Depth Estimation by Combining Stereo Matching and Coded Aperture

Chun Wang, Erdem Sahin, Olli Suominen, Atanas Gotchev

Tampere University of Technology, Tampere, Finland

***Abstract*—We investigate possible improvements that can be achieved in depth estimation by merging coded apertures and stereo cameras. We analyze several stereo camera setups which are equipped with different sets of coded apertures to infer about such possibilities. The demonstrated results of this analysis are encouraging in the sense that coded apertures can provide valuable complementary information to stereo based depth estimation in some cases. In particular, by utilizing the inherent relation between stereo cue and defocus cue, we extract depth information more robustly, especially for problematic scene regions.**

***Index Terms*—Depth estimation, stereo matching, depth from defocus, coded aperture, point spread function**

I. Introduction

Depth estimation is a comprehensively studied problem in the computer vision society. The developed depth estimation algorithms utilize the binocular depth cue obtained from stereo vision and/or monocular depth cues such as texture gradient and defocus. Algorithms based on stereo vision works effectively in the most of cases. The fundamental stage of such algorithms is the stereo matching in the sense that their performance mainly depends on the success of finding pixel correspondences in stereo views. In stereo matching, for each pixel in one view of the scene, its corresponding pixel in the other view is somehow found. Then, assuming that the camera settings are known (via calibration etc.), the depth of the scene can be estimated by triangulation. Several stereo matching algorithms have been proposed in the literature [1]. In spite of such variety that makes trading off between the accuracy and the computation time possible, stereo vision can not yet provide satisfactory depth estimates for some problematic scene regions such as ones having periodic textures, no textures, occluded parts or edges along epipolar lines.

Monocular depth cues are also widely utilized in depth estimation. Among several monocular depth cues, the defocus cue is the most exciting one. By using it, Pentland [2] initiated depth from defocus (DfD), which then becomes a popular passive depth estimation method. In DfD, the depth estimation is done by identifying the degree of blur, which is characterized by the extent of point spread function (PSF), throughout the image. In order to overcome the ill-posedness of the problem, usually two or more defocused images are captured from the same view with different but known camera settings, so the same object is blurred to different degrees. The resulting different measurements, together with known camera parameters, are sufficient to determine the amount of blur throughout the image and the corresponding depth [3].

Besides DfD, there is another depth estimation method utilizing defocus cue that is named as coded aperture (CA) due to the insertion of a special mask in the aperture position. CA was originally used in Astronomy to increase angular resolution and signal-to-noise ratio (SNR). In computer vision, it has also been utilized for different purposes such as light field capture [4] and deblurring [5]. Here we emphasize its application to depth estimation. The principle of CA for depth estimation is that the inserted mask modifies the frequency response of aperture filter so that it becomes easier to discriminate different filter scales (which correspond to different depths). Compared to DfD, CA can relieve the burden of having at least two images from the same view if a proper mask is in use, but better results can be expected if a pair of complementary masks are involved. Significant work has been done to find an optimal aperture mask or a pair of complementary masks [6], [7], [8]. In addition, several depth estimation algorithms are proposed as well, both for single mask [6], [9] and mask pair [8].

Instead of being used independently, monocular and binocular depth cues can also be utilized jointly. Indeed, stereo cue together with various monocular cues such as defocus and texture gradient are used by the human visual system to perceive depth information. Based on the motivation that stereo and monocular cues can provide complementary information, both cues began to be used together in the same system [10] [11], to improve depth estimation.

In this paper, we investigate depth estimation from a stereo camera equipped with CA. We particularly utilize CA to be able to get defocus cue effectively. Recently, Takeda et al. [12] presented a system employing a similar idea of merging CA and stereo. They focus the cameras to different depths to increase DfD performance. However, utilization of CA is not optimized in the sense of depth estimation, indeed they use the same mask in both cameras that is actually chosen according to its deblurring performance. Here we use two identical cameras to avoid undesired effects (e.g. zooming) of different camera parameters to stereo matching and we utilize aperture masks, either the same or not, which are optimized for depth estimation.

