Confocal Stereo

Samuel W. Hasinoff* and Kiriakos N. Kutulakos*

Dept. of Computer Science, University of Toronto
{hasinoff,kyros}@cs.toronto.edu

Abstract. We present *confocal stereo*, a new method for computing 3D shape by controlling the focus and aperture of a lens. The method is specifically designed for reconstructing scenes with high geometric complexity or fine-scale texture. To achieve this, we introduce the *confocal constancy* property, which states that as the lens aperture varies, the pixel intensity of a visible in-focus scene point will vary in a scene-*independent* way, that can be predicted by prior radiometric lens calibration. The only requirement is that incoming radiance within the cone subtended by the largest aperture is nearly constant. First, we develop a detailed lens model that factors out the distortions in high resolution SLR cameras (12MP or more) with large-aperture lenses (e.g., f1.2). This allows us to assemble an $A \times F$ aperture-focus image (AFI) for each pixel, that collects the undistorted measurements over all A apertures and F focus settings. In the AFI representation, confocal constancy reduces to color comparisons within regions of the AFI, and leads to focus metrics that can be evaluated separately for each pixel. We propose two such metrics and present initial reconstruction results for complex scenes.

1 Introduction

Recent years have seen many advances in the problem of reconstructing complex 3D scenes from multiple photographs [1, 2, 3]. Despite this progress, however, there are many common scenes for which obtaining detailed 3D models is beyond the state of the art. One such class includes scenes that contain very high levels of geometric detail, such as hair, fur, feathers, miniature flowers, etc. These scenes are difficult to reconstruct for a number of reasons—they create complex 3D arrangements not directly representable as a single surface; their images contain fine detail beyond the resolution of common video cameras; and they create complex self-occlusion relationships. As a result, many approaches either sidestep the reconstruction problem [2], require a strong prior model for the scene [4], or rely on techniques that approximate shape at a coarse level.

Despite these difficulties, the high-resolution sensors in today's digital cameras open the possibility of imaging complex scenes at a very high level of detail. With resolutions surpassing 12Mpixels, even individual strands of hair may

* Part of this work was done while the authors were visiting Microsoft Research Asia.

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part I, LNCS 3951, pp. 620–634, 2006.

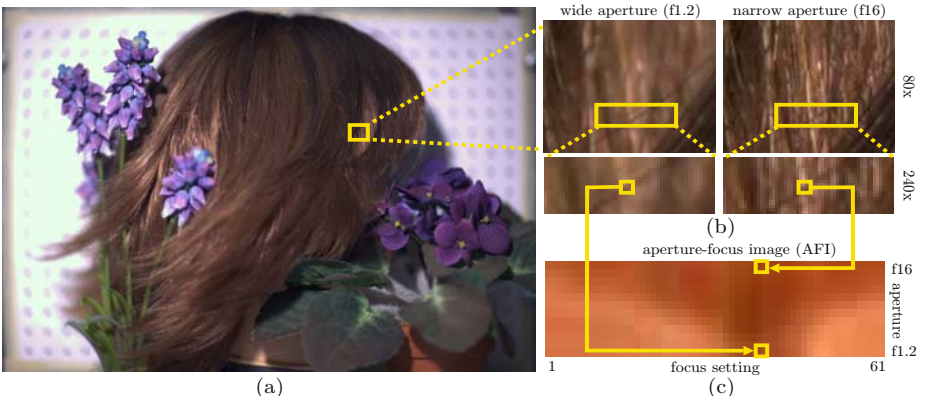

Fig. 1. (a) Wide-aperture image of a complex scene. (b) *Left:* Successive close-ups of a region in (a), showing a single in-focus strand of hair. *Right:* Narrow-aperture image of the same region, with everything in focus. Confocal constancy tells us that the intensity of in-focus pixels (e.g., on the strand) changes predictably between these two views. (c) The aperture-focus image (AFI) of a pixel near the middle of the strand. A column of the AFI collects the intensities of that pixel as the aperture varies with focus fixed.

be one or more pixels wide (Fig. 1a,b). In this paper, we explore the possibility of reconstructing such scenes with a new method called *confocal stereo*, which aims to compute depth maps at sensor resolution. The method is designed to exploit the capabilities of high-end digital SLR cameras and requires no special equipment besides the camera and a laptop. The only key requirement is the ability to actively control both the aperture and focus setting of the lens.

At the heart of our approach is a property we call *confocal constancy*, which states that as the lens aperture varies, the pixel intensity of a visible in-focus scene point will vary in a scene-*independent* way, that can be predicted by prior radiometric lens calibration. To exploit confocal constancy for reconstruction, we develop a detailed lens model that factors out the geometric and radiometric distortions observable in high resolution SLR cameras with large-aperture lenses (e.g., f1.2). This allows us to assemble an $A \times F$ *aperture-focus image (AFI)* for each pixel, that collects the undistorted measurements over all A apertures and F focus settings (Fig. 1c). In the AFI representation, confocal constancy reduces to color comparisons within regions of the AFI and leads to focus metrics that can be evaluated separately for each pixel.

Our work is closely related to depth-from-focus methods [5, 6, 7, 8], with the important difference that rather than defining our focus criterion over a spatial window, we consider pixels individually and manipulate a second, independent camera parameter (i.e., aperture). To our knowledge, aperture control has been considered only in the context of depth-from-defocus methods [9, 10, 11, 12], but these methods also rely on spatial windows and, hence, are unsuitable for reconstructing scenes at the resolutions we consider. Our work is also related to

recent approaches employing finite or synthetic apertures for image-based rendering [13] and for 3D reconstruction [14, 15]. Unlike these methods, our approach requires only a single camera, and requires no special illumination or scene model.

