

Depth Estimation by Combining Stereo Matching and Coded Aperture

Chun Wang ^{#1}, Erdem Sahin ^{#2}, Olli Suominen ^{#3}, Atanas Gotchev ^{#4}

[#] *Department of Signal Processing, Tampere University of Technology
Korkeakoulunkatu 10, Tampere, Finland, 33720*

² *erdem.sahin@tut.fi*

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 explore such possibilities. The demonstrated results of this analysis are encouraging in the sense that coded apertures can provide valuable complementary information to stereo vision based depth estimation in some cases. In addition to that, we take advantage of stereo camera arrangement to have a single shot multiple coded aperture system. We show that with this system, it is possible to extract depth information robustly, by utilizing the inherent relation between the disparity and defocus cues, even for scene regions which are problematic for stereo matching.

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

I. INTRODUCTION

The computer vision algorithms developed for the depth estimation problem usually utilize the binocular (disparity) depth cue and/or monocular depth cues such as texture gradient and defocus. Those which are based on stereo vision employ the binocular cue and work effectively in most cases. The fundamental stage of these algorithms is the stereo matching in the sense that their performance mainly depends on the success of finding pixel correspondences in stereo views. 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.

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, 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].

Coded aperture (CA), which refers to insertion of a special mask in the aperture position, is another depth estimation method which utilizes the defocus cue. 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 in depth estimation. The principle of CA for depth estimation is that the inserted mask modifies the PSF of the imaging system 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 is involved. Significant work has been done to find an optimal aperture mask or a pair of complementary masks [6], [7], [8]. In addition to that, several depth estimation algorithms are proposed for both single masks [6], [9] and mask pairs [8].

Based on the motivation that disparity cue and monocular cues can provide complementary information, both cues have also been used together in the same system [10] [11], to improve depth estimation. With the same motivation, here 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. In that work, the cameras are focused to different depths to increase DfD performance. However, utilization of CA is not optimized in the sense of depth estimation. Indeed, the mask used in both cameras is chosen according to its deblurring performance. Here we use two identical cameras to avoid undesired effects (e.g. zooming) caused by using different camera parameters on stereo matching and we utilize aperture masks, either the same or not, which are optimized for depth estimation.

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 r , i.e. $r \in S$.

The captured image for such a 3D scene can be written as

$$f(x, y) = \iint_S u(\mathbf{r}) p_r(x, y) dS, \quad (1)$$

where $u(\mathbf{r})$ is the light intensity at \mathbf{r} on S and $p_r(x, y)$ is the PSF which is determined by the aperture shape of the camera as well as the distance of the point \mathbf{r} to the camera plane. Please note that Eq. 1 can be extended to more generic scenes by specially treating occlusions.

Considering the imaging equation given by Eq. 1, DfD and CA approaches obtain depth information by estimating the correct PSFs throughout the image. In this paper, we employ two of those approaches that are introduced in [8] and [9], taking into account their demonstrated effectiveness. Zhou et al. [8] impose the locally fronto-planar assumption on the scene and thus model the imaging locally as a convolution

$$f = f_0 * p_d + \eta, \quad (2)$$

where $*$ is the convolution operation; f_0 is the latent sharp image, p_d is the PSF corresponding to depth d and η is noise. Then, depth information is obtained by using a deconvolution based maximum a posteriori formulation where image prior information is utilized. On the other hand, Favaro et al. [9] employ a set of training image patches as input for Eq. 1 and estimate the range spaces of the linear imaging systems corresponding to different depths. Depth estimation is then achieved by comparing the projections of the captured image patches onto these range spaces, without requiring inversion of system matrices. In implementation, both approaches require a set of PSFs sampled at discrete depths. While Favaro’s approach can work with a single captured image, Zhou’s approach is particularly effective with a pair of complementary masks on a single view. Indeed, the pair of masks proposed in [8] is designed to optimize both deconvolution and depth discrimination performance, which is hard to achieve with a single mask.

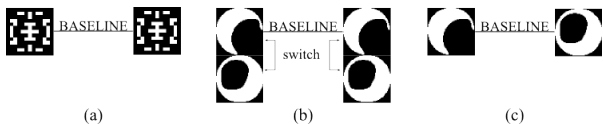


Fig. 1. Three camera setups. (a) two cameras with masks, e.g. the Levin’s mask, in stereo setup; (b) two cameras with one of Zhou’s mask pair (denoted as Zhou 1) in stereo setup, one more camera with other mask of Zhou’s pair (denoted as Zhou 2) is used to capture one more image on the right view; (c)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

Based on the motivation of merging CA and stereo for improved depth estimation performance, here in this paper we discuss about three different camera setups that employ Levin’s mask, as in Fig. 1(a), and Zhou’s pair of complementary masks, as in Fig. 1(b) and Fig. 1(c). Please note that we choose to utilize Levin’s mask and Zhou’s mask pair for their superb depth discrimination capability. Two questions lead us to come up with the first two setups shown in Fig. 1(a) and Fig. 1(b). One is whether using aperture

masks in a stereo camera seriously affects the performance of ordinary stereo matching, the other is whether CA can give us useful information where stereo matching fails. The simulation results, presented in Sec. III, demonstrate that if the captured stereo images are taken by cameras equipped with the same mask, the performance of ordinary stereo matching is not severely affected. In addition to that, more importantly, coded aperture can provide complementary information to stereo in some cases. These observations make the proposed setups attractive for the depth estimation problem.

