Fast Stereo with Background Removal using Phase Correlation

S. Effendi[1], R. Jarvis[1], D. Suter[2]

[1]IRRC, Monash University, Clayton, Vic 3168, Australia.
[2] School of Computer Science, University of Adelaide, North Terrace SA 5005, Australia.

Email: {sutono.effendi, ray.jarvis}@eng.monash.edu.au, david.suter@adelaide.edu.au

Abstract

Segmenting a scene into meaningful regions is a classical problem. One of the methods is segmentation from Stereo (on condition that the object has sufficient texture). Most of the time, our interest is foreground objects. However, in dense stereo matching the entire image is processed, thereby wasting time. Hence, in order to process only the foreground objects, we insert a preliminary process to discard the background using dominant motion estimation. Our method estimates the background from stereo images using the phase correlation method. The dominant background area, which has high phase correlation, is then eliminated leaving the foreground to be processed by the traditional Sum of Squared Differences (SSD) algorithm. Thereby, the time taken for the overall process is reduced.

Keywords: Stereo, Segmentation, Dominant Motion, Phase Correlation

1 Introduction

Scene Segmentation or the extraction of meaningful regions, has many important applications, such as object tracking, obstacle avoidance in robotics and driving assistance in automotive industry [1]. Many of these applications commonly employ the stereo approach discussed in [2] and [3]. In particular, obstacle avoidance applications demand real time processing. Hence, a fast stereo method is of enormous importance. Although most of the time the regions of interest are associated with the foreground, many stereo algorithms process the entire image.

A method, in which stereo matching is performed progressively was proposed in [4]. Instead of executing dense matching over the entire image, the algorithm starts with a few reliable points as seeds that satisfy a certain confidence threshold. The matching process then spreads to neighbouring pixels the disparity gradient of which does not exceed a certain threshold value. This method is called the disparity gradient limit that affirms that an object should have a disparity map which varies smoothly. Since it starts with reliable seeds, this scheme potentially reduces many false matches.

Our proposed method is similar to dominant motion estimation between two frames [5]. This dominant motion estimation calculates the global motion followed by the alignment. Subsequently, the residuals are estimated as motion. In [6], the phase correlation method is employed to give a good initial estimation of the global motion in order to speed up the iteration process of the dominant motion estimation.

Similarly, our method uses phase correlation to measure the global motion between left and right images. In this way, those pixels that are successfully aligned can be removed and labelled as background. This is possible because the background does not have zero disparity.

The purpose of background removal is to process the estimated foreground objects only rather than the entire image in order to obtain a disparity map. The foreground objects are obtained by cancelling the background, using phase correlation of both images to identify the translation between each image. Here, assumptions are made that both images are aligned vertically, such that only horizontal translation is present, and that both images are rectified.

2 Phase Correlation

The phase correlation method was introduced in [7] and it has long been known to be superior to the cross correlation method. This method yields a higher peak than the traditional cross correlation method at the match location, and it is also able to achieve sub-pixel accuracy where the two images align [8], [9]. Due to its superior performance, this method was used for 3D face recognition [10], and in image registration [11].

The reason why the phase correlation method is superior to the cross correlation method is simply an image phase component from the Fourier Transform (FT) carries more information than its magnitude component, as shown in [12]. In addition, the correlation in the frequency domain using the Fast Fourier Transform (FFT) yields a faster computation than in the spatial domain.

Let the left and right image be $f_1(x,y)$ and $f_2(x,y)$ respectively, and hence their FT are given as follows

$$F_1(u,v)=\sum_m\sum_n f_1(x,y)e^{-j(ux+vy)} \tag{1}$$

$$F_2(u,v)=\sum_m\sum_n f_2(x,y)e^{-j(ux+vy)} \tag{2}$$

where m and n are the height and the width of the image. To simplify equations (1) and (2), they can be written as

$$F_1(u,v)=A_{F_1}(u,v)e^{j\theta_{F_1}(u,v)} \tag{3}$$

$$F_2(u,v)=A_{F_2}(u,v)e^{j\theta_{F_2}(u,v)} \tag{4}$$

where A_{F_1} and A_{F_2} denote the amplitude components and $e^{j\theta_{F_1}(u,v)}$ and $e^{j\theta_{F_2}(u,v)}$ denote the phase components of the left and the right image respectively