Our work has five main contributions. First, unlike existing depth-from-focus or depth-from-defocus methods, our confocal constancy formulation shows that we can assess focus without modeling a pixel's spatial neighborhood or the blurring properties of a lens. Second, we show that depth-from-focus computations can be reduced to a pixel-matching problem, in the spirit of traditional stereo techniques. Third, we develop a method for the precise geometric and radiometric alignment of images taken at multiple focus and aperture settings, particularly suited for the case where the standard thin-lens model breaks down. Fourth, we introduce the aperture-focus-image representation as a basic tool for focus- and defocus-based 3D reconstruction. Fifth, we show that together, confocal constancy and accurate image alignment lead to a reconstruction algorithm that can compute depth maps at resolutions not attainable with existing techniques.

2 Confocal Constancy

Consider a camera whose lens contains multiple elements and has a range of known focus and aperture settings. We assume that no information is available about the internal components of this lens (e.g., the spacing of its elements). We therefore model the lens as a "black box" that redirects incoming light toward a fixed sensor plane, with the following idealized properties:

- **Negligible absorption:** light that enters the lens in a given direction is either blocked from exiting or is transmitted with no absorption.
- **Perfect focus:** for every 3D point in front of the lens there is a unique focus setting that causes rays through the point to converge to a single pixel on the sensor plane.
- **Aperture-focus independence:** the aperture setting controls only which rays are blocked from entering the lens; it does not affect the way that light is redirected.

These properties are well approximated by lenses used in professional photography applications, and we use such a lens to collect images of a 3D scene for A aperture settings, $\{\alpha_1, \ldots, \alpha_A\}$, and F focal settings, $\{f_1, \ldots, f_F\}$. This acquisition produces a 4D set of pixel data, $I_{\alpha f}(x, y)$, where $I_{\alpha f}$ is the image captured with aperture α and focal setting f.

Suppose that a 3D point $\mathbf{p}$ on an opaque surface is in perfect focus in image $I_{\alpha f}$ and suppose that it projects to pixel (x, y). In this case, the light reaching the pixel is restricted to a cone from $\mathbf{p}$ determined by the aperture setting (Fig. 2). For a sensor with a linear response, the intensity $I_{\alpha f}(x, y)$ at the pixel is proportional to the integral of outgoing radiance over the cone, i.e.,

$$I_{\alpha f}(x, y) \;=\; \kappa \int_{\omega \in \mathcal{C}_{xy}(\alpha, f)} L(\mathbf{p}, \omega)\, \mathrm{d}\omega \;, \tag{1}$$

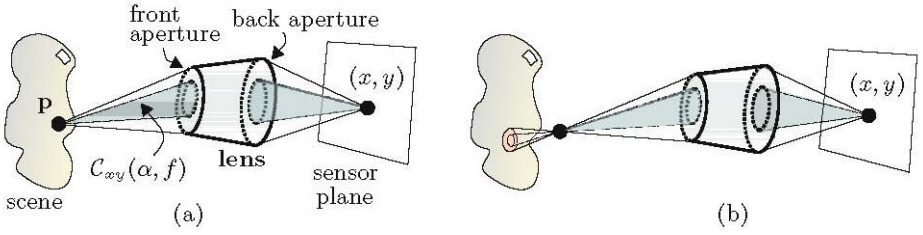

Fig. 2. Generic lens model. (a) At the ideal focus setting of pixel (x, y), the lens collects outgoing radiance from a scene point $\mathbf{p}$ and directs it toward the pixel. The 3D position of point $\mathbf{p}$ is uniquely determined by pixel (x, y) and its ideal focus setting. The shaded cone of rays, $\mathcal{C}_{xy}(\alpha, f)$, determines the radiance reaching the pixel. This cone is a subset of the cone subtended by $\mathbf{p}$ and the front aperture because some rays may be blocked by internal components of the lens, or by its back aperture. (b) For non-ideal focus settings, the lens integrates outgoing radiance from a region of the scene.

where ω measures solid angle, $L(\mathbf{p}, \omega)$ is the radiance for rays passing through $\mathbf{p}$, κ is a constant that depends only on the sensor's response function [16], and $\mathcal{C}_{xy}(\alpha, f)$ is the cone of rays that reach (x, y). In practice, the apertures on a real lens correspond to a nested sequence of cones, $\mathcal{C}_{xy}(\alpha_1, f) \subset \ldots \subset \mathcal{C}_{xy}(\alpha_A, f)$, leading to a monotonically-increasing intensity at the pixel.

If the outgoing radiance at the in-focus point $\mathbf{p}$ remains constant within the cone of the largest aperture, and if this cone does not intersect the scene elsewhere, the relation between intensity and aperture becomes especially simple. In particular, the integral of Eq. (1) disappears and the intensity for aperture α is proportional to the solid angle subtended by the associated cone, i.e.,

$$I_{\alpha f}(x, y) \;=\; \kappa \,\|\, \mathcal{C}_{xy}(\alpha, f) \,\|\, L(\mathbf{p}) \;, \tag{2}$$

where $\|\, \mathcal{C}_{xy}(\alpha, f) \,\| \;=\; \int_{\mathcal{C}_{xy}(\alpha,f)} \mathrm{d}\omega$. As a result, the ratio of intensities at an in-focus point for two different apertures becomes independent of the scene:

Confocal Constancy Property

$$\frac{I_{\alpha f}(x, y)}{I_{\alpha_1 f}(x, y)} \;=\; \frac{\|\, \mathcal{C}_{xy}(\alpha, f) \,\|}{\|\, \mathcal{C}_{xy}(\alpha_1, f) \,\|} \;\stackrel{\text{def}}{=}\; E_{xy}(\alpha, f) \;. \tag{3}$$

Intuitively, the constant of proportionality, $E_{xy}(\alpha, f)$, describes the relative amount of light received from an in-focus scene point for a given aperture. This constant, which we call the *relative exitance* of the lens, depends on lens internal design (front and back apertures, internal elements, etc.) and varies in general with aperture, focus setting, and pixel on the sensor plane.

