O. Ghita and P.F. Whelan (1999), "Real time 3-D estimation using depth from defocus", Proc. Irish Machine Vision and Image Processing Conference 1999, Dublin City University, September 8-9, pp 167-181.

Real time 3-D estimation using depth from defocus

Ovidiu Ghita

Vision Systems Laboratory
School of Electronic Engineering
Dublin City University
Dublin 9, Ireland

Email: *ghitao@eeng.dcu.ie*

Paul F. Whelan

Vision Systems Laboratory
School of Electronic Engineering
Dublin City University
Dublin 9, Ireland

Email: *whelanp@eeng.dcu.ie*

Real time 3-D estimation using depth from defocus

Abstract. In the recent times a great amount of interest has been shown in the area of range data acquisition for supporting 3-D scene interpretation. This paper presents an interesting approach to obtain depth information using defocusing techniques. This method involves calculating distance to points in a scene using the relative blurring between two images detected with different focal settings. These images are obtaining by splitting the image of the scene captured with two CCD sensors with a known physical distance between sensor planes. The proposed algorithm uses only simple filters and operators to compute the amount of defocus according to the optical settings. Nevertheless, textureless images might introduce significant errors and in order to minimise this problem a practical solution is to project a structured light on the scene. Magnification is maintained invariant to changes in focal settings because in this implementation a telecentric lens is used. Compared to other methods (e.g. stereo and motion parallax) which require solving the correspondences between different features and also suffer from occlusions or missing parts, this approach has many advantages such as reliable accuracy at low computational cost and provides easy camera calibration. This approach has been implemented and successfully tested on several real world scenes.

Keywords: depth from defocus, image blurring model, active illumination, inverse filtering, real- time.

1 Introduction

Recovering the depth information of the scene is one of the most important tasks in machine vision. Depth information plays a key role in machine vision and has a strong relationship with the real world in robotic applications. The 3-D information can be obtained in various ways. Several 3-D vision systems have been developed to solve a specific task while others are more general and consequently more complex. Among other approaches for 3-D recovery, depth from defocus (*DFD*) techniques has recently attracted a great amount of interest. Originally developed by Pentland [1987], the depth from defocus method uses the direct relationship between the depth, cameras parameters, and the degree of blurring in several images (in the current implementation only two are used). In contrast with other techniques such as stereo or motion parallax where solving the correspondences between different features is a major disadvantage, depth from defocus relies only on simple local algorithms. However, these methods are complementary. Stereo and motion parallax is used for outdoor scenes where the depth discontinuities are important while depth from defocus performs better for indoor scenes where the target is situated nearby. Another popular method used in 3-D estimation is based on triangulation. In terms of precision methods based on triangulation appear to perform better but the major drawback is the amount of computation involved. Some speed improvements have been obtained using grey or colour-coded patterns. Ideally the number of independent coloured stripes should be large and geometrically very dense but in this case the colour-structured pattern is very difficult to be manufactured. Also, different reflection properties of the object surface can introduce some errors in 3-D estimation (i.e. when the colour of the stripe is the same as the colour of the object's surface). An interesting method to generate colour-structured pattern is proposed by Chen et al. [1997]. The main idea here is to design a pattern that have strong contrast at the borders of any two adjacent stripes and the correlation between any two segments of a consecutive sequence of light stripes should be as small as possible in order to minimise the mismatch. Despite these achievements a real-time sensor based on triangulation has not yet been implemented.

This paper addresses the implementation of a real-time 3-D sensor based on depth from defocusing. As we mentioned before this method requires only two images acquired using different focal settings. This method performs badly in case if the scene does not provides high frequency textures. A practical solution for this problem is to project a structured light on the scene while the scene will have in this case a dominant frequency for texture (Nayar et al.[1995]). Furthermore, using active illumination minimises the shadows when all surfaces are visible. Xiong and Shafer [1995] propose a novel approach to determine dense and accurate depth estimation based on maximal resemblance estimation. This implementation uses a large bank of filters with a different window size tuned for all dominant texture's frequency. Using a large bank of filters makes this approach unsuitable for a real-time implementation. Subbarao and Surya [1994] proposed the Spatial-Domain Convolution/Deconvolution Transform (S Transform) when they try to estimate the depth using an analysis in frequency domain.

This implementation does not perform as well as those mentioned previously. Watanabe and Nayar [1995] proposed a small bank of broadband rational filters able to handle arbitrary textures. This implementation is simple and performs reasonable well even in case of weak textures. This approach represents a certain improvement but still fails when the scene is textureless. Therefore, considering these aspects for this present implementation the optimal solution is using structured (active) illumination. An important problem is determining the illumination pattern. Nayar et al. [1995] proposed a method for optimisation in the Fourier domain. The optimal pattern maximises the sensitivity of the focus measure in order to enhance the high spatial resolution. Keeping in mind that the CCD sensor can be approximated with an array of square elements (cells), thus the optimal pattern is a rectangular spatial grid (chessboard). The next step is tuning this filter with the CCD parameters (distance between two adjacent cells). Another advantage given by using active illumination is minimising the shadows.

