Enhanced Real-time Stereo Using Bilateral Filtering*

Adnan Ansar Andres Castano Larry Matthies
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109
{ansar, andres, lhm}@jpl.nasa.gov

Abstract

In recent years, there have been significant strides in increasing quality of range from stereo using global techniques such as energy minimization. These methods cannot yet achieve real-time performance. However, the need to improve range quality for real-time applications persists. All real-time stereo implementations rely on a simple correlation step which employs some local similarity metric between the left and right image. Typically, the correlation takes place on an image pair modified in some way to compensate for photometric variations between the left and right cameras. Improvements and modifications to such algorithms tend to fall into one of two broad categories: those which address the correlation step itself (e.g., shiftable windows, adaptive windows) and those which address the preprocessing of input imagery (e.g. band-pass filtering, Rank, Census). Our efforts lie in the latter area. We present in this paper a modification of the standard band-pass filtering technique used by many SSD- and SAD-based correlation algorithms. By using the bilateral filter of Tomasi and Manduchi [1], we minimize blurring at the filtering stage. We show that in conjunction with SAD correlation, our new method improves stereo quality at range discontinuities while maintaining real-time performance.

1. Introduction

Range from stereo is an area of ongoing interest and activity in computer vision. It spans applications from autonomous navigation and robotics to medical imaging and visualization for virtual and augmented environments. It also underlies many research areas such as large-baseline, omnidirectional and multi-view stereo. For each case we must select from a large ensemble of stereo algorithms the one which best balances the accuracy and fidelity of the range estimate against its computational cost.

In many areas such as medical imaging or generation of digital elevation models (DEMs), there is a need for the highest quality range possible; runtime is secondary or inconsequential. Stereo algorithms with these goals tend to favor optimization schemes that propagate global information to refine range estimates which cannot be estimated robustly from local information. Examples include algorithms based on graph cuts [2] and dynamic programming [3].

On the other hand, applications such as perception for autonomous navigation (e.g., robotics, automotive industry) and virtual reality require fast updates of the range estimate. These applications require algorithms with low run-times. They cannot afford the expense of a global optimization and must fall back instead on the best possible analysis of local information.

The key advantage of local approaches is speed and suitability for hardware implementation. Global optimization algorithms commonly require 2 to 3 orders of magnitude more time than even the software versions of the local methods [4]. Our own SAD implementation runs at 16 fps on a Pentium IV 2 GHz processor for images of size 320×240 pixels. In general, if such an algorithm runs in the order of tenths of seconds in software implementations, it can comfortably reach video rates using DSP and FPGA implementations [5, 6]. At the moment, there is no technique for achieving simultaneously the high quality range obtained from global optimization with the fast run-times of local schemes.

This local analysis typically takes the form of correlation based matching of blocks between the left and right image. The two flavors of correlator generally employed are SSD/SAD[7] and normalized cross correlation[8]. In the former, the sum of squared or absolute differences[1] of image intensities between local windows is computed, and the lowest such score corresponds to a match. In the latter, the cross correlation between the windows is highest at a match. In general, SAD is easier to compute and is less sensitive to outliers than both SSD and cross correlation.[9]

*This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

[1] In the case of the census algorithm, it is the Hamming distance rather than difference in intensities that is summed.

We validate this in Section 4 by showing that normalized cross correlation produces weaker matches than SAD.

All such local techniques must account in some way for photometric variation between cameras in the stereo rig. One method commonly employed with cross correlation is image normalization, in which each image is modified to have local statistics with zero mean and standard deviation equal to one. For SSD/SAD, some form of band-pass filtering is typically used. This may take the form of a Laplacian of Gaussian convolution, a difference of Gaussians or a difference of averaging filters. These amount to a spatial filtering in which texture information is preserved while low frequency background intensity and very high frequency noise are suppressed. In practice, only the high pass component which accounts for the photometric balance is needed. A fundamentally different approach is found in the Rank and Census algorithms [10]. Here the original image is replaced by one which directly encodes local image statistics. In the Rank case, each pixel is replaced by the number of neighboring pixels of lower intensity. In the Census case, each pixel is replaced by a bit string encoding the intensity of all neighboring pixels relative to the central pixel.