Confocal constancy is an important property for evaluating focus for four reasons. First, it holds for a very general lens model that covers the lenses commonly used with high-quality SLR cameras. Second, it requires no assumptions about the appearance of out-of-focus points. Third, it holds for scenes with general reflectance properties, provided that radiance is nearly constant over the

cone subtended by the largest aperture.[1] Fourth, and most important, it can be evaluated at *pixel resolution* because it imposes no requirements on the spatial layout (i.e., depths) of points in the neighborhood of **p**.

3 The Confocal Stereo Procedure

Confocal constancy allows us to decide whether or not the point projecting to a pixel (x, y) is in focus by comparing the intensities $I_{\alpha f}(x, y)$ for different values of aperture α and focus f. This leads to the following reconstruction procedure:

1. **(Relative exitance estimation)** Compute the relative exitance of the lens for the A apertures and F focus settings (Sect. 4).
2. **(Image acquisition)** For each of the F focus settings, capture an image of the scene for each of the A apertures.
3. **(Image alignment)** Warp the captured images to ensure that a scene point projects to the same pixel in all images (Sect. 5).
4. **(AFI construction)** Build an $A \times F$ aperture-focus image for each pixel, that collects the pixel's measurements across all apertures and focus settings.
5. **(Confocal constancy evaluation)** For each pixel, process its AFI to find the focus setting that best satisfies the confocal constancy property (Sect. 6).

4 Relative Exitance Estimation

In order to use confocal constancy for reconstruction, we must be able to predict how changing the lens aperture affects the appearance of scene points that are in focus. Our approach is motivated by three basic observations. First, the apertures on real lenses are non-circular and the f-stop values describing them only approximate their true area (Fig. 3a,b). Second, when the aperture diameter is a relatively large fraction of the camera-to-object distance, the solid angles subtended by different 3D points in the workspace can differ significantly.[2] Third, vignetting and off-axis illumination effects cause additional variations in the light gathered from different in-focus points [17] (Fig. 3b).

To deal with these issues, we explicitly compute the relative exitance of the lens, $E_{xy}(\alpha, f)$, for all apertures α and for a sparse set of focal settings f. This can be thought of as a radiometric lens calibration step that must be performed just once for each lens. In practice, this allows us to predict aperture-induced intensity changes to within the sensor's noise level (i.e., within 1–2 gray levels).

To compute relative exitance for a focus setting f, we place a diffuse white plane at the in-focus position and capture one image for each aperture, $\alpha_1, \ldots, \alpha_A$. We then apply Eq. (3) to each pixel (x, y) to recover $E_{xy}(\alpha_i, f)$. To

[1] For example, a $70mm$ diameter aperture located $1.2m$ from the scene corresponds to 0.5% of the hemisphere, or a cone whose rays are less than 3.4° apart.

[2] For a $70mm$ diameter aperture, the solid angle subtended by scene points 1.1–$1.2m$ away can vary up to 10%.

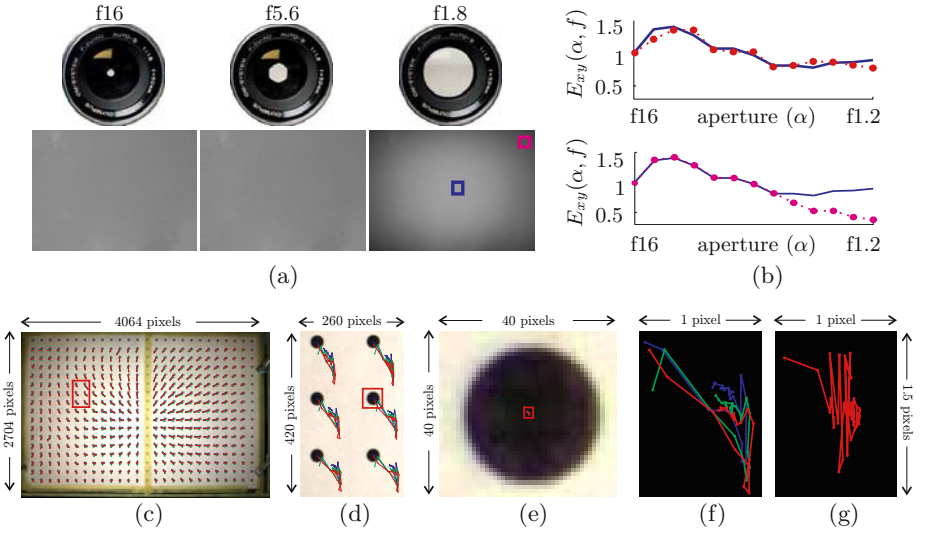

Fig. 3. (a) Images of an SLR lens showing variation in aperture shape with corresponding images of a diffuse plane. (b) *Top:* comparison of relative exitances for the central pixel indicated in (a), as measured using Eq. (3) (solid graph), and as approximated using the f-stop values (dotted) according to $E_{xy}(\alpha, f) = \alpha_1^2/\alpha^2$ [16]. *Bottom:* comparison of the central pixel (solid) with the corner pixel (dotted) indicated in (a). The agreement is good for narrow apertures (i.e., high f-stop values), but for wider apertures, spatially-varying effects are significant. (c–g) To evaluate non-deterministic lens distortions, we computed centroids of dot features for images of a static calibration pattern. (c–f) Successive close-ups of a centroid's trajectory for three cycles (red, green, blue) of the 23 aperture settings. In (c–d) the trajectories are magnified by a factor of 100. As shown in (f), the trajectory, while stochastic, correlates with aperture setting. (g) Trajectory for the centroid of (e) over 50 images with the same lens settings.