In the third setup shown in Fig. 1(c), each camera is equipped with one of Zhou’s complementary masks. As it is evident from using different masks in stereo cameras, the motivation behind this setup is different from the one for the first two setups mentioned above. Indeed, the purpose is to take advantage of the effectiveness of Zhou’s complementary masks without requiring additional cameras (other than a stereo camera) as in the second setup, furthermore to have a single shot system which does not require changing the camera or mask as in Zhou’s approach [5]. For our proposed stereo setup shown in Fig. 1(c), the images are not taken from the same view as in the case of [5]. Therefore, we develop a variation of Zhou’s approach that employs the inherent relation between disparity and defocus on the stereo images. Intuitively, if the shifting between stereo images is done by the 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 to them. Ideally, there exists a one-to-one mapping between disparity and defocus, as employed in [12]. However, in most of 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-to-one. Theoretically, the correct disparity-defocus pair will give the minimum error. The proposed approach is summarized in Table I.

III. SIMULATION RESULTS

The synthetic scene that we use in our simulations includes three fronto-parallel planes 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 $3.3 \mu m$ pixel pitch, is put in the middle of the baseline of a normal stereo setup and focused at 1.5 m. The baseline length is set to 5 cm. The left and right view images are generated from the middle view image by means of shifting. For each pixel, the amount of shift 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

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 at different depths; each pair is denoted as (p_L^d, p_R^d) ;
STEPS:	
1 :	For each (p_L^d, p_R^d) in PSFs
2 :	Find its associated disparity range S ;
3 :	For each s in S
4 :	$f'_L = f_L(x - s, y)$
5 :	$\hat{F}_0 = \frac{F'_L \overline{P_L^d} + F_R P_R^d}{ P_L^d ^2 + P_R^d ^2 + C ^2}$
6 :	$E(p_L^d, p_R^d, s) = \sum_{i=L,R} \left f_i - \mathcal{F}^{-1} \left\{ \hat{F}_0 P_i^d \right\} \right ^2$
7 :	End for
8 :	End for
9 :	$(defocus, disparity) = \arg \min_{p_L^d, p_R^d, s} E, \forall pixel$
NOTATIONS:	
F :	the Fourier transform of f ;
\overline{F} :	the complex conjugate of F ;
C :	a matrix of noise to signal ratio [5];
\mathcal{F}^{-1} :	inverse Fourier transform operator.

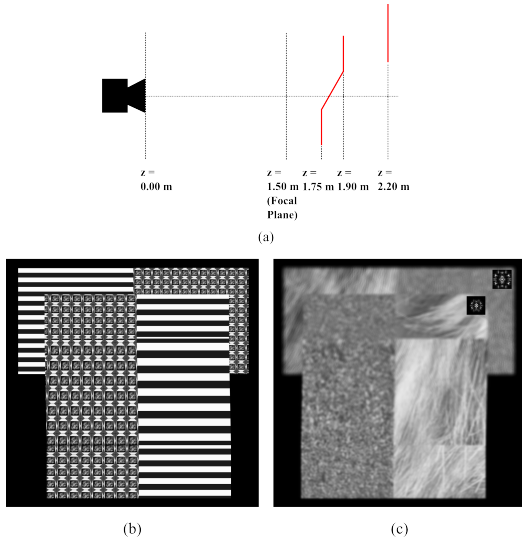


Fig. 2. Simulation environment. (a) the arrangement of the virtual camera and the scene; (b) a captured image on the left view with *ideal* pinhole aperture and the problematic texture; (c) a captured image on the right view with the Levin's mask and the good texture, together with two example PSFs at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

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 $\lambda = 534\text{nm}$. 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. We consider pinhole, circular mask, Levin's mask and one of Zhou's mask pair (at a time) cases. We also test the situation that 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.

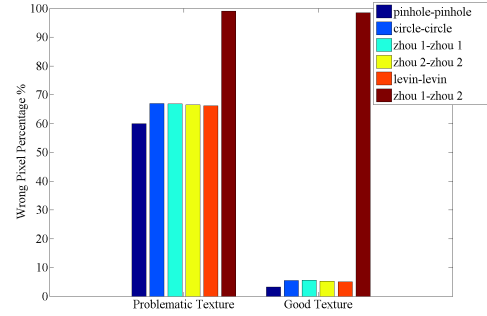


Fig. 3. The error percentages of stereo matching for different aperture masks, for both the problematic texture case and the good texture case. Please note that the pixels belonging to the black background are not considered in comparisons.

