ON THE USE OF DEPTH-FROM-FOCUS IN 3D OBJECT MODELLING FROM MULTIPLE VIEWS

Ning Xu and Narendra Ahuja

Department of Electrical and Computer Engineering and Beckman Institute
University of Illinois at Urbana-Champaign, Urbana, IL61801, USA

ABSTRACT

This paper is concerned with depth estimation from focus for acquiring 3D models of objects from multiple views. Depth-from-focus usually involves a single viewpoint but many different focus settings. Our approach uses multiple viewpoints, but for each viewpoint only two different-aperture settings, one small, showing the entire visible surface in focus, and the other large, showing only a narrow depth slice of the object in focus. Since only the aperture value is changed, the camera needs to be calibrated only once for each viewpoint. Depth estimate follows directly for those parts of the image that are in focus in both images for the same viewpoint. The depth estimates from different viewpoints are merged to obtain a complete surface estimate. This estimate serves as the initial estimate for the next stage which refined the estimate using multi-baseline stereo matching. This design has been implemented using a single camera, with a controllable aperture, and aimed at an object located on a rotary platform that sequentially captures the required set of images. Criteria are discussed for selecting camera parameters to achieve the narrowest possible depth-of-field, to obtain the best depth-from-focus estimates. Experimental results for two simple objects are presented to validate our approach.

1. INTRODUCTION

3D models of real world objects are often needed for virtual reality (VR) and multimedia systems. There are two main approaches to the acquisition of real world object models. One uses structured lighting to acquire images from multiple views, while the other uses regular lighting. The first approach [1, 2] projects a structured light pattern, usually a stripe of laser beam generated by a semiconductor laser, onto the surface of an object and captures the reflected image with a CCD camera. The image yields a depth map of the visible part of the surface. If such a scan is obtained from different viewpoints around the object, the resulting depth maps can be combined to compute a 3D surface model of the object. Laser scanners typically produce accurate results but are very expensive. The second approach uses regular images obtained from multiple viewpoints and recovers from them a 3D description of the object. 3D model reconstruction from regular images is still an open problem. Two basic classes of such approaches can be distinguished. The first class computes depth maps from individual viewpoints and then registers them into a single 3D surface model. To obtain the depth maps from different viewpoints, various methods such as depth from focus [3] or defocus [4, 5] stereo vision [6, 7], structure from motion [8] and shape from shading [9] have been proposed, each of which has different degrees of applicability in different conditions. The second class of algorithms is based on volumetric representations and is often referred to as shape estimation from silhouettes [10, 11] and shape from space carving [12].

This paper presents a method that belongs to the first class of the second approach mentioned above. The method is aimed at reconstructing 3D object models in a tabletop environment using multiple views with the help of depth-from-focus. When we are using depth-from-focus with multiple views, only two different-aperture images are required from each viewpoint, and these two different-aperture images share the same calibration parameters, which means the camera needs to be calibrated only once for each viewpoint. Depth estimates using focus from different viewpoints are integrated as a rough object surface estimate. Since we have images from many viewpoints, multi-baseline stereo can be used to refine this rough surface, while the search range for stereo are narrowed. In addition, we derive the criteria for camera parameters selection in order to achieve the narrowest possible depth-of-field which is useful to improve the accuracy of the focus based depth estimates.

Section 2 presents a brief review of other work related to the proposed method. Section 3 through 6 describe our approach in detail. Section 7 presents experimental results. Section 8 contains concluding remarks.

2. RELATED WORK

Below we briefly review the work done on the different surface estimation methods related to our approach.

2.1. Depth from focus or defocus

For a camera with focal length f, an object point at a distance u from the lens can be in focus only if the image sensor plane is at a distance v from the lens on the other side, and u, v and f satisfy the equation: $\frac{1}{f} = \frac{1}{u} + \frac{1}{v}$.

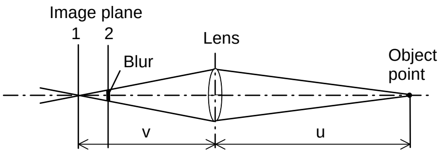

Fig. 1. A schematic diagram of the process of image formation.