Real-time stereo can be improved by modifying the correlator, by modifying the pre-processing step to supply the correlator with better information, or by some combination of the two. Various adaptations of the basic correlation scheme have been proposed. These include shiftable [11], overlapping [12] and adaptive [13] windows. However, any of these techniques will benefit from a better pre-processing of the image. We show this explicitly in Section 4.

We develop in this paper a technique for improved pre-filtering of imagery for SAD-based stereo. The technique consists of replacing the normal band-pass stage, which introduces an inherent image smoothing, with an adaptive process based on the bilateral filter (similar in concept to the SUSAN smoother [14]), introduced by Tomasi and Manduchi [1]. We show that the results are superior to band-pass filtering with SAD as well as to normalized cross correlation. We do not compare directly to Rank or Census, but these are known to suffer from the same problems at discontinuities as SSD/SAD and normalized cross correlation. [6] Furthermore, in our experience rank suffers from low information content relative to the other algorithms and performs poorly on fine structures. A fair comparison with Census would require computation on imagery with bit-depth equal to the size of a filter window. While this is ideal for hardware implementations, it is less suitable for our software based tests.

In Sec. 2 we discuss how the smoothing effect of the standard SAD pre-filter is partially responsible for the low quality at range discontinuities. In Sec. 3 we provide the necessary background on the bilateral filter and describe our adaptations for its use in real-time stereo. We also show explicitly the effect of different pre-filtering schemes on a synthetic image pair. In Sec. 4 we provide experimental results with real data. Finally, we draw our conclusions in Sec. 5.

2. Standard SSD/SAD Pre-Processing

Any stereo algorithm must compensate for photometric variations between the cameras of the stereo rig. The usual approach for SSD/SAD algorithms is to apply a Laplacian of Gaussian filter, which suppresses high frequency noise (intrinsic Gaussian smoothing) while simultaneously normalizing the intensity information and preserving texture information. This can be well approximated by a Difference of Gaussians (DOG) [15] in which the original intensity image I is replaced by I', the difference of its convolution with a large and small Gaussian kernel, i.e.

$$I' = I * G(\sigma_{small}) - I * G(\sigma_{large})$$

In effect, the small Gaussian serves as a low pass filter and the differencing serves as a high pass filter. For imagery of good quality, the noise suppression provided by the low pass filter is generally unnecessary. Thus, we only require a high-pass filter, which can be achieved by background subtraction, i.e.,

$$I' = I - I * G(\sigma_{large}) \tag{1}$$

We have found the difference in stereo quality between background subtraction, an averaging bandpass filter, and convolution with a Laplacian of Gaussian to be negligible. In the remainder of this paper, we will use background subtraction as the basis for comparison with our new approach.

Regardless of the variant used, any of the above methods introduces a blurring across image discontinuities. The effect is a ringing around foreground objects which results in a weakening of correlation match and a bleeding of range across the discontinuity. We demonstrate this in Fig. 1 using a 15x15 kernel for background subtraction.[2] Note the ringing or halo effect near the trees. This does not correspond to any real image content and is simply a side effect of the background subtraction. However, it does result in mis-estimation of disparity near the trees.

A pre-processing step that does not blur across range discontinuities is an obvious step towards improved stereo. However, until recently there has been no low-cost mechanism for smoothing in homogeneous image regions while sharply preserving discontinuities. Complex schemes to extract this information would conflict with the real-time requirement. In the next section, we show that the bilateral filter solves this problem without incurring high computational costs.

[2] This is a kernel size we frequently use for real applications and is not intended solely to highlight the ringing phenomenon.

a)  b)  c) 

Figure 1: a) Image of outdoor scene. b) Averaging with 15x15 filter. c) Background subtracted image. Note the ringing artifact around the trees.