Asada et al. [1998] using the *reversed projection blurring* (RPB) model. The RPB model is a technique used by ray tracing algorithms widely used in computer graphics. This model uses photometric properties of occluding edges when the object's surface behind nearer object is partially observed. Therefore the blurring model using convolution becomes inconsistent around the occluding edges. To compensate this problem they use the radiance of the near and far surfaces and then is mapped the occluded region. In this implementation the occluded region is assign to be equal to that from a nearer side of the depth discontinuity that in most of the situations is a correct assumption.

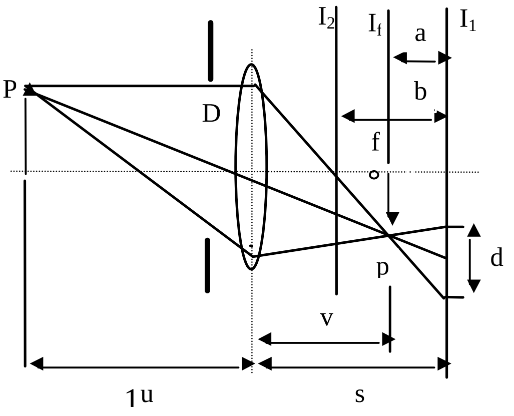
2 Theoretical approach of depth from defocus

The depth from defocus method uses the relationship between the depth, camera parameters, and the degree of blurring between near and far focused images. In other words, depth from defocus means calculating the depth of the scene in the image from the degree of image blurring.

Let *P* be a point that belongs to an object's surface and *p* the focused point refracted by the lens. The relationship between the object distance *u*, focal length *f* and image formation distance *v* is given by the lens law.

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

The Figure 1 shows the optical settings and the basic image formation geometry for convex lens.



P: Object point
p: Image point
I$_1$: Image plane for camera 1
I$_2$: Image plane for camera 2
I$_f$: Image in focal plane
D: Aperture of lens
f: Focal length
s: Sensor distance (camera1)
u: Object distance
v: Focal distance

Figure1: The camera geometry and the image formation.

If the CCD sensor is not placed in the focal plane the image is distributed over a circular patch on the sensing element. The diameter of the blur circle *d* is given by use of similar triangles:

$$\frac{D/2}{v} = \frac{d/2}{s-v} \Rightarrow \quad d = Ds\left(\frac{1}{v} - \frac{1}{s}\right) \Rightarrow \quad d = Ds\left(\frac{1}{f} - \frac{1}{u} - \frac{1}{s}\right) \qquad (2)$$

The blurring effect is seen as a convolution between the focused image and the blurring function

$$I(x,y) = \int\int I_0(u,v)h(x-u, y-v)dudv \qquad (3)$$

where I_0 is the focused image and *h* is the blurring function.

The blurring function also called as the *Point Spread Function* (*PSF*) has the following expression:

$$h_p(x,y) = \begin{cases} \dfrac{4}{pd^2} & if \quad x^2 + y^2 \leq \dfrac{d^2}{4} \\ 0 & otherwise \end{cases} \qquad (4)$$

where h_p is called the pillbox function and can be seen as a cone of light emerging from the lens with the point of the cone in focal plane. If the sensor plane is shifted from the focal plane then, cuts the cone in a circle with the diameter *d*.

If within this circle the brightness is not uniform the *PSF* is better approximated by a *two dimensional Gaussian* function (Pentland [1987]):

$$h(x,y) = \frac{1}{2ps^2} e^{-\frac{x^2+y^2}{2s^2}} \qquad (5)$$

where s is the standard deviation of the distribution of the *two-dimensional Gaussian*.

In practice we can assume that the brightness is constant over a region of the image projected onto CCD element, the result is an invariant shift from the focal plane. The blurring is better modelled by the two-dimensional Gaussian than the blur circle (another advantage is that the Fourier transform of a Gaussian is also a Gaussian). If the brightness is uniform over a small region of the image (this assumption approximate very well the practical case) σ is proportional to *d*

$$s = kd \qquad (6)$$

where *k* is constant of proportionality characteristic for every camera and can be determined from a previous camera calibration.

Unless we know *a-priori* information about the scene one image is not enough to estimate the depth (see equation (2) where *d* and *u* are unknown). Therefore, minimum two images acquired with different camera settings are necessary. Clearly, are two distinct options, either the aperture *D* is maintained constant and the sensor position *s* is modified (Nayar et al. [1995]) or the sensor is fixed and the aperture is changing when the images are taken (Pentland [1987], Subbarao and Surya [1994]). The first case has an important advantage, because it does not require any user intervention while the images are acquired but unfortunately have different magnification caused by focusing.