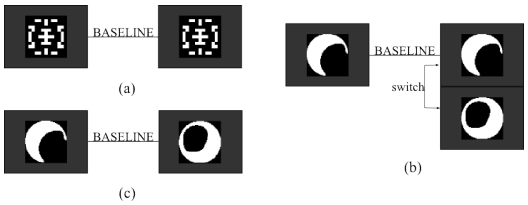


Fig. 1. Three camera setups. (a) two cameras with masks, e.g. the Levin's mask, in stereo setup; (b) two cameras with one of Zhou's mask pair (denoted as Zhou 1) in stereo setup, one more camera with other mask of Zhou's pair (denoted as Zhou 2) is used to capture one more image on the right view; (c) two cameras with Zhou's mask pair in stereo setup.

II. Camera Setups and Methods

A camera records the projection of a three-dimensional (3D) scene onto a sensor plane. Let us consider a Lambertian scene without any occlusion that can be represented by a curved surface $S \subset \mathbb{R}^3$ which is traced by the vector $\boldsymbol{r}$, i.e. $\boldsymbol{r} \in S$. The captured image for such a 3D scene can be written as

$$f(x,y) = \iint_S u(\boldsymbol{r}) p_{\boldsymbol{r}}(x,y) dS, \tag{1}$$

where $u(\boldsymbol{r})$ is the light intensity at $\boldsymbol{r}$ on S and $p_{\boldsymbol{r}}(x,y)$ is the PSF which is determined by the aperture shape of the camera as well as the distance of the point $\boldsymbol{r}$ to the camera plane. Hence, once the aperture shape is fixed, we can obtain the depth information by estimating the correct PSF throughout the image. Please note that Eq. 1 can be extended to more generic scenes by specially treating occlusions.

DfD and CA approaches try to somehow invert the relation given by Eq. 1 to estimate depth. In this paper we employ two of them that are introduced in [9] and [8], considering their demonstrated effectiveness. Favaro et al. [9] use least squares formulation to invert Eq. 1. On the other hand, Zhou et al. [8] further impose the locally fronto-planar assumption on the scene and thus model the imaging locally as a convolution

$$f = f_0 * p_d + \eta \tag{2}$$

where $*$ is the convolution operation; f_0 is the latent sharp image, p_d is the PSF at depth d and η is noise. To solve Eq. 2, they use deblurring based maximum a posteriori formulation where they utilize image prior information. In implementation, both approaches require a set of PSFs sampled at discrete depths. While Favaro's approach need one image, Zhou's approach requires two images captured with a pair of masks on a single view.

Based on the motivation of merging CA and stereo, here in this paper we discuss about three different camera setups, as shown in Fig. 1, and their purposes. In the first setup, we use two cameras with masks, either the same or not, e.g. the Levin's mask in both as shown in Fig. 1(a). In the second setup shown in Fig. 1(b), one more camera is added to the right view where we shall have a pair of images captured with Zhou's complementary masks pair. Finally, in the third setup shown in Fig. 1(c), each camera is equipped with one of Zhou's mask pair so that no extra camera is required. Please note that we choose to utilize Levin's mask and Zhou's mask pair for their superb depth discrimination capability.

TABLE I
Summary of the Stereo Version of Zhou's Approach

INPUTS:	
(f_L, f_R) :	captured left and right images;
PSFs :	a set of pre-sampled PSF pairs on different depths; each pair is denoted as $\{p_L, p_R\}$;
STEPS:	
1 :	For each (p_L, p_R) in PSFs
2 :	Find its associated disparity range D;
3 :	For each d in D
4 :	$f'_L = f_L(x-d, y)$
5 :	$\hat{F}_0 = \frac{F'_L \overline{P}_L + F_R \overline{P}_R}{\lvert\overline{P}_L\rvert^2 + \lvert\overline{P}_R\rvert^2 + \lvert C\rvert^2}$
6 :	$\mathrm{E}(p_L, p_R, d) = \sum_{i=L,R} \left\lvert f_i - ifft\left(\hat{F}_0 P_i\right)\right\rvert^2$
7 :	End for
8 :	End for
9 :	$(defocus, disparity) = \underset{p_L,\, p_R,\, d}{\arg\min}\ E, \forall pixel$
NOTATIONS:	
F_i :	the Fourier transform of f_i;
$\overline{P}_i$:	the complex conjugate of P_i;
C :	a matrix of noise to signal ratio;