Thus, the cross phase spectrum of the left and right images, $f_1(x,y)$ and $f_2(x,y)$ respectively, is given as

$$\hat{R}(u,v)=\frac{F_1(u,v)F_2^*(u,v)}{|F_1(u,v)F_2(u,v)|}=e^{j\theta(u,v)} \tag{5}$$

where F_2^* is a complex conjugate of F_2, $e^{j\theta(u,v)}=e^{j\theta_{F_1}(u,v)+j\theta_{F_2}(u,v)}$. The denominator just serves as a normalisation factor.

Therefore, the phase correlation of two images is obtained from the inverse Fourier Transform of their cross phase spectrum $\hat{R}$ given as

$$\hat{r}(x,y)=\frac{1}{MN}\sum_m\sum_n \hat{R}(u,v)e^{-j\theta_{F_1}(u,v)}e^{-j\theta_{F_2}(u,v)} \tag{6}$$

where M and N are the height and the width of the image. The dominant peak occurs at the location where both images are in alignment.

Since the phase correlation is a Fourier transform of an image as whole rather than a point, its correlation peak suggests a translation distance for most of the area in the image. Thereby, computing the intensity difference for both images at this distance will yield minimum square error.

Both images are passed through a Laplacian of Gaussian (LoG) filter in order to reduce the intensity difference acquired from different stereo cameras. Hence, instead of performing phase correlation on the image absolute brightness, the phase correlation is performed on the difference of the brightness.

$$LoG(x,y)=\frac{x^2+y^2}{\sigma^4}g_{2D}(x,y)-\frac{2}{\sigma^2}g_{2D}(x,y) \tag{7}$$

where g_{2D} is a Gaussian filter in 2D given as

$$g_{2D}=\left(\frac{1}{\sqrt{2\pi\sigma}}e^{\frac{-x^2}{2\sigma^2}}\right)\left(\frac{1}{\sqrt{2\pi\sigma}}e^{\frac{-y^2}{2\sigma^2}}\right) \tag{8}$$

The window size of the filter used in our experiment is the same size as our SSD window size which is 9x9 and σ is chosen to be 1 in our experiments.

3 Background Estimation

Having successfully registered the right image to the left (as a reference), we then seek to eliminate background pixels. When two images are perfectly aligned, the result of subtraction from each other will be zero.

The left image is aligned and subtracted from the right image. The remaining residue indicates improper alignment which consists of occlusion areas, edges, and foreground objects, as shown later from our experiments in figures 4 and 7. Hence, the background pixel is estimated as follows

$$\hat{b}=\left|\hat{f}_1-f_1\right|^l<\delta \tag{9}$$

where $\hat{f}_1$ is the right image shifted to the left determined by the location of the dominant peak from equation (6), l is 1 or 2 for L1 and L2 distance metric, and δ is the threshold value. Therefore, a pixel is classified as background if the difference is less than certain threshold $\delta=1$. $l=1$ is used in this experiment.

The higher the threshold is, the larger the background area will be. Hence, the foreground area will be smaller which means that the processing time needed for stereo matching is reduced. At the same time, the accuracy of the estimation is also desired to be as accurate as possible. The accuracy is calculated by comparing the estimated background from the Phase Correlation method with its ground truth. So, the percentage of bad pixel estimation can be computed as follows

$$p=\frac{1}{N}\sum_N\left(\hat{b}-b\right) \tag{10}$$

where b is the true background, and N is the total number of the pixels.

4 Disparity Map Estimation

Following background removal, in our experiment the SSD algorithm is used to obtain a disparity map. The reasons are two folds. The first is that it is high speed, so that it is suitable for real time applications. Additionally, the digiclops stereo camera employed in our experiments provides a built-in function for the SSD algorithm: generating the disparity map below

$$\min_{d_{\min}}^{d_{\max}} \sum_{(u,v)\in W} \left(f_1(u,v) - f_2(x+u+d, y+v)\right)^2 \quad (11)$$

where W is the window size, and d is the disparity.

