DEPTH ESTIMATION FROM FOCUS AND DISPARITY

Arnav Acharyya, Dustin Hudson, Ka Wai Chen, Tianjia Feng, Chih-Yin Kan and Truong Nguyen

University of California San Diego
Department of Electrical and Computer Engineering
La Jolla, CA

ABSTRACT

This paper explores how focusing and defocusing can be used in combination with stereoscopy to enhance the accuracy and speed of depth estimation. The proposed method for combining focus information with stereo is inspired from how the human brain perceives depth using both. The outline of the method is to first estimate depth from focus using a high focal length camera. Next, we perform disparity estimation from the stereo images by only searching in the neighborhood of the disparity from focus (derived from depth from focus) for matching features. The matching is faster as the search range is only a small neighborhood of the expected disparity map. It also increases the accuracy because searching over a small range of high confidence rules out the possibility of wrong matching, especially if there are multiple areas with similar features. Depth from focus can be coarsely approximated by depth from defocus using multiple cameras focused at complementary distances. Hence, the proposed method also works with depth from defocus.

Index Terms— Stereoscopy, Focus, Defocus, Stereogram

1. INTRODUCTION

Depth estimation with image processing is becoming increasingly ubiquitous in the field of robotics. Of all the different ways to estimate depth from image processing, stereoscopy is the most common. Ideally, it employs a pair of horizontally separated cameras having parallel lines of sight, perpendicular to the line joining the cameras. For any object, the shift in the position of its image in the left and right camera is called disparity of that object. The disparity of an object is inversely proportional to the distance of that object from the camera lens, projected along the line of sight. So, depth can be computed from disparity, where disparity can be estimated by matching features to find correspondence between the pixels of left and right image [1] [2]. In contrast, depth from focus and defocus are derived from optical properties of lenses. *Depth from focus* is calculated by tuning the focus setting to the get the object of choice exactly in focus. In contrast, *depth from defocus* is calculated when the focus setting is not dynamically tunable. In that case, depth is computed from the amount of blur in the image and the optical properties of the lens. In [3], Xiong and Shafer proposed an efficient algorithm and precise blurring model for estimating depth from focusing and defocusing. Depth from focus can be approximated by depth from defocus if multiple cameras with different focus settings are employed. Then, it's as if these cameras collectively perform a low resolution scan on a focusing range. Perhaps, the best example of stereoscopy is the depth perception by the human brain. Moreover, illusions like stereograms reveal how the brain also uses focus information for performing stereoscopy. This is the motivation for using focusing/defocusing in combination with stereo vision, especially if we want to implement the same principle of human eyes in robots.

This work is supported by Qualcomm Inc.

2. RELATED WORK

There has been a lot of prior research related to estimating depth using stereo and defocusing. In [4], Gheta *et al.* modeled blurring as a space-variant linear system with a Gaussian point spread function (impulse response) and they minimized an additive combination of the energy (cost) functional for stereo and defocus. However, quantifying the blurriness to estimate depth is dependent on the presence of edges or rich texture in that portion. Texture-less portion of an image might not give any focus information. In [5], Takeda *et al.* used coded aperture for working around this problem. A coded aperture is when the aperture of the lens is masked with a known pattern. So, a blurred image of a point object is a patterned disk instead of a Gaussian circular disk. This helps quantify the blur more accurately even if there is no texture in the image.

3. DEPTH FROM FOCUS

A convex lens converges parallel beam of light at its focal length (f), which characterizes most of the optical properties of the lens. The distance of an object from the lens is called its object distance (v) and the distance (on the opposite side)

where the lens converges light emanating from the object is called its image distance (u). Any object can be thought of as a multitude of point objects. For a simple thin lens, the relation between object and image distance is given by the following *thin lens equation* [6].