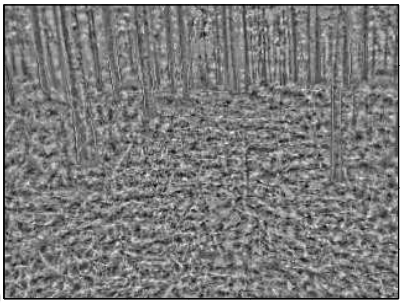
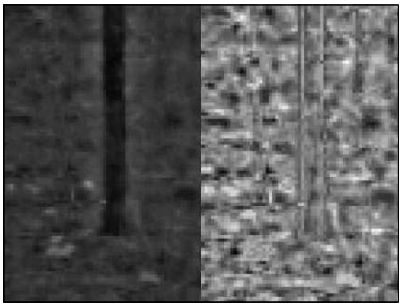
a)  b)  c) 

Figure 2: a) Result of applying 15x15 bilateral filter with $\sigma_d = 5$ and $\sigma_d = 10$ to Fig. 1.a. Note that homogeneous areas such as the ground are blurred, but fine detail and edges are preserved. b) Bilateral background subtraction. Texture is evident, but ringing is less prominent. c) Detail of Fig. 1.c and Fig. 2.b. illustrates removal of ringing artifact

Note that correlation itself introduces an additional error at range discontinuities. Since the correlation window has non-zero extent, it will span objects at two different depths if they are adjacent in the image. The result is an averaging of correlation scores across boundaries. Larger correlation windows result in greater density in the range image because they provide a larger support for the correlation function. However, this increased density is at the expense of accuracy, especially at range discontinuities. Since our goal is to analyze improvements in the pre-filtering stage, We wish to minimize this effect as much as possible and, therefore, restrict ourselves to 7x7 correlation windows.

We now introduce the bilateral filter and show how it can be applied to stereo.

3 The bilateral filter and its application to stereo

The bilateral filter[1] computes the weighted average of the pixels within a neighborhood with the weights depending on both the spatial and intensity difference between the central pixel and its neighbors. Expressed formally, the filter takes a signal $f(x)$ and returns

$$h(x) = \frac{\int_\Omega f(\xi)\, c(\xi, x)\, s(f(\xi), f(x))\, d\xi}{\int_\Omega c(\xi, x)\, s(f(\xi), f(x))\, d\xi} \qquad (2)$$

where Ω is the filter support. The weight functions c and s are typically Gaussian distributions of the form

$$c(\xi,\, x) = e^{-\frac{1}{2}\left(\frac{|\xi - x|}{\sigma_d}\right)^2} \qquad (3)$$

$$s(f(\xi), f(x)) = e^{-\frac{1}{2}\left(\frac{|f(\xi) - f(x)|}{\sigma_r}\right)^2} \qquad (4)$$

For the case of images, $f(x)$ is the intensity at pixel x, σ_d is the standard deviation of the spatial component of the blurring function and σ_r is the standard deviation of the intensity component.

The bilateral filter can be used as an edge-preserving smoother, removing high-frequency components of an image without blurring its edges. We can control the the spatial support of the filter, and thus the level of blurring, by varying σ_d. By varying σ_r, we can adapt the sensitivity of the filter to changes in image intensity. In Fig. 2, we show the same greyscale image of an outdoor scene as in Fig. 1, but now using a 15x15 bilateral filter with $\sigma_r = 5, \sigma_d = 10$. Observe that tree edges are preserved by the bilateral filter while homogeneous regions are blurred. In the background subtracted image, texture is apparent without the noticeable ringing of the standard background subtraction.

For stereo, the bilateral filter takes the place of Gaussian averaging in the background subtraction step. Thus, if the original intensity image is I and its bilaterally filtered ver-

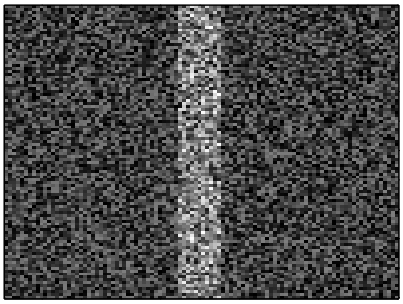


Figure 3: Left image of stereogram of pillar in front of background plane.