An elegant and effective solution was proposed by Watanabe and Nayar [1995] by using *telecentric* lens. They suggest an optical solution to obtain constant magnification. It is well known that using *telecentric* optics magnification remains constant despite the focus changes. Most of the popular commercial lenses can be transformed to *telecentric* only by adding a small extra aperture. The aperture will be placed in the front focal plane of the lens. Using *telecentric* lens, they demonstrate that the magnification changes can be reduced to as low as 0.03%. Because the aperture has to be small the only drawback of this approach is the severe reduction in brightness. Therefore, to compensate this issue is necessary to use a brighter source of illumination.

The second possible implementation is not hampered by this issue but the depth estimation by far is not as precise. Certainly, a third possibility can consider the modification of both parameters but the depth estimation is not significantly improved.

3 Estimating the depth of the scene

As we mentioned above, the depth can be estimated by taking a small number of images (usually two) under different camera or optical settings. Subbarao and Surya [1994] proposed the Spatial-Domain Convolution/Deconvolution Transform (S Transform). They modelled an image as a cubic polynomial in spatial domain and the image is developed in a *Taylor* series. Since the *PSF* is a circularly symmetrical function therefore, the final expression is greatly simplified

$$f(x,y) = g(x,y) - \frac{s^{2}}{4}\nabla^{2} g(x,y) \qquad (7)$$

where *f* is the focused image, *g* is the defocused image, σ is the standard deviation for *PSF* and ∇^2 the Laplacian operator. The equation (7) represents the deconvolution formula. If are taken two images g_1 and g_2 under different camera settings and the term *f (x,y)* from the first equation is replaced in the second equation, the result is a simple expression:

$$g_1(x,y) - g_2(x,y) = \frac{1}{4}(s_1^2 - s_2^2)\nabla^2 g, \qquad \nabla^2 g = \frac{\nabla^2 g_1 + \nabla^2 g_2}{2} \qquad (8)$$

From equation (8) can be observed that no terms depend on scene’s texture frequency. Furthermore the depth can be estimated using the difference between the standard deviation of the near focused image (g_1) and far focused image (g_2). The use of the Laplacian as a focus operator is very convenient because has a simple kernel but the depth map resulted by using equation (8) is accurate only if the depth discontinuities in the scene are important. Also, if the scene has only weak texture the depth estimation is poor. Certainly, in order to obtain a dense and robust depth map a more sophisticated approach for modelling *PSF* has to be involved.

Nevertheless, the focus operator plays an important role in depth estimation stage. Therefore, this goal of this paper is to study the accuracy of depth estimation when are used different operators. Because the defocus function (*PSF*) acts like a low pass filter therefore the focus operator has to perform inverse filtering.

The next step is determining depth from two images. The simplest solution is to use the ratio between the defocus function of the near focused image and far focused image. Nayar et al. [1995] proposed a normalised ratio M/P that is a monotonic and bounded function.

$$\frac{M}{P} = \frac{g_1(x,y) - g_2(x,y)}{g_1(x,y) + g_2(x,y)} = \frac{H\ (p,q,s_1) - H\ (p,q,s_2)}{H\ (p,q,s_1) + H\ (p,q,s_2)} \tag{9}$$

where H is the Fourier transform of the *PSF* and σ_1 is the standard deviation of near focused image (σ_2 is the standard deviation for far focused image).

Now, the depth can be determined by mapping the M/P ratio into a look-up table that returns directly the depth.

4 Active illumination

If the scene is highly textured the depth estimation will be precise and reliable. Unfortunately, if the scene has a weak texture or is textureless (like a blank sheet of plain paper) the depth recovery is very far from accurate. An effective and relatively simple solution is based on the use of structured (active) light. Initially, suggested by Pentland et al. [1994], then Nayar et al.[1995] developed an symmetrical pattern as a rectangular spatial grid optimised for a specific type of camera. Therefore, the illumination pattern has a single dominant frequency in direct correlation with the pattern’s arrangement for transparent and opaque regions. When the structured light is projected onto scene the spectrum will have the same dominant frequency.

The resulting pattern is very dense and rotational symmetrical in order to obtain spatial invariance. A problem caused by using a dense spatial pattern is the reduction in illumination caused by the filter’s opaque regions, thus a very powerful source of light is required. Nevertheless, a very precise pattern is difficult to fabricate and in our testing we discovered that this issue is not as very restrictive as it seems. For the current implementation a simple stripes grid (10 lines/mm) used in Moire contour detection was used.

5 The focus operator

The goal of this operator is determining the defocus function (σ) by inverse filtering near and far focused images. Our efforts in this paper were concentrated in evaluating the efficiency of different focus operators. Because the blur circle is uniform only for small regions, the kernel of focus operator has to be small in order to preserve locality but on the other hand the windowing introduces supplementary errors. Xiong and Shafer [1994] proposed a solution to select the window size for Gabor filters. They used a simple criterion when the window size is selected to be as small as possible while the error caused by noise and windowing is smaller than a preset value. Aside from window size every focus operator must be rotationally symmetric and must not respond to any DC component (a DC component can be a change in image brightness). This condition is satisfied if the sum of all elements of the focus operator is equal to zero.