The first two setups are considered based on two questions. One is whether using aperture masks in a stereo camera seriously affects the performance of stereo matching; the other is whether CA can give us useful information where the stereo matching fails. For the first question, we design the first setup by which testing the influence of different masks on ordinary stereo matching is possible. Our observations, which we present in Sec. III, lead to the proposition of integrated systems as the first setup (with the Levin's mask in both) and the second setup. Since in those two system both stereo matching and CA can be applied independently, they are used to explore the second question.

In the third setup, we want to take advantage of the effectiveness of Zhou's complementary masks without requiring an additional camera as in the second setup. Facing the problem that the requirement of Zhou's approach that two images are captured on the same view is not satisfied in this setup, we develop a variation of Zhou's approach for stereo images which employs the inherent relation between disparity and defocus. Intuitively, if the shifting between stereo images is done by correct disparity value (for a particular depth), the corresponding pixels in two images will be well aligned so that Zhou's approach will be able to be applied for them. Ideally, there exists a one-one mapping between disparity and defocus, as used in [12]. However, in most practical cases the depth resolution that can be achieved by CA is lower than the resolution provided by stereo. As a consequence of this resolution mismatch, here we set the relation between disparity and defocus as multi-one. Theoretically, the correct disparity-defocus pair will give the minimum error. The procedure is summarized in Table I.

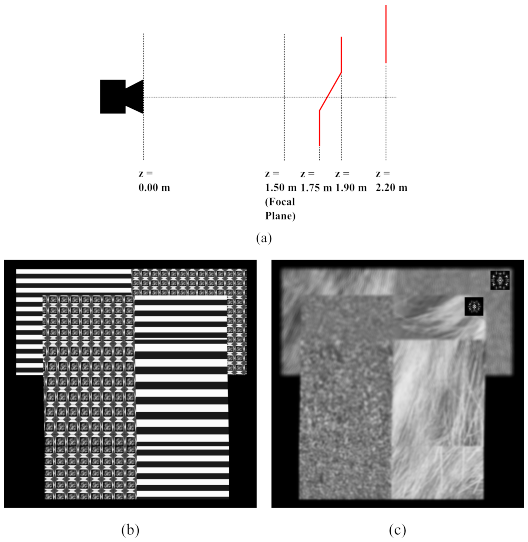


Fig. 2. Simulation environment. (a) the arrangement of the virtual camera and the scene; (b) a captured image on the left view with *ideal* pinhole aperture and the problematic texture; (c) a captured image on the right view with the Levin's mask and the good texture, together with two example PSFs at depths z=1.9m and z=2.2m (PSFs are scaled by a factor of 3 for visualization). Please note there exists an occluded area in our scene.

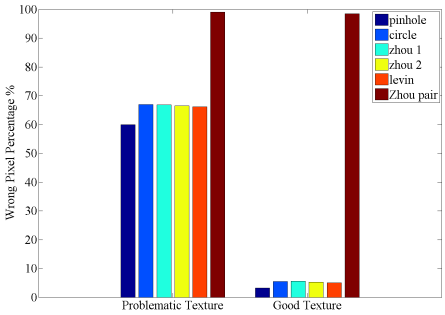


Fig. 3. The error percentages of stereo matching for different aperture masks, for both the problematic texture case and the good texture case. Please note that the pixels belonging to the black background are not considered in comparisons.

III. Simulation Results

The synthetic scene that we use in our simulations includes three fronto-parallel plane and a slanted plane as shown in Fig. 2(a). Two cases are considered. One is problematic textures including repetitive pattern and stripes. The other includes gravel and rabbit's fur which are good textures in the sense of randomness. A virtual camera, with a 35mm lens and 3.3 μm pixel pitch, is put in the middle of the baseline of a normal stereo setup and focused at 1.5 meter. The baseline length is set to 5 cm. The left view and right view images are generated by shifting the middle view image where the shift amount is calculated by triangulation. An example of captured image, for the left view, with the *ideal* pinhole camera model and the problematic texture is shown in Fig. 2(b).