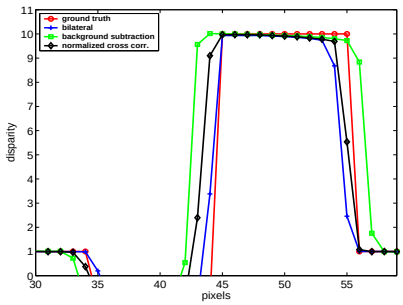


Figure 4: Stereo results averaged over all rows of Fig. 3. Displayed are ground truth(red), background subtraction(green), bilateral filtering(blue) and normalized cross correlation(black). Observe that bilateral filtering and cross correlation are closer to ground truth near the column edges.

sion is B, we replace the image with I' subject to

$$I' = I - B$$

The resulting process achieves the same normalization effect as background subtraction in homogeneous areas, but minimizes the blurring artifact at discontinuities.

We study the effect on stereo with a synthetic example which allows us to control ground truth and which is explicitly designed to illustrate the effect of the new prefilter on edges. In Fig. 3 we show the left image

of a stereogram consisting of uniform random noise. The image is of an 11 pixel wide column in front of a background plane. The background has a disparity of 1 pixel from left to right image and the column has a disparity of 10 pixels. The column is on average brighter than the background. We compute stereo using background subtraction and bilateral filtering, both with 15x15 kernels. In the case of the bilateral filter, we use $\sigma_d = 5$, and σ_r is computed by a heuristic described below. For comparison, we also compute stereo using normalized cross correlation. In all cases, a 7x7 window is used for correlation, and left-right line of sight checking is enabled. In Fig. 4, we show the result of averaging computed disparities over all rows (recall that they should be equal) using the three algorithms just mentioned. We also show ground truth. Observe that both bilateral pre-filtering with SAD and normalized cross correlation are less susceptible to edge effects than the standard background subtraction. We will show in Section 4 that the bilateral approach also preserves the range density typical of SAD and performs better than cross correlation on homogeneous regions.

Figure 5: "Separable" bilateral filter consisting of two passes of 1d bilateral filter (vertical and horizontal) applied to image in Fig. 1.a. Observe that the result is very similar to the 2d filter.

Figure 6: Left image of cones with ground truth disparity.

We now address the crucial issue of runtime. The bilateral filter is not a filter in the traditional sense because the kernel actually depends on the function f in Eqn. 2. In particular, this complicates computation because the bilateral filtering process is not separable. However, we have found that approximating with a separable filter is adequate. In Fig. 5, we show the result of applying a separable approximations consisting of a pair of 1d bilateral filters, one horizontal and one vertical, to the image in 2. Observe that the results are quite similar to the true 2d filter. With this "separable" version of the filter, our real-time system runs at 10 fps on 320x240 imagery using a 2 GHz P4 processor. We anticipate that further optimization is possible.

Selection of σ_d, the standard deviation of the spatial distribution, is dictated in part by the correlation window size and is largely independent of image content. However, σ_r necessarily depends on the image. We offer a simple heuristic. For each pixel, we compute local image variance. We then take the mode of this variance over the whole image as