Watanabe and Nayar [1995] suggested an approach based on the use of rational filters. They proposed a method to compute a set of broadband rational operators. The first operator performs prefiltering (for removing DC components) and then another three operators are involved in depth estimation. Finally, the depth errors caused by spurious frequencies are minimised by applying a smoothing operator.

This paper investigates the performance of Laplacian (4 and 8 neighbourhood) and rational operators (3x3 and 7x7 kernels). The 3 by 3 operators are shown in Figure 2 and followed by the 7 by 7 operator in Figure 3.

$$\begin{array}{ccc} \begin{matrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{matrix} & \qquad \begin{matrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{matrix} & \qquad \begin{matrix} 0.55 & -1 & 0.55 \\ -1 & 1.8 & -1 \\ 0.55 & -1 & 0.55 \end{matrix} \\ \\ \text{(a)} & \qquad \text{(b)} & \qquad \text{(c)} \end{array}$$

Figure 2: Focus operator kernels. (a) Laplacian (4), (b) Laplacian (8) and (c) rational operator (3x3)

– 0.143	– 0.1986	– 0.1056	– 0.07133	– 0.1056	– 0.1986	– 0.143
– 0.1986	– 0.1927	0.01795	0.07296	0.01795	– 0.1927	– 0.1986
– 0.1056	0.01795	0.2843	0.4601	0.2843	0.01795	– 0.1056
– 0.07133	0.07296	0.4601	0.6449	0.4601	0.07296	– 0.07133
– 0.1056	0.01795	0.2843	0.4601	0.2843	0.01795	– 0.1056
– 0.1986	– 0.1927	0.01795	0.07296	0.01795	– 0.1927	– 0.1986
– 0.143	– 0.1986	– 0.1056	– 0.07133	– 0.1056	– 0.1986	– 0.143

Figure 3: Focus operator kernel of an rational operator (7x7).

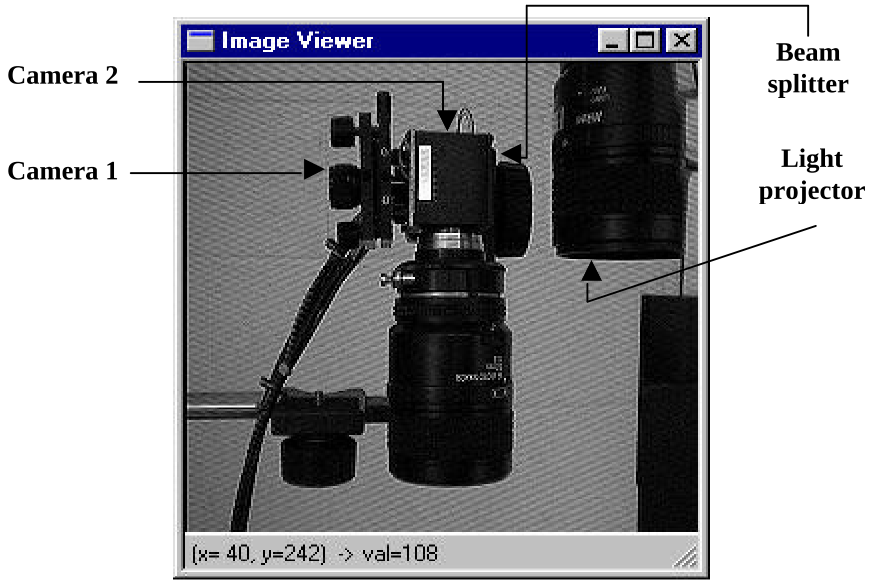
Because the image is discrete the focus operator will introduce errors (apart those caused by windowing). Furthermore, supplementary errors are caused by misalignment between the cells of the CCD sensor and the illumination pattern. In order to minimise the problems mentioned above, a post-filtering operator is used after the focus operator is applied to near and far focused images.

6 Physical implementation

The main goal of this implementation is to build a real-time depth estimator. Therefore, the near and far focused images have to be acquired in the same time. For this purpose were used two OFG VISION*plus* – AT frame grabbers. The scene is imaged using an AF MICRO NIKKOR 60mm F 2.8 (Nikon). Between the NIKKOR lens and the sensing equipment (CCD sensors) is placed a 22mm beam splitter cube. Then, the near and far focused images are acquired using two low cost 256 by 256 CCD sensors VVL 1011C (VLSI Vision Ltd.). These sensors are precisely fixed to ensure that one will acquire the near focused image and the other the far focused image. The physical displacement between these sensors is approximately 0.8mm.

The structured light is projected onto scene using MP-1000 Moire Projector with MGP-10 Moire gratings (stripes grid with density of 10 lines/mm). The lens attached to the projector is the same type that one used to image the scene. All the sensing equipment required by this implementation is at low cost (except lenses) and furthermore the calibration procedure is relatively simple. The set up involved in this present implementation is described in Figure 4.