As in Fig.1, the image sensor must be located at position 1 to capture a focused image of the object point. If the image sensor is at a different position, say 2, the image of the object will be blurred. Thus, if an image point is in focus, the depth of the corresponding object point can be estimated using the equation given above, and the known focal length f and image distance v [3]. Depth estimates can also be obtained for the out-of-focus points if the degree of defocus can be measured and a model of defocusing is available [4, 5, 13]. An illumination pattern that is projected via the same optical path used to acquire images helps to realize a real time range sensor for estimating both textured and textureless surfaces [14].

2.2. Multi-baseline stereo

Stereo vision [6] involves using two calibrated cameras to take images of the same scene from two slightly different viewpoints. The 3D surface is estimated by finding pairs of points in the two images, each corresponding to a point on a 3D surface, and computing the depth of the scene point as a function of the positional disparity between its two images. The task of matching points between the two images is known as the correspondence problem. Multi-baseline stereo [7] uses multiple images of a scene taken from different known viewpoints, yielding multiple stereo pairs of images which provide redundancy in surface estimation. The further apart a given pair of stereo viewpoints are, the higher the accuracy of the resulting surface estimate.

2.3. Integrating focus and stereo

There are also some approaches trying to integrate focus and stereo [15, 16]. The approach in [15] estimates surface with the integrated use of focus, camera vergence and stereo disparity information. A vision system that using shape from focus and rotational stereo to model 3D object is proposed in [16]. In these works, depth-from-focus from each viewpoint requires many different focus settings.

3. THE USE OF DEPTH-FROM-FOCUS IN MULTI-VIEW CASES

In this section, we present how to use depth-from-focus in 3D modelling from multiple views, and how it differes from traditional depth-from-focus methods.

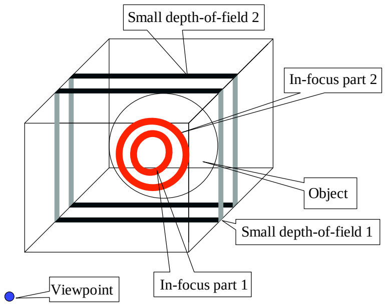

Fig. 2. Traditional depth-from-focus considers single viewpoint but many different focus settings. Two in-focus parts are shown for two of the focus settings.

Traditional depth-from-focus methods compute depth maps from a single viewpoint, and require many different focus settings for this viewpoint. Each focus setting yields depth estimate of a small in-focus part. For example, Fig. 2 shows two in-focus parts for two different focus settings for the same viewpoint. The complete depth map is obtained by merging all these in-focus parts. For each focus setting used, the projection matrix will be different since the camera parameters (including focal length f, object distance u, and image distance v) will change. Thus, camera should be calibrated for each focus setting.

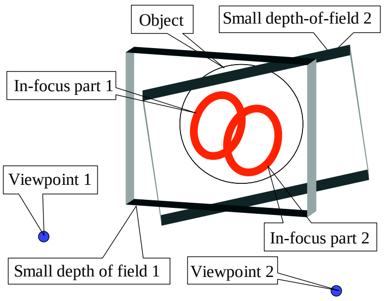

Fig. 3. Depth-from-focus considered in multiple-viewpoint case. Two viewpoints and their corresponding in-focus parts and depth-of-fields are shown.

In our approach, only two different focus settings are required for each viewpoint. These two settings are acquired by changing aperture which does not affect the camera calibration parameters. Thus, we only need to calibrate the camera once for each viewpoint. The input image captured with a small aperture shows the entire visible surface in focus, and the other one with large aperture shows in focus only a small part of the object surface contained within the narrow depth-of-field. The depth map of the in-focus part for each viewpoint is estimated and all in-focus parts from different viewpoints are merged as a rough object surface model. Fig. 3 shows an example of two in-focus parts estimated from two different viewpoints. If many viewpoints are used, the shaded in-focus parts will cover the surface of the object.

Those images obtained with small aperture cameras from different viewpoints can then be used as input to a multi-baseline stereo algorithm to refine the result, while the search ranges are narrowed by the rough surface estimate from focus.