As shown in Fig. 3, in the good texture case, the effects of using the same aperture masks on the performance of stereo matching are not severe; while in the problematic texture case, stereo matching already fails even if no mask is used. In the case that two cameras are equipped with different masks (Zhou's pair), the performance of stereo matching decreases dramatically even with the good texture. That is actually inevitable, since different mask shapes result in different PSFs which affects the captured images differently and this complicates stereo matching. These observations 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 work independently with acceptable performance.

In order to see whether CA can give useful information where the stereo matching fails, we particularly consider the problematic texture case. We test two cases: Favaro's 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 the result of stereo matching in pinhole case, are shown in Fig. 4. Although the depth resolution provided by CA is lower than that of stereo matching, we get more reliable information as can be observed in the figure.

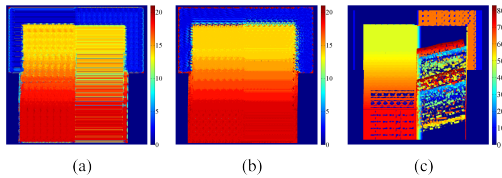


Fig. 4. Three depth maps produced by three approaches on the problematic texture case. (a) the defocus map obtained by the Favaro's approach with the Levin's mask; (b) the defocus map obtained by the Zhou's approach with Zhou's mask pair on the single view; (c) the disparity map obtained by stereo matching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

Zhou's approach gives the best result for the problematic texture as shown in Fig. 4. However, it is worth to point out that it takes two images from the same view, which might be a limitation in practice. Therefore, the proposed setup shown in Fig. 1(c) together with the algorithm given in Table I is of critical importance. 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. Moreover, it simultaneously provides a disparity map which has a significant improvement compared to the disparity map produced by pure stereo matching shown in Fig. 4(c). In spite of these improvements, it is worth to mention that the proposed setup suffers from the occlusion problem introduced by the stereo vision that is usually higher than the case in single view CA (the extent of occlusion is determined by the baseline in stereo, whereas it is determined by the aperture width in CA).

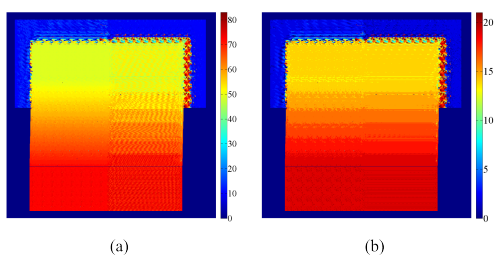


Fig. 5. The results of the proposed approach on the problematic texture case. (a) the disparity map; (b) the defocus map.

IV. CONCLUSIONS

Based on the presented preliminary results, for the proposed setups shown in Fig. 1(a) and Fig. 1(b), the degradation in the performance of stereo matching (compared to the case where no mask is used) is tolerable considering the superior

performance of CA approaches to stereo matching for problematic scene regions. Thus, having such two independently working systems, it is possible to improve stereo vision based depth estimation by merging the defocus and disparity maps produced by CA and stereo matching, respectively. Although it is not discussed in the paper, here it is worth to point out that CA can also be utilized to provide complementary information to stereo regarding occluded scene regions. Indeed, one of our future plan is to develop a merging algorithm which complements stereo based depth estimation with the information produced by CA in problematic texture and occlusion cases.

For the setup shown in Fig. 1(c), we propose a stereo version of Zhou's CA approach that produces a disparity map and a defocus map simultaneously. It provides convincing results even for the problematic textures we tested, where ordinary stereo matching fails. Furthermore, the setup possesses the one shot property which removes the requirement of changing cameras (or replacing masks) that presents in the case of single view, multiple mask CA.

In conclusion, the demonstrated preliminary results are promising and encourage further studies regarding the combination of stereo vision and CA for improved depth estimation.

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, May 2002.
- [2] A. P. Pentland, "A new sense for depth of field," *IEEE Trans. Pattern Anal. Mach. Intell.*, 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, USA: Springer 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'07*. New York, NY, USA: ACM, 2007.
- [5] C. Zhou and S. Nayar, "What are good apertures for defocus deblurring?" in *Computational Photography (ICCP), 2009 IEEE International Conference on*, Apr. 2009, pp. 1–8.
- [6] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," in *ACM SIGGRAPH'07*. New York, NY, USA: ACM, 2007.
- [7] A. Sellent and P. Favaro, "Optimized aperture shapes for depth estimation," *Pattern Recognition Letters*, 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 *Computer Vision (ICCV), 2009 IEEE 12th International Conference on*, Sept 2009, pp. 325–332.
- [9] P. Favaro and S. Soatto, "A geometric approach to shape from defocus," *IEEE Trans. Pattern Anal. Mach. Intell.*, 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. Artificial 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 Anal. Mach. Intell.*, 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 *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, Jun. 2013, pp. 209–216.
- [13] J. Goodman, *Introduction to Fourier Optics*, 2nd ed. McGraw-Hill, 1996.
- [14] W. Abbeeloo, "Real-time stereo vision," Master's thesis, Karel de Grote-Hogeschool University College, Belgium, May 2010.