obtain $E_{xy}(\alpha_i, f)$ for focus settings that span the entire workspace, we repeat the process for multiple values of f and use interpolation to compute the in-between values. Since Eq. (3) assumes that pixel intensity is a linear function of radiance, we linearize the images using the inverse of the sensor response function [16].

5 High-Resolution Image Alignment

The intensity comparisons needed to evaluate confocal constancy are only possible if we can locate the projection of the same 3D point in multiple images taken with different settings. The main difficulty is that real lenses map in-focus 3D points onto the image plane in a non-linear fashion that cannot be predicted by ordinary perspective projection. To enable cross-image comparisons, we develop an alignment procedure that reverses these non-linearities and warps the input images to make them consistent with a reference image.

Since our emphasis is on reconstructing scenes at the maximum possible spatial resolution, we aim to model real lenses with enough precision to ensure

sub-pixel alignment accuracy. This task is especially challenging because at resolutions of 12MP or more, we begin to approach the optical and mechanical limits of the camera. In this domain, the commonly-used thin lens (i.e., magnification) model [6, 7, 8, 18, 15] is insufficient to account for observed distortions.

Deterministic second-order radial distortion model. To model geometric distortions caused by the lens optics, we use a model with $F + 5$ parameters for a lens with F focal settings. The model expresses deviations from an image with reference focus setting f_1 as an additive image warp consisting of two terms—a pure magnification term m_f that is specific to focus setting f, and a quadratic distortion term that amplifies the magnification:

$$\mathbf{w}_f^{\mathrm{D}}(x, y) = \big[\, m_f + m_f(f - f_1)(k_0 + k_1 r + k_2 r^2) - 1 \,\big] \cdot \big[(x, y) - (x_c, y_c)\big] \;, \quad (4)$$

where k_0, k_1, k_2 are the quadratic distortion parameters, (x_c, y_c) is the estimated image center, and $r = \|(x, y) - (x_c, y_c)\|$ is the radial displacement. Note that when the quadratic distortion parameters are zero, the model reduces to pure magnification. Also note that the quadratic distortion term depends linearly on the focus setting as well. Empirically, we have found that the model of Eq. (4) is necessary to obtain sub-pixel registration at high resolutions.

Non-deterministic first-order distortion model. We were surprised to find that significant misalignments can occur even when the camera is controlled remotely without any change in settings, and is mounted securely on an optical table (Fig. 3g). While these motions are clearly stochastic, we also observed a reproducible, aperture-dependent misalignment of about the same magnitude (Fig. 3c–f). In order to achieve sub-pixel alignment, we approximate these motions by a global 2D translation, estimated independently for every image:

$$\mathbf{w}_{\alpha f}^{\mathrm{ND}}(x, y) = \mathbf{t}_{\alpha f} \;. \quad (5)$$

Offline geometric lens calibration. We recover the full distortion model of Eqs. (4–5) in a single optimization step, using images of a calibration pattern taken over all F focus settings at the narrowest aperture, α_1. This optimization simultaneously estimates the $F + 5$ parameters of the deterministic model and the $2F$ parameters of the non-deterministic model. To do this, we solve a non-linear least squares problem that minimizes the squared reprojection error over a set of features detected on the calibration pattern:

$$E(x_c, y_c, \mathbf{m}, \mathbf{k}, \mathbf{T}) = \textstyle\sum_{(x,y)} \sum_f \, ||\mathbf{w}_f^{\mathrm{D}}(x, y) + \mathbf{w}_{\alpha_1 f}^{\mathrm{ND}}(x, y) - \Delta_{\alpha_1 f}(x, y)\,||^2 \;, \quad (6)$$

where $\mathbf{m}$ and $\mathbf{k}$ are the vectors of magnification and quadratic parameters, respectively; $\mathbf{T}$ collects non-deterministic translations; and $\Delta_{\alpha_1 f}(x, y)$ is the displacement between a feature location at focus setting f and its location at the reference focus setting, f_1. To increase robustness, we fit the model iteratively, removing features whose reprojection error is more than 3.0 times the median.

Online alignment. While the deterministic warp parameters need only be computed once for a given lens, we cannot apply the non-deterministic translations

computed during calibration to a different sequence. Thus, for a new capture we identify (potentially different) features in the scene and redo the optimization of Eq. (6), with all parameters except $\mathbf{T}$ fixed to the values computed offline.

6 Confocal Constancy Evaluation

Together, image alignment and relative exitance estimation allow us to establish a pixel-wise geometric and radiometric correspondence across all input images, i.e., for all aperture and focus settings. Given a pixel (x, y), we use this correspondence to assemble an $A \times F$ *aperture-focus image*, describing the pixel's intensity variations as a function of aperture and focus (Fig. 4a):

Aperture-Focus Image (AFI)

$$AFI_{xy}(\alpha, f) \;=\; \frac{1}{E_{xy}(\alpha, f)} \;\hat{I}_{\alpha f}(x, y) \;, \tag{7}$$

where $\hat{I}_{\alpha f}$ denotes the images after geometric image alignment.