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f} \tag{1}$$

The distance of the screen from the lens is called the screen distance (s). Now, for a point object if $u = s$, then the image is also a point and the object is said to be at focus. Otherwise the image is a blurred circular disk (uniform spread function) with a radius (σ) given by equation 2, where A is the radius of the aperture.

$$\sigma = \left| \frac{A(s-u)}{u} \right| \tag{2}$$

A focus ring of a camera affinely controls the position of the lens without changing its focal length (effective focal length for a compound lens). A schematic diagram of a simplified model of the camera is shown in Fig. 3. The depth of an object is defined as $Z = s + v$. If the focus ring setting is x, then $s(x) = mx + c$ for some m and c depending on the camera. We change x such that for a particular object, $s(x_0) = u$. Then Z can be given as $Z = s(x_0) + \frac{s(x_0)f}{s(x_0)-f}$ from equation 1. To find x_0 we must employ blur detection because the image is sharpest (blur-free) when $s(x) = u$ but blurred otherwise. The resolution of the blur detection algorithm specifies the *circle of confusion* (σ_c) which means that the algorithm outputs no-blur for all $\sigma \leq \sigma_c$. Now we find the (first order) sensitivity of depth from focus with error in blur detection. Later on, we asymptotically compare this sensitivity with that of stereo. It is noted that when an object is exactly in focus i.e. $s = u$ and $\sigma = 0$, then the following holds from equation 1.

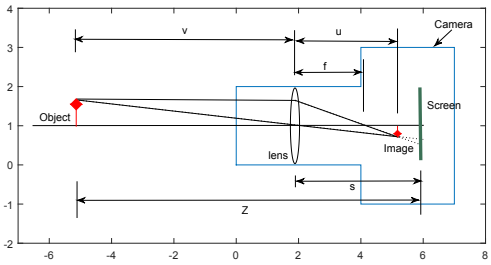


Fig. 1. Simple model of a camera

$$\frac{1}{Z-s} + \frac{1}{s} = \frac{1}{f} \tag{3}$$

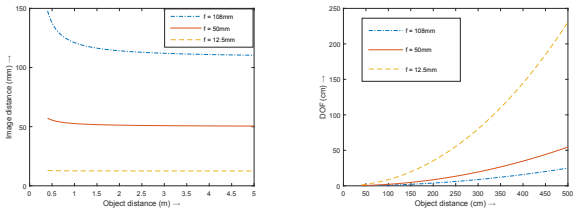
From the definition of spread function (σ) in equation 2, we observe that $\frac{d_+\sigma}{ds} = -\frac{d_-\sigma}{ds} = \frac{A}{s}$ at $u = s$. Combining this with equation 3 we obtain that

$$\left| \frac{dZ}{d\sigma} \right| = \frac{s}{A} \left(\left(\frac{Z-s}{s} \right)^2 - 1 \right) \tag{4}$$

But $s \to f$ as $Z \to \infty$ (from equation 3) implies

$$\left| \frac{dZ}{d\sigma} \right| \approx \frac{Z^2}{Af} \quad \text{if } Z \gg f \tag{5}$$

For a small circle of confusion (σ_c), the depth of field (DOF) can be approximated in first order as $DOF \approx \sigma_c \left| \frac{dZ}{d\sigma} \right|$ where s and Z follow the relation in equation 3.



(a) (b)

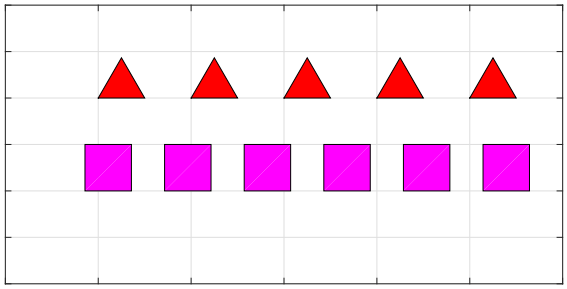
Fig. 2. (a) Image distance vs Object distance for different focal lengths. (b) DOF vs Object distance at different focal lengths

Fig. 2(a) delineates the thin lens equation 1 for different focal lengths and Fig 2(b) shows the growth of depth of field with object distance for different focal lengths. It is clear from Figs. 2(a-b) that higher focal length cameras can perform better to estimate depth from focus. However, higher focal length entails higher magnification ($\propto \frac{v}{s}$) and consequently, this high zoom factor makes it difficult to work with very high focal lengths because a significant portion of the objects can hardly fit in the field of view.