4. CAMERA PARAMETER SELECTION

The large aperture camera is used to distinguish those points on the object surface within the depth-of-field from those outside the depth-of-field — the former are in focus and the latter are blurred. The narrower the depth-of-field, the higher the accuracy of the depth estimate desired from focus, and therefore, we need to select camera parameters to yield narrowest depth-of-field. The exact depth-of-field depends on imaging parameters including aperture size, focal length and object distance. It is calculated [17] as

$$d = \frac{2A \cdot p \cdot u \cdot f \cdot (u-f)}{A^2 \cdot f^2 - p^2 \cdot (u-f)^2}$$

,where f is the focal length, u is the distance between object point and lens, A is the radius of aperture and p is the largest possible radius of the blur circle around a single pixel before it crosses into the next pixel in the image sensor. In the equation above, A, p, u and f are the four parameters that can be adjusted to minimize d. Obviously, p is given for given sensor and independent of other parameters. Since we always have $u > f$,

$$\frac{\partial d}{\partial p} = \frac{2A \cdot u \cdot f \cdot (u-f) \cdot [A^2 \cdot f^2 + p^2 \cdot (u-f)^2]}{[A^2 \cdot f^2 - p^2 \cdot (u-f)^2]^2} > 0,$$

we need to minimize p to make d as small as possible. The rest of the parameters need to be selected under two other constraints. First, the maximum value of A is limited. Every lens has its own f-number range, whose smallest value corresponds to the largest aperture of the lens. For different lenses, their smallest f-numbers are approximately the same, say s. Then $s \cdot A/2 \leq f \Rightarrow A \leq 2/s \cdot f$. Let $a = 2/s$, then we have $A \leq a \cdot f$. Second, if O denotes object size, I denotes image sensor size and $M = O/I$, we must have $u \geq (1+M) \cdot f$ in order to capture the image of the entire object. Since

$$\frac{\partial d}{\partial A} = -\frac{2p \cdot u \cdot f \cdot (u-f) \cdot [p^2 \cdot (u-f)^2 + A^2 \cdot f^2]}{[A^2 \cdot f^2 - p^2 \cdot (u-f)^2]^2} < 0$$

and

$$\frac{\partial d}{\partial u} = \frac{2A^3 \cdot p \cdot f^3 \cdot (2u-f) + 2A \cdot p^3 \cdot f^2 \cdot (u-f)^2]}{[A^2 \cdot f^2 - p^2 \cdot (u-f)^2]^2} > 0,$$

we know that for a given f, we need to maximize A and minimize u to make d as small as possible. Substituting $A = a \cdot f$ and $u = (1+M) \cdot f$, we get

$$\frac{\partial d}{\partial f} = \frac{4a \cdot p \cdot (1+M) \cdot M \cdot f^3 \cdot (a^2 \cdot f^2 - 2p^2 \cdot M^2]}{[a^2 \cdot f^2 - p^2 \cdot M^2]^2}$$

and $\frac{\partial^2 d}{\partial f^2} > 0$. Thus the optimal f value is given by $f = \sqrt{2} \cdot p \cdot M/a$. Considering some typical values of p, M and a, e.g. $p = 6.7/2 \mu m$, $M = 60$ and $a = 2/1.6$, it can be seen that the resulting f value of $0.23mm$ is quite small in comparison with f values of common lenses. Therefore, we need to select the smallest available focal length lens and adjust u and v accordingly to be able to capture the full object on an image sensor of a given size.

5. SYSTEM SETUP AND CAMERA CALIBRATION

5.1. System setup

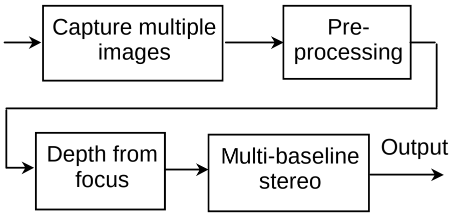

Fig. 4. The flowchart of our approach.

Fig. 4 presents an overview of the entire system. Ideally, cameras will be placed such that the depth-of-field of those cameras with large apertures will roughly cover the object surface, i.e., all parts of the object surface appear in focus in one or more blurred (large aperture) images. However, in our implementation, we will assume, without loss of generality, that a single stationary camera acquires the multiple required images of the object as it rotates about a vertical axis. Two images are captured for each object orientation, one using a small aperture and the other using a large aperture. This configuration corresponds to our current implementation, as shown in Fig.5. A CCD camera is used to take both sets of images as the object is rotated on a rotary stepper. A positioning system is used to adjust the optical axis of the camera so it intersects with, and is perpendicular to, the rotation axis of the rotary stepper.