Since there is no modification on the SSD algorithm, we can expect its ceiling performance to follow the ranking described in [13-16]. The only difference with our method is that we process only those pixels that are labelled as belonging to the foreground in order to reduce processing time while still being able to fulfil its purposes, i.e. for obstacle avoidance, etc.

5 Experimental Results

Our experiments used several images taken from the Middlebury datasets provided in [13-16]. These images are Tsukuba, Art, Baby, Flower Pot, Midd and Cone image. The window size and the maximum disparity distance for each image are tabulated in Table 1 as these parameters significantly affect the stereo processing time.

Table 1: Window Size & Maximum Disparity.

Image	Window Size	Maximum Disparity
Tsukuba	9x9	19
Art	9x9	85
Baby	9x9	50
Flower Pot	9x9	75
Midd	9x9	75
Cone	9x9	60

Before the phase correlation method from equation (5) is performed, it is necessary to be apply a preliminary process employing LoG filter to both the left and the right images using equation (7).

As an example of computing the peak of the phase correlation using equation (6), figure 1 shows the Tsukuba images' cross phase spectrum, and the dominant peak is found at column 6. This means that these two images are globally shifted 6 pixels away.

The performance of the background estimation using equation (9) in order to calculate the percentage of bad pixels using equation (10) are plotted against several values for threshold δ in figure 2. Besides achieving lower error rate, we also want to remove the background areas as much as possible, shown in figure 3. In these experiments, the threshold value is chosen to be a value of 1.

In figure 2, for almost all images except the baby image, the percentage of error rises as the threshold increases. This means that more overestimation occurs by falsely labelling more foreground pixels as background pixels.

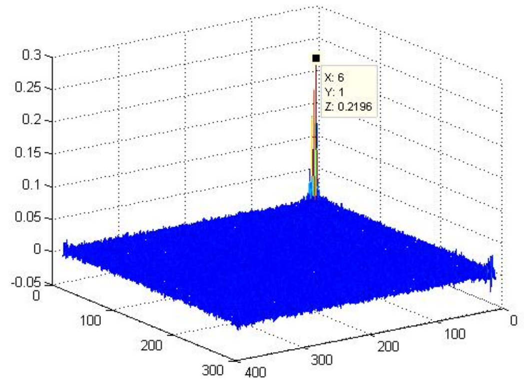


Figure 1: Inverse FT of Tsukuba image pair phase correlation.

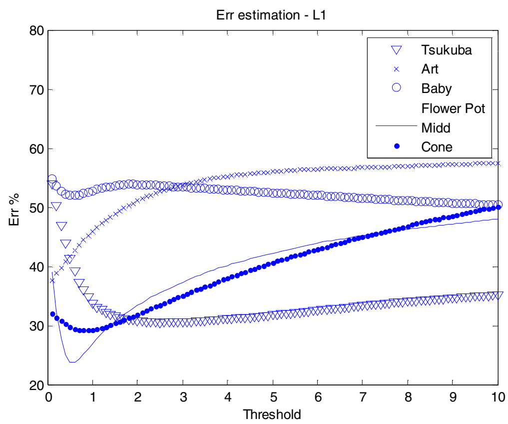


Figure 2: % error at different threshold.

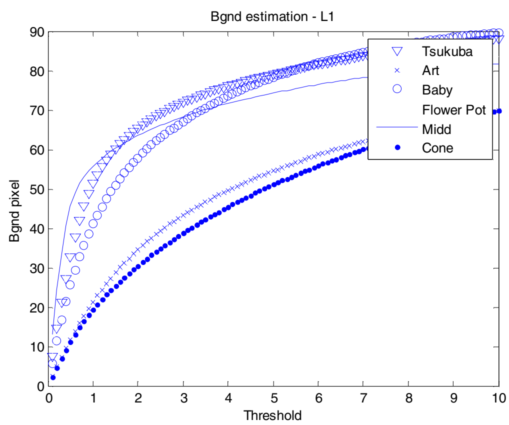


Figure 3: Num. of background pixels.