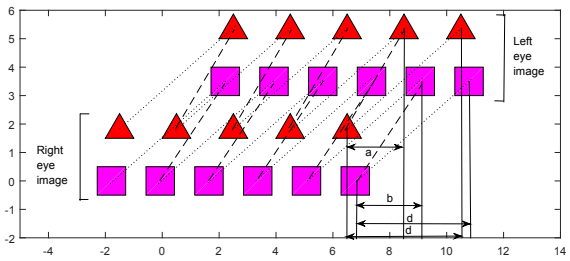
4. ROLE OF FOCUSING IN STEREOSCOPY IN HUMAN BRAIN

The human eyes cannot measure depth from focusing alone. It can be verified by trying to touch a hanging needle with one eye closed, in a single attempt, without knowing its position beforehand. We perceive depth from disparity using stereo vision. But, stereogram is an illusion which implies that the brain combines focus information with stereo vision to estimate the depth. Fig. 3(a) shows a simple stereogram. If we focus behind the plane of this image [1] we perceive a relative depth between the two geometric objects. It seems as if the square objects are closer (elevated) than the triangles. Ideally, the disparity of both the squares and the triangles are the same, denoted as d in Fig. 3(b). The ideal matching of objects between the left and right eye images is shown with

[1] Refer online to see steps on how to view stereograms.



(a)



(b)

Fig. 3. (a) Stereogram. (b) Schematic diagram of matching between left and right eye images

dotted lines in Fig. 3(b). But by intentionally focusing at a farther distance we tell the brain that the depth from focus is high, i.e. the disparity from focus is low. Because the brain takes this information into account, the matching happens between low-disparity clones, shown with dashed lines in Fig. 3(b). It is clear from Fig. 3(b) that because $b > a$, we perceive the squares to be nearer to us than the triangles.

5. PROPOSED IDEA

We propose a particular method for combining focus information with stereoscopy which serves as an explanation to the stereogram illusion. It speeds up the stereo-disparity estimation step and gives more accurate depth estimate. Before presenting this method we compare the error margin of stereoscopy and depth from focus, and set the rationale for combining the two.

5.1. Comparison of stereoscopy and depth from focus

Ideal stereo vision can estimate depth Z from disparity D by the formula $Z = \frac{Bf}{D} + f$, where the baseline B is the horizontal separation between the stereo camera pair and f is the focal length of each stereo camera. The sensitivity of the depth estimate with small error in disparity estimation is given by

$$\left|\frac{dZ}{dD}\right| = \frac{(Z-f)^2}{Bf} \implies \left|\frac{dZ}{dD}\right| = \frac{Z^2}{Bf} \text{ when } Z \gg f \tag{6}$$

If f_s is the focal length of the stereo pair and f_f and A are the focal length and aperture of the focusing lens, respectively, then the error margin of stereo ΔZ_s and that of depth from focus ΔZ_f are related as

$$\frac{\Delta Z_s}{\Delta Z_f} \approx \frac{Af_f \Delta D}{Bf_s \Delta\sigma} \quad \text{for } Z \gg \max\{f_f, f_s\} \tag{7}$$

where $\Delta\sigma$ is the circle of confusion σ_c and ΔD is the error in disparity estimation limited by accuracy of the feature matching algorithm. In general, the sub-pixel precision of feature matching is better than that of blur detection. Moreover, for general stereo setup the baseline B is also considerably bigger than the aperture A of the focusing camera. Consequently, in spite of $f_f > f_s$, we get $Af_f\Delta D < Bf_s\Delta\sigma$, in practice. This implies that stereoscopy works better for far away objects ($Z \gg f_f$). But in close range, depth from focus is more accurate than stereo. Furthermore, in real world images, often there is considerable repetition of features, like parallel edges, similar texture and so on. As the result, ΔD may have sporadic high values at some portions of the scene. Here, depth from focus can be used with stereo to correct these sporadic errors, as well as provide accurate information in close range. We also note another interesting property of the sensitivity of disparity from focus with error in blur detection. Disparity from focus is defined as $D_f = \frac{Bf_s}{Z_f - f_s}$ where Z_f is the depth from focus. The sensitivity, given by equation 8 is bounded

$$\left|\frac{dD_f}{d\sigma}\right| \approx \frac{Bf_s}{As} < \frac{Bf_s}{Af_f} \tag{8}$$