When the images are acquired are necessary few operations to determine the scene's depth map. For sake of computation efficiency the depth is estimated directly from g_1 and g_2 using a pre-computed look-up table (Figure 5).



(x= 40, y=242) -> val=108

Figure 4: The 3-D sensor and its principal components.

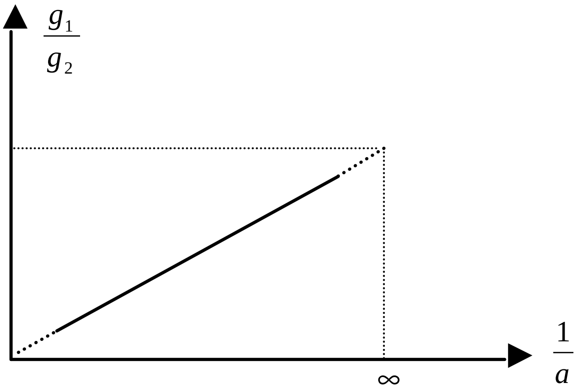


Figure 5: The defocus function.

Unfortunately, this function is not bounded but this is not a major drawback because the extreme values cannot be reached. Is well known that with a real aperture is almost impossible to obtain a perfect focused image (more details in Krotkov [1989]). A simple solution of avoiding the case g_2 to be equal with zero, a simple solution is to add a small constant value to g_1 and g_2. As we mentioned before this function can be evaluated using the ratio (g_1-g_2) / (g_1+g_2), the defocus function being bounded in this case but for this implementation while the depth is investigated only within a small range (0-9 cm) was proven not being sensitive enough. Certainly, the defocus function illustrated in Figure 4 is more sensitive to external noise, therefore the depth was smoothed by using a 3 by 3 smoothing operator.

The flowchart illustrated in Figure 6 describes these operations.

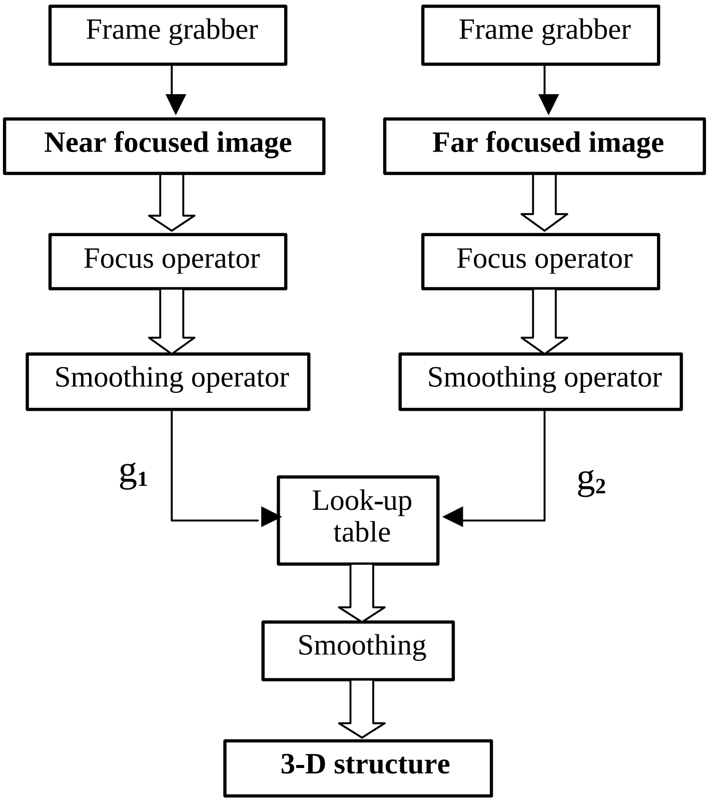
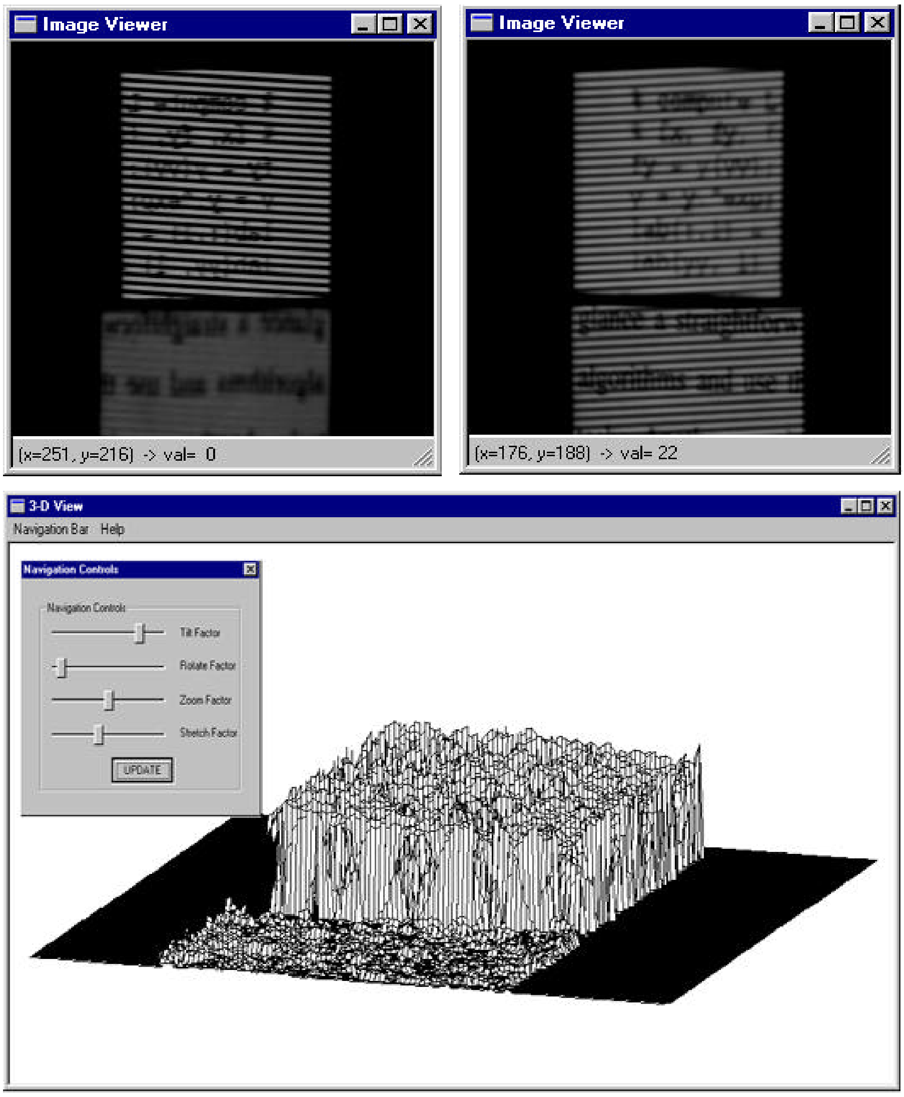