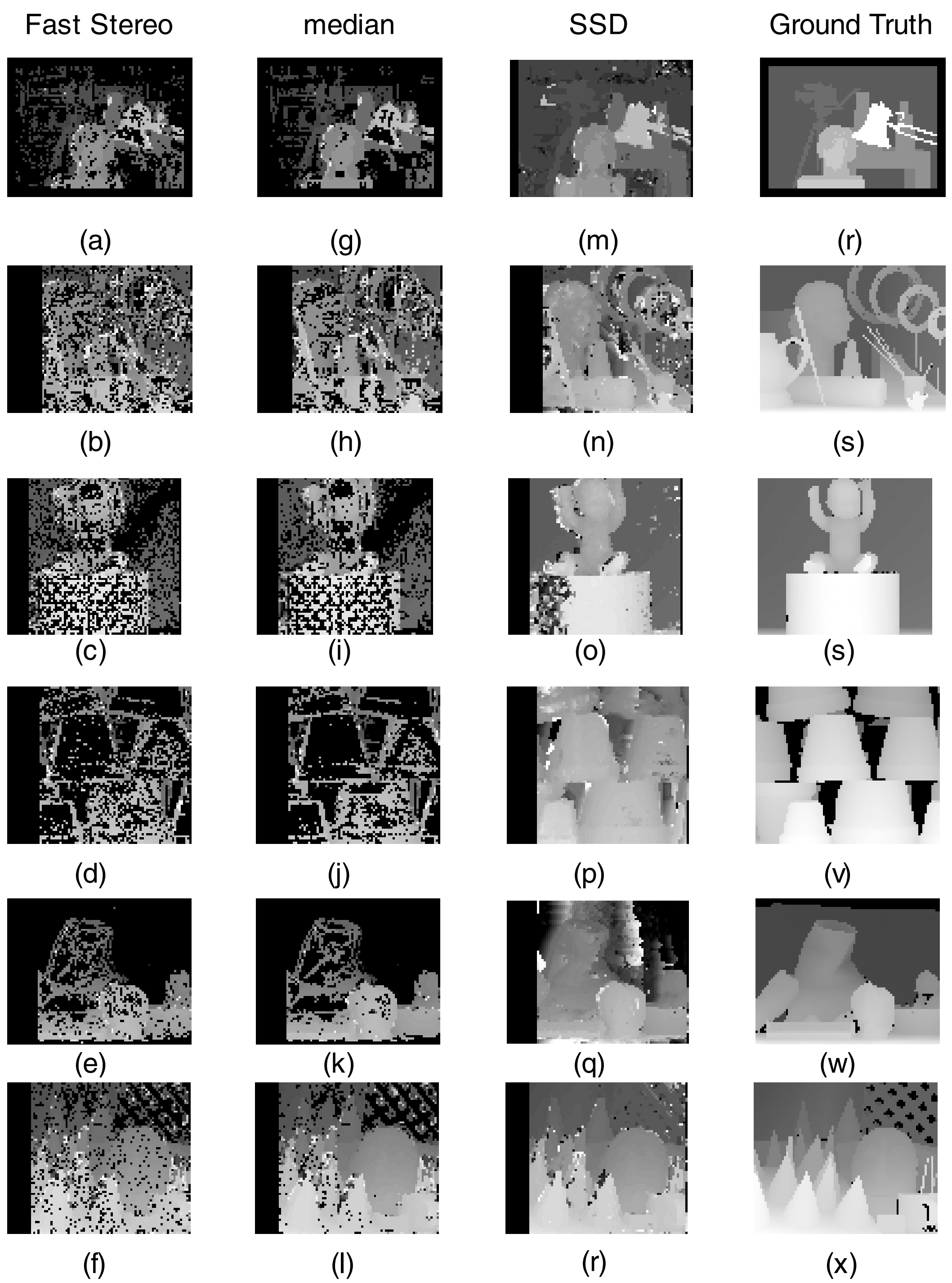


Figure 4: (a) – (f) Fast disparity map, (g) – (l) median filter 3x3, (m) – (r) traditional SSD, (r) – (x) ground truth. Fast stereo with background removal using phase correlation for Tsukuba image, Art image, Baby image, Flower Pot image, Midd image and Cone image [13], [14] , [15] and [16].