AFIs are a rich source of information about whether or not a pixel is in focus at a particular focus setting f. We make this intuition concrete by developing two functionals that measure how well a pixel's AFI conforms to the confocal constancy property at f. Since we analyze the AFI of each pixel (x, y) separately, we drop subscripts and use $AFI(\alpha, f)$ to denote its AFI.

Direct Evaluation of Confocal Constancy. Confocal constancy tells us that when a pixel is in focus, its relative intensities across aperture should match the variation predicted by the relative exitance of the lens. Since Eq. (7) already corrects for these variations, confocal constancy at f implies constant intensity within column f of the AFI (Fig. 4b). Hence, to find the ideal focus setting we can simply find the column with minimum variance:

$$f^* \;=\; \arg\min_f \,\mathrm{Var}\,\{\, AFI(1, f),\, \ldots,\, AFI(A, f)\,\} \;. \tag{8}$$

The reason why the variance is higher at non-ideal focus settings is that defocused pixels integrate regions of the scene surrounding the true surface point (Fig. 2b), which generally contain "texture" in the form of varying geometric structure or surface albedo. Hence, for confocal constancy to be discriminative as a focus measure, such texture must be present in the scene.

Evaluation by AFI Model-Fitting. A disadvantage of the previous method is that most of the AFI is ignored when testing a given focus hypothesis f, since only one column participates in the calculation of Eq. (8). In reality, the 3D location of a scene point determines both the column of the AFI where confocal constancy holds as well as the degree of blur that occurs in the AFI's remaining, "out-of-focus" regions.[3] By taking these regions into account, we can create a focus detector with more resistance to noise and higher discriminative power.

[3] While not analyzed in the context of confocal constancy or the AFI, this is a key observation exploited by *depth-from-defocus* approaches [9, 10, 11, 12].

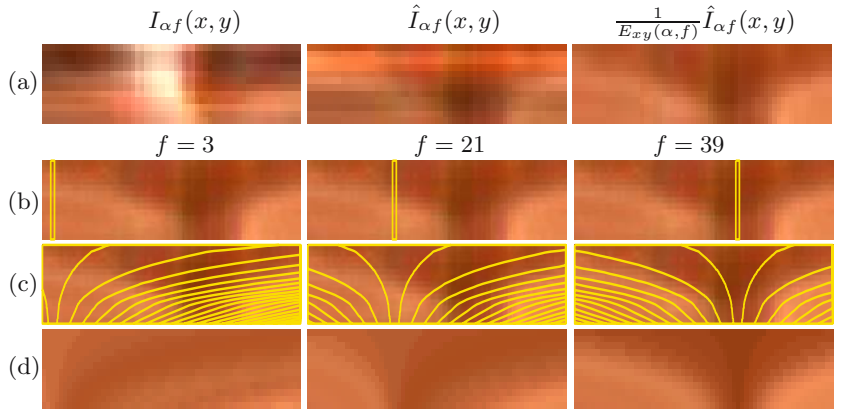

Fig. 4. (a) The $A \times F$ measurements for the pixel shown in Fig. 1. *Left:* prior to image alignment. *Middle:* after image alignment. *Right:* after accounting for relative exitance (Eq. (7)). Note that the AFI's smooth structure is discernible only after both corrections. (b) Direct evaluation of confocal constancy for three focus hypotheses. (c) Boundaries of the equi-blur regions, superimposed over the AFI (for readability, only a third are shown). (d) Results of AFI model fitting, with constant intensity in each equi-blur region, from the mean of the corresponding region in the AFI. Observe that for $f = 39$ the model is in good agreement with the measured AFI ((a), rightmost).

In order to take into account both in- and out-of-focus regions of a pixel's AFI, we develop an idealized, parametric AFI model that generalizes confocal constancy. This model is controlled by a single parameter—the focus hypothesis f—and is fit directly to a pixel's AFI measurements. The ideal focus setting is chosen to be the hypothesis that maximizes agreement with these measurements.

Our AFI model is based on two key observations. First, the AFI can be decomposed into a set of F disjoint *equi-blur* regions that are completely determined by the focus hypothesis f (Fig. 4c). Second, under mild assumptions on scene radiance, the intensity within each equi-blur region will be constant when f is the correct hypothesis. These observations suggest that we can model the AFI as a set of F constant-intensity regions whose spatial layout is determined by the focus hypothesis f. Fitting this model to a pixel's AFI leads to a focus criterion that minimizes intensity variance in every equi-blur region (Fig. 4d):

$$f^* \;=\; \arg\min_f \sum_{i=1}^{F} \Big(w_i^f \;\mathrm{Var}\left\{ \; AFI(\alpha,\phi) \;\middle|\; (\alpha,\phi) \in \mathcal{R}_i^f \; \right\} \;\Big) \;, \tag{9}$$

where $\mathcal{R}_i^f$ is the i-th equi-blur region for hypothesis f, and w_i^f weighs the contribution of region $\mathcal{R}_i^f$ ($w_i^f = \mathrm{area}(\mathcal{R}_i^f)$ in our experiments).