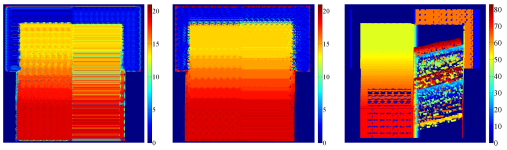
For a (physically valid) camera model having a physical aperture, the parts of the scene that are out of focus are blurred by a depth dependent PSF. Under thin lens and paraxial optics approximations together with aberration free lens and perfectly incoherent light assumptions, we derive the PSF for a single lens imaging system using wave optics [13] as

$$p_d(x,y) = \frac{1}{d^2}\left| \int\int a(\xi,\eta) \exp\left\{ j\frac{\pi}{\lambda} z_d(\xi^2+\eta^2) \right\} \times \exp\left\{ -j\frac{2\pi}{\lambda l}(x\xi+y\eta) \right\} d\xi d\eta \right|^2, \quad (3)$$

where $a(\xi,\eta)$ is the lens aperture function or the mask for CA, d is the depth of the point, l is the distance between lens and sensor plane, $z_d = \frac{1}{d} + \frac{1}{l} - \frac{1}{f}$ (f is the focal length) and λ is the wavelength of the light. We work with the green channel and thus take $\lambda = 534nm$. An example of right view image, captured with the Levin's mask for the good texture, together with examples of two PSFs at different depths are shown in Fig. 2(c). We use Eq. 3 to also determine the depth resolution of CA. We set the resolution to 2.5 cm for which we observe that two PSFs are discriminable enough and thus form the set of PSFs to be used in depth estimation algorithms accordingly.

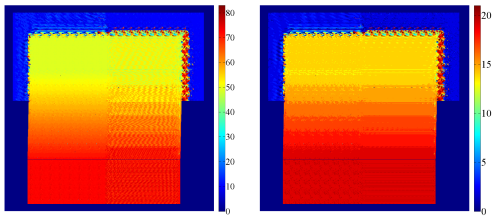
In order to observe whether using aperture masks in a stereo camera seriously affects the performance of stereo matching, we apply the same stereo matching algorithm [14], to different stereo image pairs captured with identical aperture masks (the first camera setup discussed in Sec. II). We consider pinhole, circular mask, Levin's mask, and one of Zhou's mask pair (at a time) cases. We also test situations where one camera is equipped with one of the Zhou's mask pair and the other camera is equipped with the other one. The raw disparity maps are all compared with the ground truth disparity map, and percentages of wrong disparity values are given in Fig. 3, for both the problematic texture case and the good texture case. As shown in Fig. 3, for the good texture case, the effects of using the same aperture masks on the performance of stereo matching are not severe; while for the problematic texture case, stereo matching already fails even if no mask is used. We also test the case that two cameras are equipped with different masks (Zhou's pair). In this case, the performance of stereo matching decreases dramatically even with the good texture. One possible reason of this degradation is that different mask shapes result in different PSFs which affects the captured images differently. Therefore, stereo matching may fail to find correspondences. The other reason is that two masks have different geometrical centres, which induces an extra baseline (the amount changes with depth). Thus, the disparity value will be incorrect even if two points are matched correctly. These results indicate that if the same mask is employed in both cameras, we can have integrated systems, as shown in Fig. 1(a) and Fig. 1(b), where both stereo matching and CA can be applied independently.

In order to see whether CA can give us useful information where the stereo matching fails, we particularly consider problematic texture case and attempt to do depth estimation by also CA approaches. There we test two cases: Favaro's



(a) (b) (c)

Fig. 4. Three depth maps produced by three approaches on the problematic texture case. (a) the defocus map obtained by the Favaro's approach with the Levin's mask; (b) the defocus map obtained by the Zhou's approach with Zhou's mask pair on the single view; (c) the disparity map obtained by stereo matching with *ideal* pinhole aperture



(a) (b)

Fig. 5. The results of the proposed approach on the problematic texture case. (a) the disparity map; (b) the defocus map.