Figure 6: Data flow during the computation process.

Like any other range sensor, apart from main operations required by depth estimation, a key step is represented by gain calibration and minimising the errors caused by the imperfection of optical equipment. The implementation presented above computes the depth map (256 by 256) in approximately 95 ms on a Pentium 133 MHz (the time required by graphical interface is not included).

7 Experiments and results

In order to verify the efficiency of this range sensor it was tested on several indoor scenes. Firstly, this sensor was tested on simple targets like planar surfaces, then on scenes with complex scenario. The accuracy and linearity of this sensor is estimated for a distance within one cubic meter workspace. Figure 7 shows the depth recovery for two planar objects situated at different distances in front of the sensor.

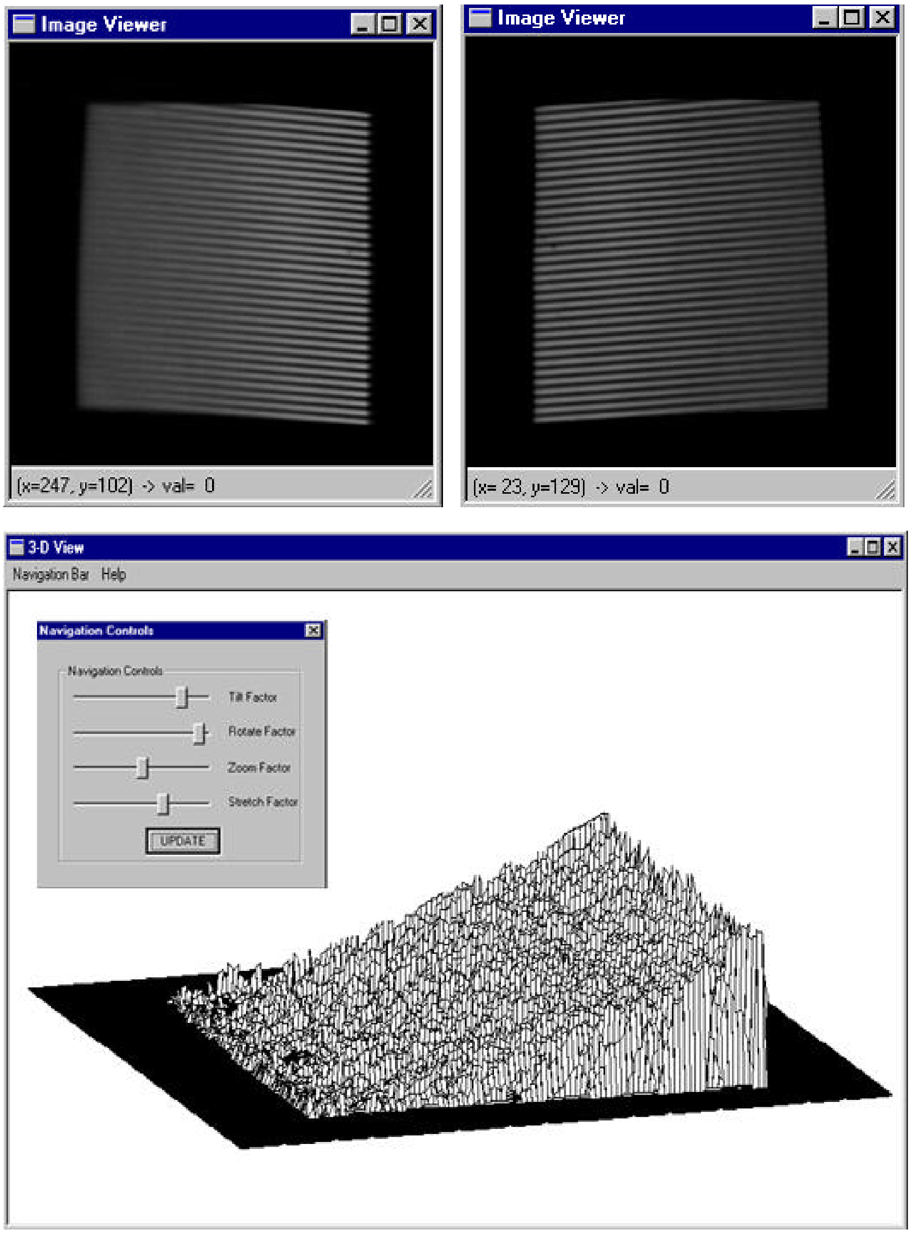


(x=251, y=216) -> val= 0

(x=176, y=188) -> val= 22

Figure 7: Near and far focused image and the depth estimation for two planar objects situated at different distance from sensor.

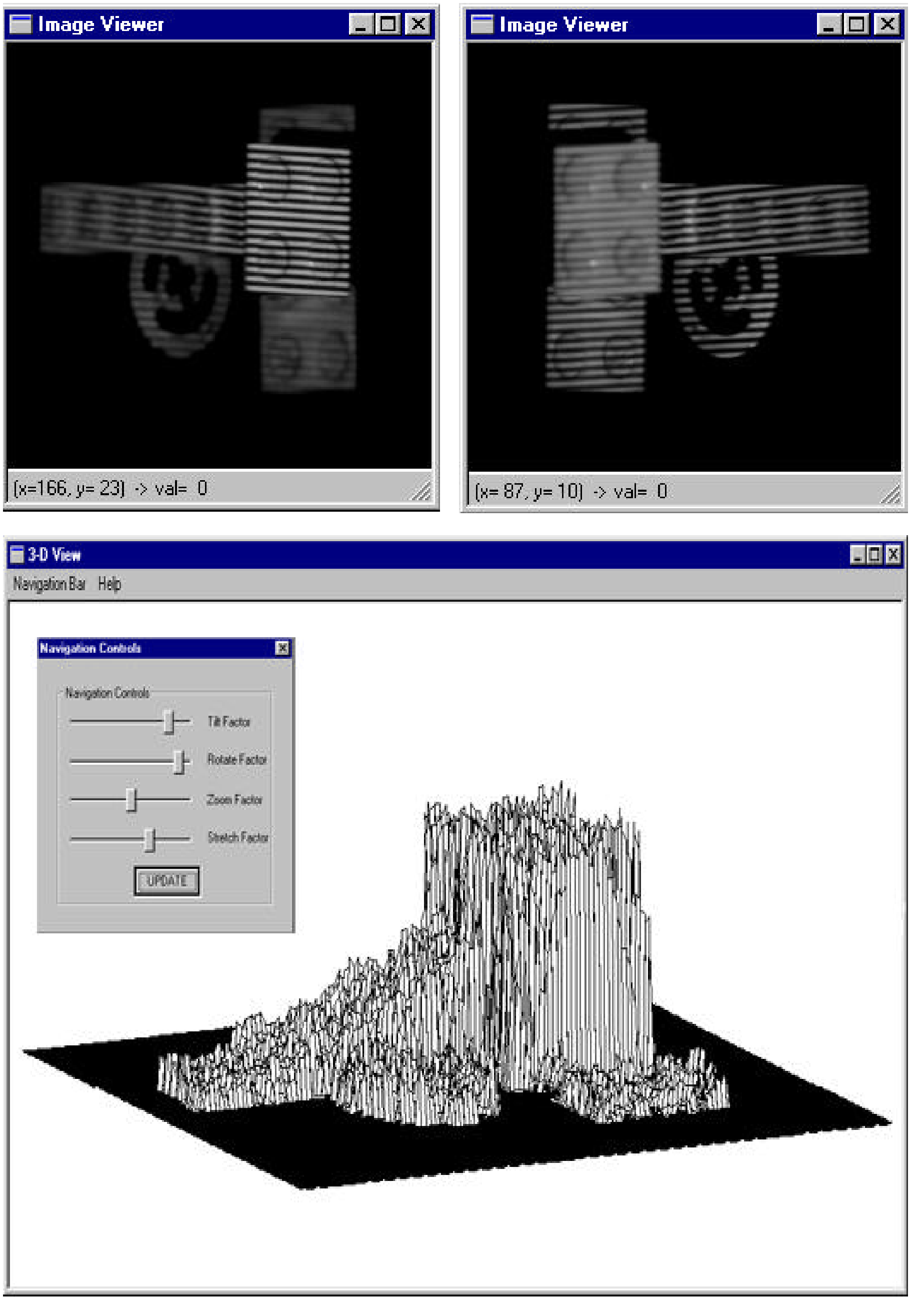
Figure 8 shows the depth map for a slanted planar object and Figure 9 shows a more complex scene containing LEGO objects with different shape and a large scale of colours.