So, for a small circle of confusion σ_c, the error in disparity is bounded (in first order) and given by $\Delta D_f \approx \left|\frac{dD_f}{d\sigma}\right| \sigma_c < \frac{Bf_s\sigma_c}{Af_f}$

5.2. Proposed method for combining stereo and focus

A general disparity estimation algorithm [7] [8] [9] [10] finds correspondence between the left and right stereo images for every small window by matching features. In a pair of stereo-calibrated images, for every small window in the left image, its corresponding match is searched over all horizontal shifts within a given range inside the right image. This range is called the disparity search range or simply, disparity range R. Let $W_L(x, y)$ be a window centered at (x, y) in the left stereo image, then for each pixel (x, y), $W_L(x, y)$ is matched with all of $W_R(x - \Delta, y)$ where $\Delta \in \{0, \ldots R - 1\}$. All disparity estimation algorithms involve conducting such search in a disparity range R. Our proposed method is two-fold:

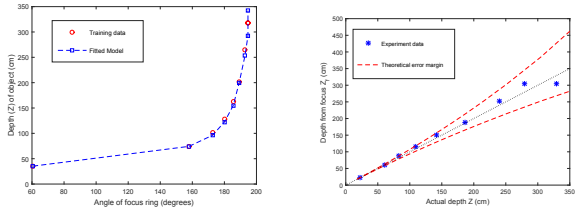
1. We obtain the depth from focus $Z_f(x, y)$ for all (x, y) in the left stereo image. Then we compute the corresponding disparity from focus $D_f(x, y) = \frac{Bf_s}{Z_f(x,y) - f_s}$.
2. We employ regular stereo disparity estimation with the only modification that for any pixel (x, y), a window

$W_L(x,y)$ is matched with all windows $\in W_R(x - D_f(x,y) \pm \Delta_d, y)$, where $\Delta_d = \sup\{\Delta D_f\} = \frac{Bf_s\sigma_c}{Af_f}$

Because the search range is reduced from R to $2\Delta_d$, the computation of disparity is faster in the proposed method.

6. RESULTS

In our experiments, we use a 70 mm camera for estimating depth from focus and two symmetric 12.5 mm cameras for stereo. We model the optics of the camera according to section 3. Fig. 4(a) validates the credibility of the assumed model. Fig. 4(b) shows that the error of depth from focus increases with object distance, as expected from theory.



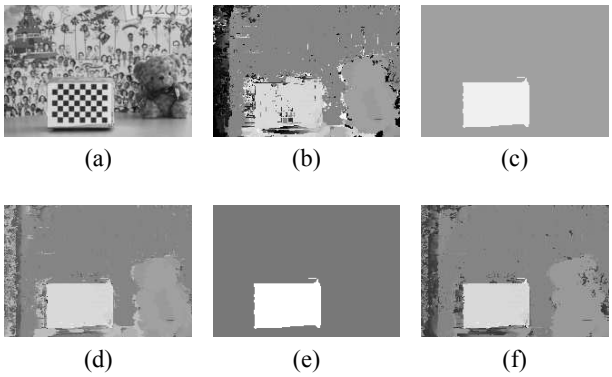
(a) (b)

Fig. 4. (a) Model fitted with data. (b) Estimated depth from focus

(a) (b) (c) (d)

Fig. 5. (a) Tsukuba image. (b) Ground truth disparity. (c) Regular disparity estimation. (d) Disparity estimation with proposed method.