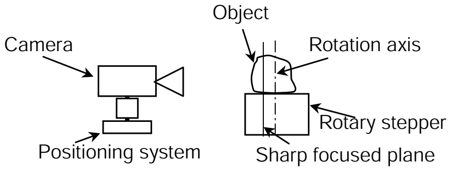

Fig. 5. Experimental layout.

5.2. Calibration

Calibration is required to estimate the various imaging parameters. This is done in two steps. First, we determine the intrinsic parameters of the camera by using a calibration grid placed at various distances from the camera, as described in [18]. Second, we adjust the camera's optical axis using a positioning system so that it intersects with the rotation axis of the rotary stepper at right angle. Then, we use the large aperture camera to take an image of the grid while it is rotated by a certain angle. By manually identifying the in-focus parts of the grid in the images, we can calculate the distance between rotation axis and the sharp-focused plane.

After the system is calibrated, a CCD camera is used to take both sets of images as the object is rotated by 720 degrees using a rotary stepper. For the first 360 degrees, the camera aperture is set to a small value, and for the second 360 degrees, it is set to a large value. The large aperture gives a blurred image, and the small aperture gives an in-focus reference image. By adjusting the shutter speed, the relative irradiances are adjusted so that both images capture the same amount of light.

6. ALGORITHM IMPLEMENTATION

From each viewpoint and for each aperture setting, multiple images are captured sequentially and then averaged to reduce additive sensor noise. The images are also used to estimate the variance of additive noise. In addition, the object boundary in each reference image is identified, which is easily done since the object is placed against a very dark background.

6.1. Depth from focus

Since the depths of the in-focus pixels in the large-aperture images can be computed from the known depth-of-field of the camera, the task in this step is to identify these in-focus points in the large-aperture images. In our approach, these in-focus pixels are located by analyzing intensity variations in their vicinity and their conformity with the corresponding pixels in the small-aperture images. Specifically, for each pixel in the large-aperture images, three questions are asked. First, is this pixel inside the object boundary? If it is, and if it is sharply focused, (i.e. the depth of this pixel in the current setting is equal to u, which is the distance between optical center and sharp focused plane), we calculate the locations of its corresponding pixels in the images from other viewpoints, and check whether they are all inside the corresponding object boundaries. Obviously, a surface point of the object will never have its image outside of the object boundary regardless of the viewpoint from which it is seen. Second, is the variance of this pixel's neighborhood greater than that of the noise estimated in the pre-processing? If the neighborhood of a pixel has a smaller variance than the noise, no difference can be perceived between its blurred image and its reference image. Third, is the correlation between a small window around this pixel and the corresponding window in its reference image higher than a given threshold? The correlation value is assumed to be higher than the threshold if the pixel is in-focus. Those pixels satisfying all three conditions are selected as focused and their depths in the current images are estimated to be u.

6.2. Using multi-baseline stereo

With input images from multiple views, multi-baseline stereo is used in our implementation to refine the surface estimates from focus. For each reference image, two previous reference images and two following reference images are selected for multi-baseline stereo analysis. The reason we select only 5 consecutive reference images is that if the cameras are further apart, the perspective deformation of the neighborhood region of the pixel is significant and affects the result of correlation. Dense correspondences are estimated within the object boundary using conventional window-based correlation methods, while the searches are constrained within a small range around the depth estimates given by depth-from-focus. The resulting 3D surface points are represented using cylindrical coordinates along the rotary axis. Then we simply average those points sharing the same height and angle, and use bilinear interpolation to obtain depth estimates for the points without depth estimates. Thus we obtain a surface estimate of the object represented using polar coordinates along the rotary axis. Finally, the whole object surface is smoothed using a sliding average window of a small size.

7. EXPERIMENTAL RESULTS

To validate our approach of using focus to help stereo matching in multi-view cases, we present our experimental results on two real objects in this section. Since it is not our purpose to fully develop a system giving the best performance, which could be achieved by improving and optimizing the techniques for depth-from-focus and multi-baseline stereo individually, we only use simple objects to experiment with, and use cylindrical coordinate representations to simplify the computation.

In the first experiment, we use a cylinder with a highly textured surface as our test object. The cylinder is put on the rotary stepper whose rotation axis coincides with the vertical cylinder axis and is rotated by 2×360 degrees. 36 different stepper positions are used, 10 degrees apart. Each of the two rotations yields 36 different viewpoints. At each viewpoint of each rotation, 5 images are taken and averaged. The first rotation is performed using a small aperture and shutter speed of $1/60$ second. In the second rotation when the aperture size is increased, the shutter speed is decreased to $1/10,000$ second so as to maintain the same total image irradiance in both rotations. Two sample image pairs are shown in Fig. 6.

