Verlag, 1999).
30. S. Z. Li, *Markov Random Field Modeling in Computer Vision* (Springer-Verlag, 1995).
31. J. Besag, "Spatial interaction and the statistical analysis of lattice systems," J. R. Stat. Soc. Ser. B (Methodol.) **36**, 192–236 (1974).
32. Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," IEEE Trans. Pattern Anal. Mach. Intell. **23**, 1222–1239 (2001).
33. V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?" IEEE Trans. Pattern Anal. Mach. Intell. **26**, 147–159 (2004).
34. A. Raj and R. Zabih, "A graph cut algorithm for generalized image deconvolution," in *Proceedings of the International Conference on Computer Vision* (IEEE, 2005), pp. 1048–1054.
35. C. Rother, V. Kolmogorov, V. Lempitsky, and M. Szummer, "Optimizing binary MRFs via extended roof duality," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition* (IEEE, 2007), pp. 1–8.
36. V. Kolmogorov and C. Rother, "Minimizing nonsubmodular functions with graph cuts-A review," IEEE Trans. Pattern Anal. Mach. Intell. **29**, 1274–1279 (2007).