approach with Levin's mask and Zhou's approach with Zhou's mask pair on the single view, as shown in Fig. 1(a) and Fig. 1(b), respectively. The results, together with that from the stereo matching in pinhole case, are shown in Fig. 4. Although the depth resolution provided by CA is lower than that from stereo matching, we get more reliable information as can be observed in the figure. It is worth to point out that Zhou's approach gives the best result, but it takes two images from the same view, which might be a limitation in practice. Fascinated by performance of Zhou's pair on problematic area and depth resolution of stereo matching, yet unwilling to take more images, we design the third setup shown in Fig. 1(c) for which we propose the stereo version of Zhou's approach. The disparity and defocus maps produced by the proposed approach are shown in Fig. 5, for the problematic texture case. The results are promising. Comparing Fig. 4(b) and Fig. 5(b), we can say that the proposed approach does not degrade the performance of original Zhou's approach. In addition, it produces an acceptable disparity map simultaneously, which has a significant improvement compared to the disparity map produced by pure stereo matching shown in Fig. 4(c). However, it is worth to mention that it still suffers from the occlusion problem introduced by the stereo vision.

IV. Conclusions

Based on our preliminary results, we observe that putting the same mask does not severely affect the performance of stereo matching, and CA approaches can give more reliable information on the problematic area where the stereo matching usually fails. In addition, we propose a stereo version of Zhou's approach which produces disparity map and defocus map simultaneously, and its results are promising even for the problematic textures we test. The performance of the proposed stereo version of Zhou's approach can be further improved, if we can form a better set of disparity-defocus mappings.

For the setups shown in Fig. 1(a) and Fig. 1(b), defocus maps produced by the CA and disparity map produced by stereo matching can be merged (knowing the inherent relation between defocus and disparity) to improve the result in more realistic scenes, including various problematic and good textures, for example by applying Markov Random Field (MRF).

Based on the demonstrated preliminary results, we conclude that the combination of stereo vision and CA is worth to study further for improved depth estimation performance.

References

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[2] A. Pentland, "A new sense for depth of field," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 9, no. 4, pp. 523–531, Jul. 1987.

[3] S. Chaudhuri and A. Rajagopalan, *Depth From Defocus: A Real Aperture Imaging Approach*. New York: Springer-Verlag New York, 1999.

[4] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, "Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing," in *ACM SIGGRAPH*, San Diego, California, USA, 2007.

[5] C. Zhou and S. Nayar, "What are good apertures for defocus deblurring?" in *IEEE Int. Conf. Computational Photography*, Apr. 2009, pp. 1–8.

[6] A. Levin, R. Fergus, F. Durand, and W. Freeman, "Image and depth from a conventional camera with a coded aperture," in *ACM SIGGRAPH*, San Diego, California, USA, 2007.

[7] A. Sellent and P. Favaro, "Optimized aperture shapes for depth estimation," *Pattern Recognition Lett.*, vol. 40, no. 0, pp. 96 – 103, Apr. 2014.

[8] C. Zhou, S. Lin, and S. Nayar, "Coded aperture pairs for depth from defocus," in *IEEE 12th Int. Conf. Computer Vision*, Kyoto, Japan, Sept. 2009, pp. 325–332.

[9] P. Favaro and S. Soatto, "A geometric approach to shape from defocus," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 406–417, Mar. 2005.

[10] A. Saxena, J. Schulte, and A. Ng, "Depth estimation using monocular and stereo cues," in *Proc. of the 20th Int. Joint Conf. Artifical Intelligence*, 2007, pp. 2197–2203.

[11] A. Rajagopalan, S. Chaudhuri, and U. Mudenagudi, "Depth estimation and image restoration using defocused stereo pairs," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1521–1525, Nov. 2004.

[12] Y. Takeda, S. Hiura, and K. Sato, "Fusing depth from defocus and stereo with coded apertures," in *IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2013, pp. 209–216.

[13] J. Goodman, *Introduction to Fourier Optics*, 2nd ed. McGraw-Hill, 1996.

[14] W. Abbeloos, "Real-time stereo vision," Master's thesis, Karel de Grote-Hogeschool University College, Belgium, May 2010.