The method in [11] has been used as the general framework for (regular) disparity estimation in the experiments. Fig. 5(b) is the ground truth disparity of the Tsukuba image (Fig. 5(a)). The highest disparity (lamp) of the Tsukuba image is $14px$. The regular disparity estimation algorithm of [11] with disparity range, $R = 20px$ takes 8.071 secs to compute the disparity, given in Fig. 5(c). Fig. 5(d) is the disparity calculated using the proposed method. Here the ground truth disparity map Fig. 5(b) is taken as the disparity from focus and the search is conducted with $2\Delta_d = 8px$. This takes 5.992 secs to complete. Even though that the ground truth disparity is used as disparity from focus in this simulation, it shows that besides taking shorter time to compute, the resulting disparity map in Fig. 5(d) is more accurate than that in Fig. 5(c). Fig. 6(a) has maximum disparity of $31px$ at the front of the box. The regular disparity estimation algorithm with disparity range, $R = 40px$ takes 4.735 secs to compute



(a) (b) (c)

(d) (e) (f)

Fig. 6. (a) Teddy bear image. (b) Regular disparity. (c) Disparity from defocus. (d) Disparity estimation with proposed method by combining with (c). (d) Erroneous disparity from defocus. (e) Disparity estimation with proposed method by combining with (d).

the disparity, given by Fig. 6(b). Fig. 6(d) is the disparity calculated by taking Fig. 6(c) as the initial disparity from defocus and then searching in the neighborhood of $\Delta_d = 8px$. This takes 3.513 secs to complete. Fig. 6(f) is the disparity calculated by taking the erroneous disparity map of Fig. 6(e) as the initial disparity from defocus and then searching in the neighborhood with $\Delta_d = 8px$. This takes 3.54 secs to complete. It is observed that using the proposed method, disparity estimates in Figs. 6(d,f) are much better than the regular disparity estimate of Fig. 6(b). Furthermore, although the disparity from defocus in Figs. 6(c,e) are little different, the final disparity estimates, Figs. 6(d,f) are almost same. This suggests that the proposed method is robust towards error in the initial disparity from focus/defocus.

7. CONCLUSION

A new method has been proposed for combining focus information with stereo vision, inspired from the human brain, to achieve better performance of depth estimation. The method is efficient which reduces the time for disparity estimation. Experimental results show that the proposed method performs better than regular disparity estimation. We also established the rationale, theoretically. However, it is difficult to work with high focal-length cameras because of their narrow field of view (high zoom). It involves image registration to find correspondence with stereo images. To account for the error in image registration we can increase Δ_d by the appropriate amount. In future, all these practical areas should also be studied to better utilize the information from focus/defocus.

8. REFERENCES

[1] M. Drumheller and T. Poggio, "On parallel stereo," in *Robotics and Automation. Proceedings. 1986 IEEE International Conference on*, Apr 1986, vol. 3, pp. 1439–1448.

[2] Jean-Philippe Pons, Renaud Keriven, and Olivier Faugeras, "Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score," *International Journal of Computer Vision*, vol. 72, no. 2, pp. 179–193, 2007.

[3] Yalin Xiong, Steven Shafer, et al., "Depth from focusing and defocusing," in *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR'93., 1993 IEEE Computer Society Conference on*. IEEE, 1993, pp. 68–73.

[4] Ioana Gheta, Christian Frese, Michael Heizmann, and Jürgen Beyerer, "A new approach for estimating depth by fusing stereo and defocus information.," *GI Jahrestagung (1)*, vol. 7, pp. 26–31, 2007.

[5] 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*, June 2013, pp. 209–216.

[6] Eugene Hecht, *Optics*, Pearson Education Limited, Edinburgh Gate, England, 4th edition, 2014.

[7] Heiko Hirschmüller and Daniel Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 9, pp. 1582–1599, 2009.

[8] Takeo Kanade and Masatoshi Okutomi, "A stereo matching algorithm with an adaptive window: Theory and experiment," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 16, no. 9, pp. 920–932, 1994.

[9] A. Fusiello, V. Roberto, and E. Trucco, "Efficient stereo with multiple windowing," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, Jun 1997, pp. 858–863.

[10] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, June 2005, vol. 2, pp. 807–814 vol. 2.

[11] Zucheul Lee, J. Juang, and T.Q. Nguyen, "Local disparity estimation with three-moded cross census and advanced support weight," *Multimedia, IEEE Transactions on*, vol. 15, no. 8, pp. 1855–1864, Dec 2013.