In figure 3, the Art and Cone image for background pixels estimation have gradual slope than the other images. In figures 4(f) and 4(l), the estimated background area at the cone image is mostly located at smooth areas on the wood surface. Therefore, the phase correlation method for this image fails to identify the background. Perhaps this is because the objects are located away progressively from the camera hence there is no clear cut between the foreground and the background.

The false detection on smooth objects can easily be understood from equation (9) because its difference is likely to be very small thereby the pixels with small difference are identified as the background. In contrast, the traditional SSD algorithm is still able to distinguish those smooth areas. The background that is falsely identified as foreground does not affect the accuracy of the performance. This, however, slows down the overall process as the algorithm computes the unnecessary matching.

Then, a traditional standard SSD algorithm from equation (11) is performed in order to produce a disparity map. The matching process takes the left image as the reference. Hence at figure 4 and 7, we can see black area on the left side of the disparity map. The width of the black area signifies the maximum disparity distance. A median filter is employed to clean up the disparity image as shown in figure 4. Lastly, a median filter 3x3 is used to clean the entire disparity map.

In this paper, we do not mean to compare the performance evaluation in real time using a commercial hardware. However, the evaluation is done in the Matlab environment using a PC 2.9 GHz with 2 GByte memory. The result of the performance for all images is shown in figure 5.

The image that has the lowest error rate and the fastest performance is the Midd image. However, lots of mislabelling foreground pixels as background occurs on almost the entire object of the white bucket area as shown in figures 4(e) and 4(k). Same false labelling also takes place in the flower pot image. Considering the obstacle avoidance purposes with low error rate and the speed, the best result is achieved on the Tsukuba image shown in figure 4(a) and 4(g). Most of the foreground objects are reasonably well detected such as the head and the camera with some area missing on the table and the lamp.

As mentioned earlier, our method overall accuracy will not exceed ordinary SSD algorithm in [15]. Hence, it is sufficient to compute the overall of this method performance as background error. We also have to take note that this error occurs on smooth areas.

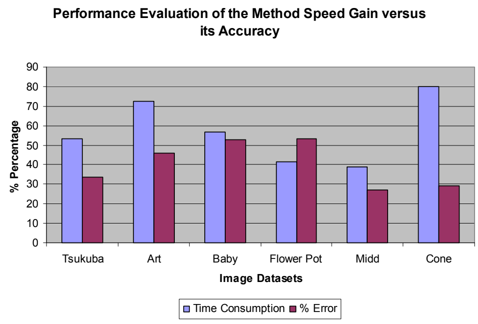
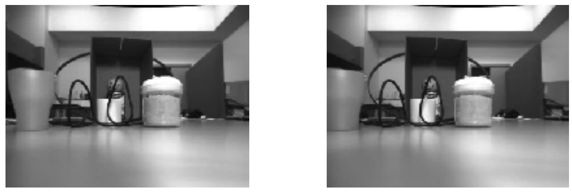


Figure 5: Performance Evaluation of the Method Speed Gain versus its Accuracy.



(a) (b)

Figure 6: (a) (b) The left and right images of our laboratory images.

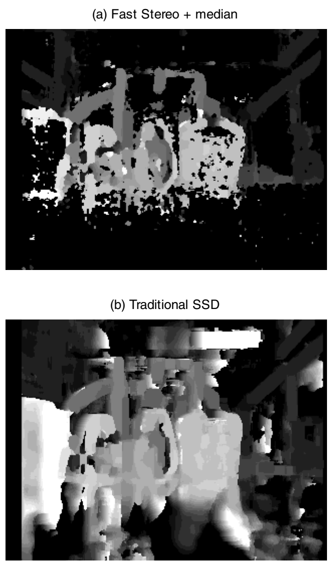


Figure 7: (a) Fast Stereo + median filter on our laboratory image (b) the traditional SSD.

We also created our own laboratory image for evaluation as shown in figures 6 and 7. The image is taken using a digiclops camera, with base line of 2 cm away. There are three cameras (Top, Right, and Left) in a Triclops, however only the right image is used. Next, the camera is shifted 2 cm away, the same right image is used again. The reason for so doing is that in the vision guided robotic grasping the distance from the robot camera to the objects is ranging 1 ~ 2 meter away. However, the smallest distance between the stereo cameras available is 10 cm.