(x=247, y=102) -> val= 0

(x= 23, y=129) -> val= 0

Figure 8: Near and far focused images and depth recovery for a scene containing a slanted planar object



(x=166, y= 23) -> val= 0

(x= 87, y= 10) -> val= 0

Figure 9: Near and far focused images and depth recovery for a scene containing various LEGO objects.

The accuracy of this sensor is estimated for a distance within one cubic meter workspace. For these scenes the lowest accuracy is 3.4% normalised in agreement with the distance from sensor. This accuracy is reported for both textured and textureless non-specular objects. We tried to identify an optimal solution for focused operator. As we mentioned in section 5 four focus operator were used. The best results in respect with the gain were obtained for a 7 by 7 rational operator but the depth estimation is not very linear. The results were more linear when the Laplacian (4) and the 3 by 3 rational were used as focus operator but the discontinuities in depth were not as well recovered. A trade-off between gain and linearity was given by Laplacian (8).

8 Conclusions

This paper presented the implementation of a real-time depth sensor. In comparison with stereo technique, the *DFD* method does not suffer from the correspondence problem. Furthermore, the *DFD* approach is not affected by occlusion or missing parts, therefore it can be used as a ranging method for various applications. The consistency between theory and experimental results has indicated that our implementation is an attractive solution to estimating the depth fast and accurate.

In contrast to other implementations based on defocusing where the depth range is relatively large, we proposed a solution to estimate depth within a small range (between 0 and 9cm). Furthermore, this present approach has another advantage over other implementations suggested by Pentland et al. [1994], Nayar et al. [1995] because does not contain any sensitive equipment to movements or vibrations, therefore can be easily involved in robotics applications.

Because *DFD* methods perform badly for textureless objects, hence the active illumination was identified as being the key issue for this implementation. The depth estimation can be further improved by using a camera with higher resolution and re-designing the illumination pattern and the focus operator.

Acknowledgements

This research was supported by Motorola BV, Dublin, Ireland.

References

1. Asada N., Fujiware H. and Matsuyama T., “Seeing behind the scene: analysis photometric properties of occluding edges by the reversed projection blurring model”, *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), vol. 20, no. 2, pp. 157-166, July 1998.*

2. Krotkov E., “Focusing*”, Intl. Journal of Computer Vision, vol. 1, pp. 223-327, 1987.*

3. Pentland A., “A new sense for depth of field”, *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), vol. 9, no. 4, pp. 523-531, July 1987*

4. Pentland A., Scherock S., Darrell T., and Girod B., “Simple range cameras based on focal error”, *Journal of Opt. Soc. of America, vol. 11, no. 11, pp. 2925-2935, November 1994.*

5. Nayar S. K. and Watanabe M. Noguchi M., “Real-time focus range sensor”, *Proc. of Intl. Conf. on Computer Vision, pp. 995-1001, June 1995.*

6. Subbarao M. and Surya G., “Depth from defocus: a spatial domain approach”, *Intl. Journal of Computer Vision, vol. 13, no. 3, pp. 271-294, 1994.*

7. Watanabe M., and Nayar S. K., “Rational filters for passive depth from defocus”, *Technical Report CUCS-035-95, Dept. of Computer Science, Columbia University, New York, NY, USA, September 1995.*

8. Watanabe M. and Nayar S. K., “Telecentric optics for constant magnification imaging”, *Technical Report CUCS-026-95, Dept. of Computer Science, Columbia University, New York, NY, USA, September 1995.*

9. Xiong Y. and Shafer S. A., “Depth from focusing and defocusing”, *Proc. of IEEE Conf. on Computer Vision and pattern recognition, pp. 68-73, June 1993.*

10. Xiong Y. and Shafer S. A., ”Moment filters for high precision computation of focus and stereo”, *Proc. of Intl. Conf. on robotics and automation, pp. 108-113, August 1995.*