To implement Eq. (9) we must compute the equi-blur regions for a given focus hypothesis f. Suppose that the hypothesis f is correct, and suppose that the current aperture and focus of the lens are α and f, respectively, i.e., a scene point $\mathbf{p}$ is in perfect focus for this setting. Now consider "defocusing" the lens by

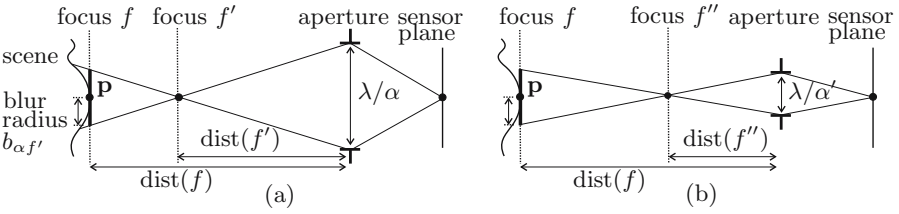

Fig. 5. (a) Quantifying the blur due to aperture α at a non-ideal focus setting f'. The aperture's diameter can be expressed in terms of its f-stop value α and the focal length λ. (b) A second aperture-focus combination with the same blur radius. In our AFI model, (α, f') and (α', f'') belong to the same equi-blur region.

changing its focus to f' (Fig. 5a). We represent the blur associated with the pair (α, f') by a circular disc centered on point $\mathbf{p}$ and parallel to the sensor plane. From similar triangles, the radius of this disc is equal to

$$b_{\alpha f'} = \frac{\lambda}{2\alpha} \frac{|\text{dist}(f) - \text{dist}(f')|}{\text{dist}(f')} , \tag{10}$$

where λ is the focal length of the lens and dist$(\cdot)$ converts focus settings to distances from the front aperture.

Given a focus hypothesis f, Eq. (10) assigns a "blur radius" to each point (α, f') in the AFI and induces a set of nested, wedge-shaped curves of equal blur radius (Figs. 4c and 5b). We quantize the possible blur radii into F bins associated with the widest-aperture settings, i.e., $(\alpha_A, f_1), \ldots, (\alpha_A, f_F)$, which partitions the AFI into F equi-blur regions, one per bin.

While Eq. (10) fully specifies our parametric AFI model, it is important to note that this model is approximate. We have implicitly assumed that once relative exitance and geometric distortion have been factored out (Sects. 4–5), defocusing is well-approximated by the thin-lens model [17]. Moreover, the intensity at two equi-blur positions in an AFI will be constant only if two conditions hold: (i) outgoing radiance remains constant within the cone of the largest aperture for *all* scene points contributing intensity to the pixel (i.e., the shaded region of the scene in Fig. 2b), and (ii) depth variations within this region do not significantly affect the defocus integral. In practice, we have found that this model matches the observed pixel variations quite well (Fig. 4d).

7 Experimental Results

To test our approach, we used a Canon EOS-1Ds digital SLR camera with a wide-aperture, fixed focal length lens (Canon EF85mm 1.2L). The lens aperture was under computer control and its focal setting was adjusted manually using a printed ruler on the body of the lens. We operated the camera at its highest resolution, capturing 4604×2704-pixel images in RAW 12-bit mode. Each image was demosaiced using Canon software and linearized using the algorithm in

[16]. We used $A = 13$ apertures ranging from f1.2 to f16, and $F = 61$ focal settings spanning a workspace that was $17cm$ in depth and $1.2m$ away from the camera. Successive focal settings therefore corresponded to a depth difference of approximately $2.8mm$. We mounted the camera on an optical table in order to allow precise ground-truth measurements and to minimize external vibrations.

To enable the construction of aperture-focus images, we first computed the relative exitance of the lens (Sect. 4) and then performed offline geometric calibration (Sect. 5). Our geometric distortion model was able to align the calibration images with an accuracy of approximately 0.15 pixels, estimated from centroids of dot features (Fig. 3e). The accuracy of online alignment was about 0.5 pixels, i.e., higher than during offline calibration but well below one pixel. This penalty is expected since far fewer features are used for online alignment.

Quantitative evaluation: "Box" dataset. To quantify reconstruction accuracy, we used a tilted planar scene consisting of a box wrapped in newsprint (Fig. 6). The plane of the box was measured with a FaroArm Gold 3D touch probe whose single-point accuracy was $\pm 0.05mm$ in the camera's workspace. To relate probe coordinates to coordinates in the camera's reference frame we used the Matlab Camera Calibration Toolbox along with further correspondences between image features and 3D coordinates measured by the probe. A similar procedure was used to estimate the mapping between focal settings and the depth of in-focus points, i.e., the dist$(\cdot)$ function in Eq. (10).

We computed a depth map of the scene for three focus criteria: direct confocal constancy (Eq. (8)), AFI model-fitting (Eq. (9)), and a depth-from-focus (DFF) method, applied to the widest-aperture images, that chooses the focus setting with the highest variance in a 3×3 window centered at each pixel. The planar shape of the scene and its detailed texture can be thought of as a best-case scenario for such window-based approaches. The plane's footprint contained 2.8 million pixels, yielding an equal number of 3D measurements. As Table 1 shows, all three methods performed quite well, with accuracies of 0.37–0.45% of the object-to-camera distance. This performance is on par with previous quantitative studies (e.g., [12]) although few results with real images have been reported in the passive depth-from-focus literature. Significantly, AFI model-fitting slightly outperforms spatial variance (DFF) in both accuracy and number of outliers even though its focus computations are performed entirely at the pixel level and, hence, are of much higher resolution. Qualitatively, this behavior is confirmed by considering all three criteria for specific pixels (Fig. 6, top).