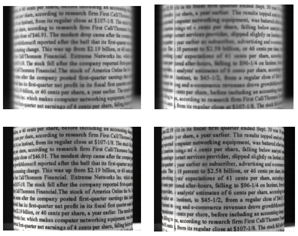

Fig. 6. (a) Two sample image pairs obtained in the first experiment. The upper row shows large-aperture images and the lower row shows small-aperture, reference images. Images in the same column are captured from the same viewpoint.

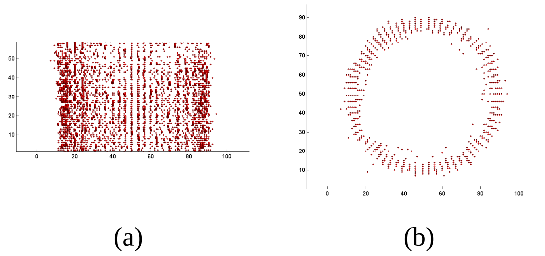

Fig. 7. Points obtained by the depth-from-focus algorithm. (a) A side view of the 3D points. (b) A top view of the 3D points.

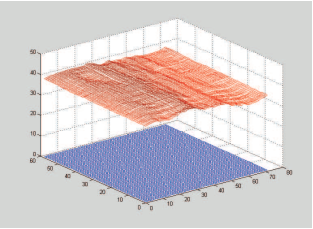

Fig. 9. Estimated distances of cylinder surface points from the axis.

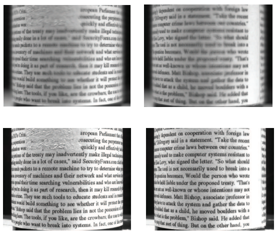

Fig. 10. Two samples of image pairs obtained in the second experiment. The upper row shows large-aperture images. Images in the same column are captured from the same viewpoint.

We first apply our depth-from-focus algorithm and obtain the depth estimates of the in-focus pixels in large-aperture images. The resulting 3D points in this step are shown in Fig. 7. Fig .7(a) shows a side view of these 3D points, while Fig. 7(b) shows the top view. It can be seen that the result of the depth-from-focus is approximately correct. The final estimated surface model is shown in Fig. 8. For comparison with the ground truth, the result is represented using polar coordinates at each height value, as shown in Fig. 9. The RMS error of the depth estimates is calculated to be 1.55% of the cylinder radius.

In the second experiment, we consider another cylinder identical to the first one, except that some parts of the cylinder surface are textureless. Camera settings are the same as in the first experiment. Two sample image pairs are shown in Fig. 10. The corresponding surface represented in the polar coordinates is shown in Fig.11. The RMS error is calculated to be about 2.49% of the cylinder radius.

In the third experiment, we consider a cubic object of a different size. We choose a suitable set of camera parameters and take the same number of images as in pervious experiments. Two sample image pairs are shown in Fig. 12. The estimated 3D surface model is shown in Fig. 13.

8. CONCLUSIONS AND DISCUSSION

We have presented an approach that uses depth-from-focus and multi-baseline stereo on multiple views of a small object to obtain its 3D surface model. For each viewpoint, our depth-from-focus algorithm requires only two different-aperture images and the camera needs to be calibrated only once. (Traditional depth-from-focus methods usually require many images using different focus settings for each viewpoint and calibrate camera for each focus setting.) The criteria for camera parameter selection to achieve the narrowest depth-of-field are also derived in this paper.

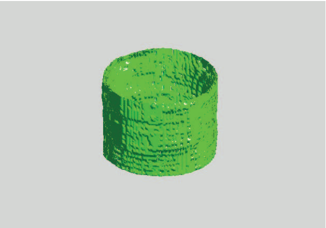

Fig. 8. 3D surface model derived in the first experiment.

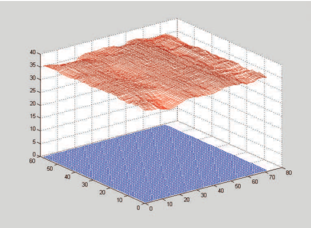

Fig. 11. Estimated distances of cylinder surface points from the axis.