Using equation (6), the displacement is 12 pixels away. Using window size of 9x9 and a maximum disparity of 24, the elapsed time of our method is reduced to 44% (9.86 seconds) than the traditional SSD method 22.17 seconds. The glass at the left hand side is failed to be identified as its area is smooth. Overall, the disparity map result is satisfactory as it can identify almost all foreground objects.

6 Conclusion

The larger the disparity range is, the better our method contributes to the reduction of computation effort. Generally, our method was about 50% faster than the traditional SSD method. However, the performance degrades if the foreground objects contain lots of smooth areas. In the absence of smooth areas, the ceiling of the performance is the same as the traditional SSD.

The phase correlation method significantly degrades when there is large different in views between images. So in order to obtain a good global displacement estimation of two images using phase correlation, a short base line stereo is desired. Furthermore, the further the background from the foreground the better the phase correlation method performs.

Our method can also be used for foreground background separation by further processing. All pixels in the background, that have disparity values less than a certain value, are eliminated.

Overall our method can reduce stereo processing time. Our future work will focus on how to reduce the background estimation error on smooth areas.

7 References

[1] Y. Huang, S. Fu, and C. Thompson, “Stereovision-based object segmentation for automotive applications,” Eurasip Journal on Applied Signal Processing, vol. 2005, no. 14, pp. 2322-2329, 2005.

[2] E. Francois, and B. Chupeau, “Depth-based segmentation,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 7, no. 1, pp. 237-240, 1997.

[3] H. Kim, D. B. Min, and K. Sohn, "Real-time stereo using foreground segmentation and hierarchical disparity estimation," Lecture Notes in Computer Science, 6th Pacific Rim Conference on Multimedia - Advances in Mulitmedia Information Processing - PCM 2005, 2005, pp. 384-395.

[4] Z. Zhang, and Y. Shan, “A progressive scheme for stereo matching,” Springer LNCS 2018: 3D Structure from Images - SMILE 2000, pp. 68-85, 2001.

[5] Y. Huang, D. Paulus, and H. Niemann, “Background-Foreground Segmentation Based on Dominant Motion Estimation and Static Segmentation,” Journal of Computing and Information Technology - CIT, vol. 8, no. 4, 2000.

[6] F. F. Edouard, Thoreau, Dominique (FR), Chevance, Christophe (FR), Viellard, Thierry (FR), Process for estimating a dominant motion between two frames,to Thomson, Multimedia SA. (FR), 2000.

[7] C. D. Kuglin, and D. C. Hinne, “The phase correlation image alignment method,” Proc International Conference on Cybernetics and Society, pp. 163-165, 1975.

[8] K. Takita, T. Aoki, Y. Sasaki et al., “High-Accuracy Subpixel Image Registration Based on Phase-Only Correlation,” IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E86-A, no. 8, pp. 1925-1934, 2003.

[9] K. Takita, M. A. Muquit, T. Aoki et al., “A sub-pixel correspondence search technique for computer vision applications,” IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E87-A, no. 8, pp. 1913-1923, 2004.

[10] N. Uchida, T. Shibahara, T. Aoki et al., "3D face recognition using passive stereo vision," IEEE International Conference on Image Processing 2005, ICIP 2005. pp. 950-953.

[11] B. Lisa Gottesfeld, “A survey of image registration techniques,” vol. 24, no. 4, pp. 325-376, 1992.

[12] A. V. Oppenheim, and J. S. Lim, “The importance of phase in signals,” Proceedings of the IEEE, vol. 69, no. 5, pp. 529-541, 1981.

[13] H. Hirschmuller, and D. Scharstein, "Evaluation of cost functions for stereo matching," 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'07.

[14] D. Scharstein, and C. Pal, "Learning conditional random fields for stereo," 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'07.

[15] 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.

[16] D. Scharstein, and R. Szeliski, "High-accuracy stereo depth maps using structured light," 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. I-195-I-202 vol.1.