Table 1. Ground-truth accuracy results. The inlier threshold was set to $11mm$. All distances were measured relative to the ground-truth plane.

	median ABS dist. (mm)	inlier RMS dist. (mm)	% inliers	RMS % dist. to camera
confocal constancy evaluation	3.18	4.61	66	0.454
AFI model fitting	2.13	3.78	84	0.373
3×3 spatial variance (DFF)	2.16	3.79	80	0.374

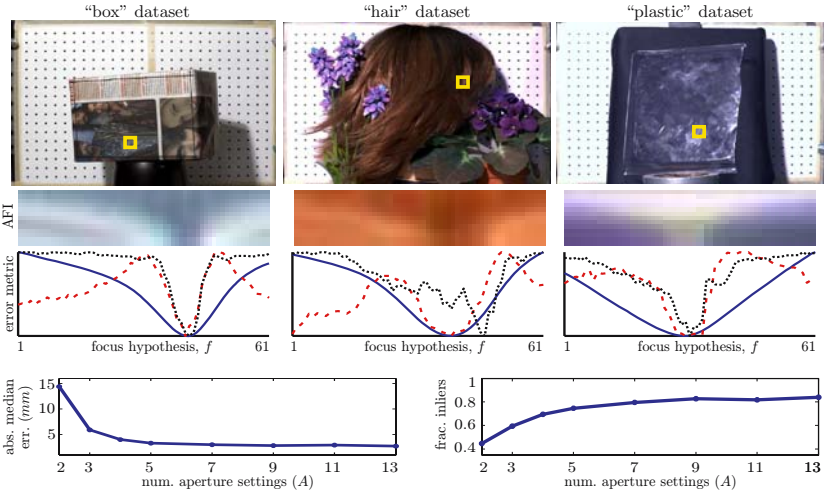

Fig. 6. *Top:* Behavior of focus criteria for a specific pixel (highlighted square) in three test datasets. The dotted graph is for 3×3 variance (DFF), dashed is for direct confocal constancy (Eq. (8)) and the solid graph is for AFI model-fitting (Eq. (9)). While all three criteria often have corresponding local minima near the ideal focus setting, AFI model-fitting varies much more smoothly and exhibits no spurious local minima in these examples. For the middle example, which considers the same pixel shown in Fig. 1, the global minimum for variance is at an incorrect focus setting. This is because the pixel lies on a strand of hair only 1–2 pixels wide, beyond the resolving power of variance calculations. *Bottom:* AFI model fitting error and inlier fraction as a function of A ("box" dataset, inlier threshold $= 11mm$).

As a final experiment with this dataset, we investigated how AFI model fitting degrades when a reduced number of apertures is used (i.e., for AFIs of size $A'\times F$ with $A' < A$). Our results suggest that reducing the apertures to five or six causes little reduction in reconstruction quality (Fig. 6, bottom).

"Hair" dataset. Our second test scene was a wig with a messy hairstyle, approximately $25cm$ tall, surrounded by several artificial plants (Figs. 1 and 6).[4] Reconstruction results for this scene (Fig. 7) show that our confocal constancy criteria lead to very detailed depth maps, at the resolution of individual strands of hair, despite the scene's complex geometry and despite the fact that depths can vary greatly within small image neighborhoods (e.g., toward the silhouette of the hair). By comparison, the 3×3 variance operator produces uniformly-lower resolution results, and generates smooth "halos" around narrow geometric structures like individual strands of hair. In many cases, these "halos" are larger than the width of the spatial operator, as blurring causes distant points to influence the results.

In low-texture regions, such as the cloth flower petals and leaves, fitting a model to the entire AFI allows us to exploit defocused texture from nearby

[4] For additional results, see `http://www.cs.toronto.edu/~hasinoff/confocal`.