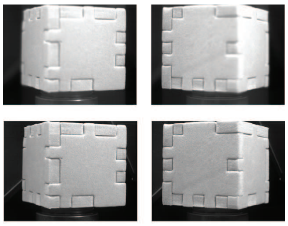

Fig. 12. Two samples of image pairs obtained in the third experiment. The upper row shows large-aperture images and the lower row contains reference images. Images in the same column are captured from the same viewpoint.

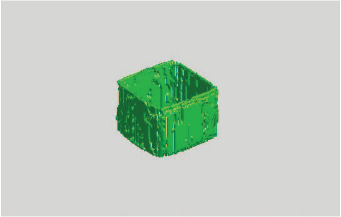

Fig. 13. 3D surface model derived in the third experiment.

When applying multi-baseline stereo, the search for matching points is simplified by using the depth-from-focus estimates as constraints. Most previous methods for 3D modelling from regular images captured using a turntable are based on stereo or motion. The proposed design uses both focus and stereo cues.

The surface parts that have weaker textures yield poorer depth estimates using both depth-from-focus and stereo methods. For such parts, optical texture may be projected on the object to improve the quality of the estimates.

9. REFERENCES

[1] R. Jarvis, “A perspective on range finding techniques for computer vision,” *PAMI*, vol. 5, no. 2, pp. 122–139, Mar. 1983.

[2] E. Mouaddib, J. Batlle, and J. Salvi, “Recent progress in structured light in order to solve the correspondence problem in stereo vision,” *ICRA*, pp. 130–136, April 1997.

[3] P. Grossmann, “Depth from focus,” *Pattern Recognition Letters*, vol. 5, pp. 63–69, Jan. 1987.

[4] A. Pentland, “A new sense of depth of field,” *IEEE Trans. Patt. Anal. Machine Intell.*, vol. PAMI-9, pp. 523–531, July 1987.

[5] G. Surya and M. Subbarao, “Depth from defocus by changing camera aperture: a spatial domain approach,” *Proceedings. of CVPR’93*, pp. 61–67, June 1993.

[6] S.T. Barnard and M.A. Fishler, “Computational stereo,” *Computing Surveys*, vol. 14, no. 4, pp. 554–572, 1982.

[7] M. Okutomi and T. Kanade, “A multiple-baseline stereo,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 4, pp. 353–363, Apr. 1993.

[8] J. Weng, N. Ahuja, and T.S. Huang, “Optimal motion and structure estimation,” *IEEE PAMI*, vol. 15, no. 9, pp. 864–884, 1993.

[9] B.K.P. Horn, *Shape from Shading: A Method for Obtaining the Shape of a Smooth Opaque Object from One View*, Ph.D. thesis, Massachusetts Inst. Of Technology, 1970.

[10] S. Sullivan and J. Ponce, “Automatic model construction, pose estimation, and object recognition from photographs using triangular splines,” *ICCV*, pp. 90–95, Jan. 1998.

[11] S. Srivastava and N. Ahuja, “An algorithm for generating octrees from object silhouettes in perspective views,” *Proc. IEEE Workshop Computer Vision*, pp. 363–365, Dec. 1987.

[12] K. N. Kutulakos and S. M. Seitz, “A theory of shape by space carving,” *International Journal of Computer Vision*, vol. 38, no. 3, pp. 199–218, July 2000.

[13] Y. Xiong and S.A. Shafer, “Depth from focusing and defocusing,” in *DARPA93 Image Understanding Workshop*, 1993, pp. 967–976.

[14] Shree K. Nayar, Masahiro Watanabe, and Minori Noguchi, “Real-time focus range sensor,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 12, pp. 1186–1198, 1996.

[15] A. L. Abbot and N. Ahuja, “Active surface reconstruction by integrating focus, vergence, stereo, and camera calibration,” in *International Conference on Computer Vision*, 1990, pp. 489–492.

[16] H. Y. Lin and M. Subbarao, “A vision system for fast 3d model reconstruction,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001, pp. 663–668.

[17] A. Krishnan and N. Ahuja, “Range estimation from focus using a non-frontal imaging camera,” *IJCV*, vol. 20, pp. 169–185, 1996.

[18] R. Y. Tsai, “An efficient and accurate camera calibration technique for 3d machine vision,” *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 364–374, May 1986.