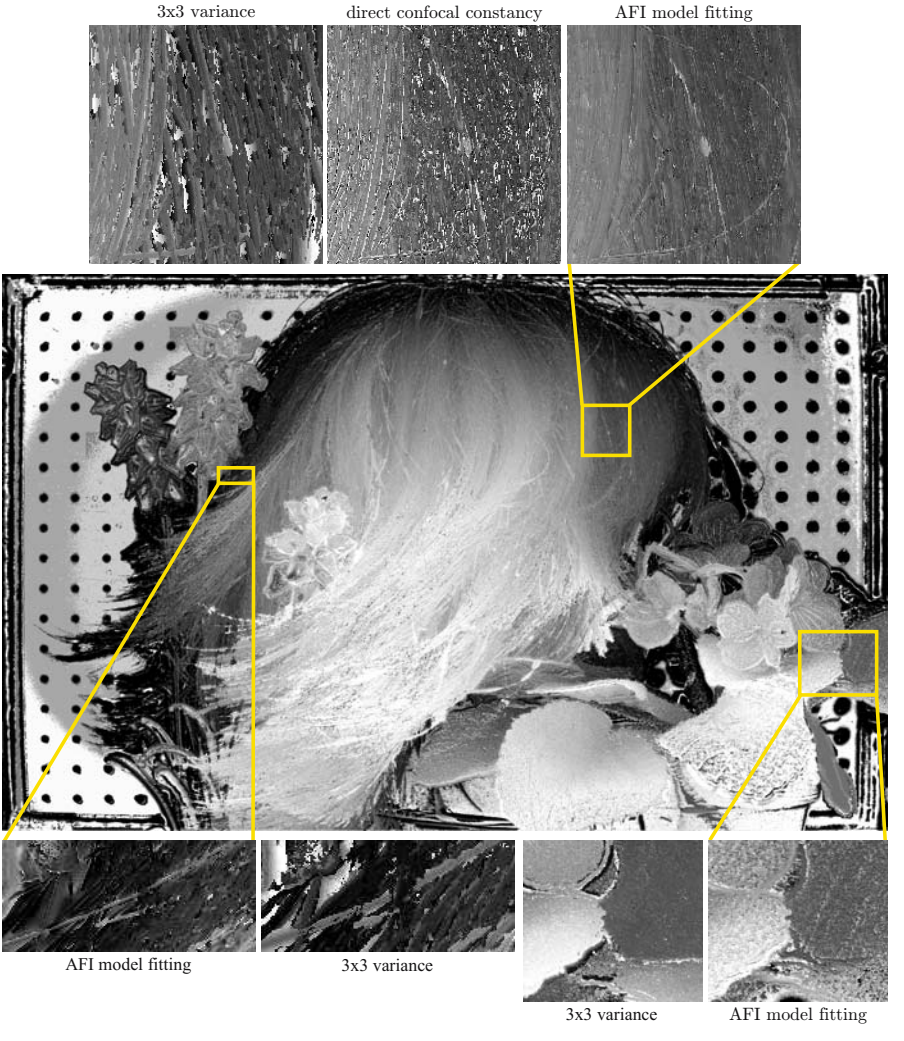

Fig. 7. *Center:* Depth map for the "hair" dataset using AFI model fitting. *Top:* Several distinctive foreground strands of hair are resolved in the AFI-based depth map. Direct evaluation of confocal constancy is also sharp but much noisier, making structure difficult to discern. By contrast, 3×3 variance (DFF) exhibits thick "halo" artifacts and fails to detect most of the foreground strands (see also Fig. 6, top). *Bottom right:* DFF yields smoother and more accurate depths for the low-texture leaves. *Bottom left:* Unlike DFF, AFI model fitting resolves structure amid significant depth discontinuities.

scene points. Window-based methods like variance, however, generally yield even better results in such regions, because they propagate focus information from nearby texture more directly. Like all focus measures, those based on confocal constancy are uninformative in extremely untextured regions, i.e., when the AFI

is constant. Such pixels may be detected using a "confidence" measure (e.g., assessing the steepness of the minimum) or by processing the AFI further.

8 Concluding Remarks

The extreme locality of shape computations derived from aperture-focus images is both a key advantage and a major limitation of the current approach. While we have shown that processing a pixel's AFI leads to highly detailed reconstructions, this locality does not yet provide the means to handle large untextured regions or to reason about global scene geometry and occlusion [18, 19, 15]. To handle low texture, we are exploring the possibility of analyzing AFIs at multiple levels of detail and for multiple pixels simultaneously. We are also investigating a space-sweep approach to analyze occlusions, analogous to voxel-based stereo.

Acknowledgements. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under the RGPIN and CGS-D programs, by a fellowship from the Alfred P. Sloan Foundation, by an Ontario Premier's Research Excellence Award and by Microsoft Research.

References

1. Zitnick, C.L., Kang, S.B., Uyttendaele, M., Winder, S., Szeliski, R.: High-quality video view interpolation using a layered representation. In: SIGGRAPH. (2004) 600–608
2. Fitzgibbon, A., Wexler, Y., Zisserman, A.: Image-based rendering using image-based priors. IJCV **63** (2005) 141–151
3. Hertzmann, A., Seitz, S.M.: Example-based photometric stereo: Shape reconstruction with general, varying BRDFs. PAMI **27** (2005) 1254–1264
4. Wei, Y., Ofek, E., Quan, L., Shum, H.Y.: Modeling hair from multiple views. In: SIGGRAPH. (2005) 816–820
5. Darrell, T., Wohn, K.: Pyramid based depth from focus. In: CVPR. (1988) 504–509
6. Nayar, S., Watanabe, M., Noguchi, M.: Real-time focus range sensor. PAMI **18** (1996) 1186–1198
7. Favaro, P., Soatto, S.: Learning shape from defocus. In: ECCV (2). (2002) 735–745
8. Favaro, P., Osher, S., Soatto, S., Vese, L.A.: 3D shape from anisotropic diffusion. In: CVPR (1). (2003) 179–186
9. Pentland, A.P.: A new sense for depth of field. PAMI **9** (1987) 523–531
10. Subbarao, M., Surya, G.: Depth from defocus: A spatial domain approach. IJCV **13** (1994) 271–294
11. Farid, H., Simoncelli, E.P.: Range estimation by optical differentiation. J. Optical Society of America, A **15** (1998) 1777–1786
12. Watanabe, M., Nayar, S.K.: Rational filters for passive depth from defocus. IJCV **27** (1998) 203–225
13. Isaksen, A., McMillan, L., Gortler, S.J.: Dynamically reparameterized light fields. In: SIGGRAPH. (2000) 297–306

14. Levoy, M., Chen, B., Vaish, V., Horowitz, M., McDowall, I., Bolas, M.T.: Synthetic aperture confocal imaging. In: SIGGRAPH. (2004) 825–834
15. Favaro, P., Soatto, S.: Seeing beyond occlusions (and other marvels of a finite lens aperture). In: CVPR (2). (2003) 579–586
16. Debevec, P., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: SIGGRAPH. (1997) 369–378
17. Smith, W.J.: Modern Optical Engineering. 3rd edn. McGraw-Hill, NY (2000)
18. Asada, N., Fujiwara, H., Matsuyama, T.: Seeing behind the scene: Analysis of photometric properties of occluding edges by the reversed projection blurring model. PAMI **20** (1998) 155–167
19. Schechner, Y.Y., Kiryati, N.: Depth from defocus vs. stereo: How different really are they? IJCV **39** (2000) 141–162