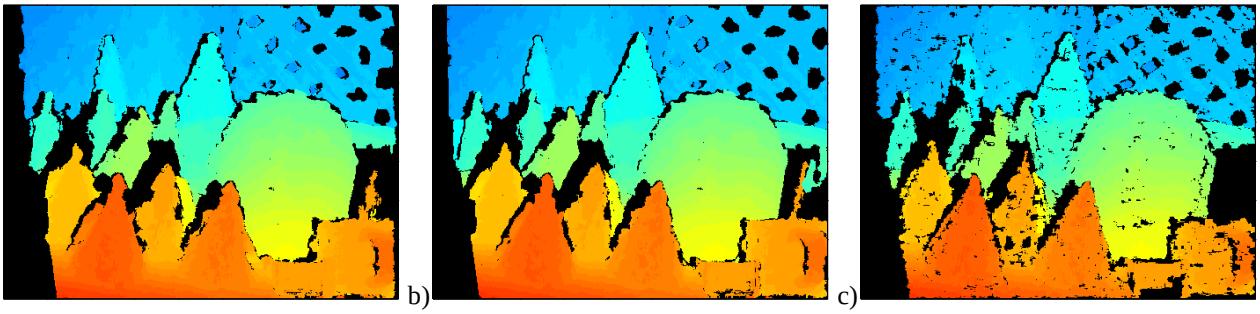


Figure 7: Stereo results on Fig. 6 using a) 11x11 background subtraction. b) 11x11 bilateral filter c) normalized cross correlation. Edges are more faithfully reproduced using both bilateral filtering and normalized cross correlation. However, there is greater loss of valid range using cross correlation.

a reasonable candidate for σ_r^2.

4. Experimental results

We show for real stereo imagery that the use of bilateral filtering with SAD has advantages over both background subtraction with SAD and normalized cross correlation[3]. We will also examine the effect of varying the filter size and show that bilateral filtering is consistently better than background subtraction. In all cases, we use 7x7 correlation windows with left-right line of sight checking enabled and set σ_d in the bilateral filter to $\frac{1}{3}$ the kernel size. We use our separable approximation of the bilateral filter throughout.

In Fig. 6 we see the left image of a stereo pair as well as ground truth disparity. This image is taken from the recent work of Scharstein and Szeliski[16] and used with permission. We begin with a comparison of stereo using background subtraction, bilateral filtering and normalized cross correlation using 11x11 kernels for the pre-filters. For the bilateral filter, we use $\sigma_r = 50$. Note that the filter size is irrelevant for the cross correlation approach. The results in Fig. 7 show sharper definition near edges for both the bilateral and cross correlation approaches. However, the latter is missing more valid range in homogeneous areas. Separate diagnostic tools indicate that most of this loss is due to failure of the left-right check, most likely arising from shallow extrema in the correlation scores.

In Fig. 8 we show that this improvement occurs primarily We make these observations more concrete in Table 1. We accept as accurate those estimates which are within 0.5 pixel of the subpixel ground truth shown in Fig. 6.

at the boundaries of objects and accounts for gross errors rather than subpixel errors. The figure shows a false color image of the absolute difference of disparities between standard SAD and SAD with bilateral filtering. Notice that the

[3]Note that for normalized cross correlation we adapt local image statistics (within a correlation window) to have zero mean and standard deviation equal to one. This corresponds to metric C_4 in [8]. Furthermore, the image is not pre-filtered in any way.

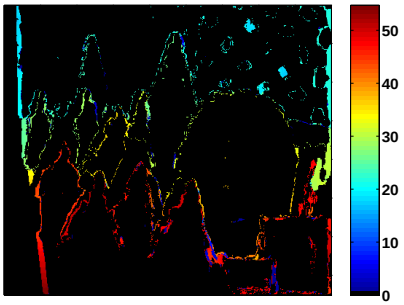


Figure 8: Absolute difference of disparity images from standard and bilaterally filtered SAD stereo. The colorbar on the right (ranging from 0 to 55) indicates the magnitude of the errors.

	% Detected	% Correct	% Incorrect
Bck. Sub.	78.3	68.5	9.8
Bilat.	80.4	72.4	8.0
Cross Corr.	73.7	66.2	7.5

Table 1: Percentage of detected, correctly detected and incorrectly detected range.

difference is typically on the order of 10s of pixels disparity. The majority of pixels corrected by bilateral filtering as reported in Table. 1 are accounted for in this difference image. Notice also that more usable range is recovered at the extreme ends of the image. This error in the standard SAD algorithm results from the background subtraction averaging over the usable edge of the rectified image. We now show that the improvement of the bilateral filter pre-process over background subtraction is independent of kernel size. In Fig. 9 we show both filters for kernel sizes of 7x7, 11x11 and 15x15. In each case, the bilateral filter (second row) produces better stereo at edges and on fine structures.

We turn now to the real scene in Fig. 1 taken from a vehicle during an autonomous navigation trial. Unlike the artificial image of the cones, this image presents a scenario more likely to be encountered by a system for which real-time stereo is crucial. The near trees present a challenge to

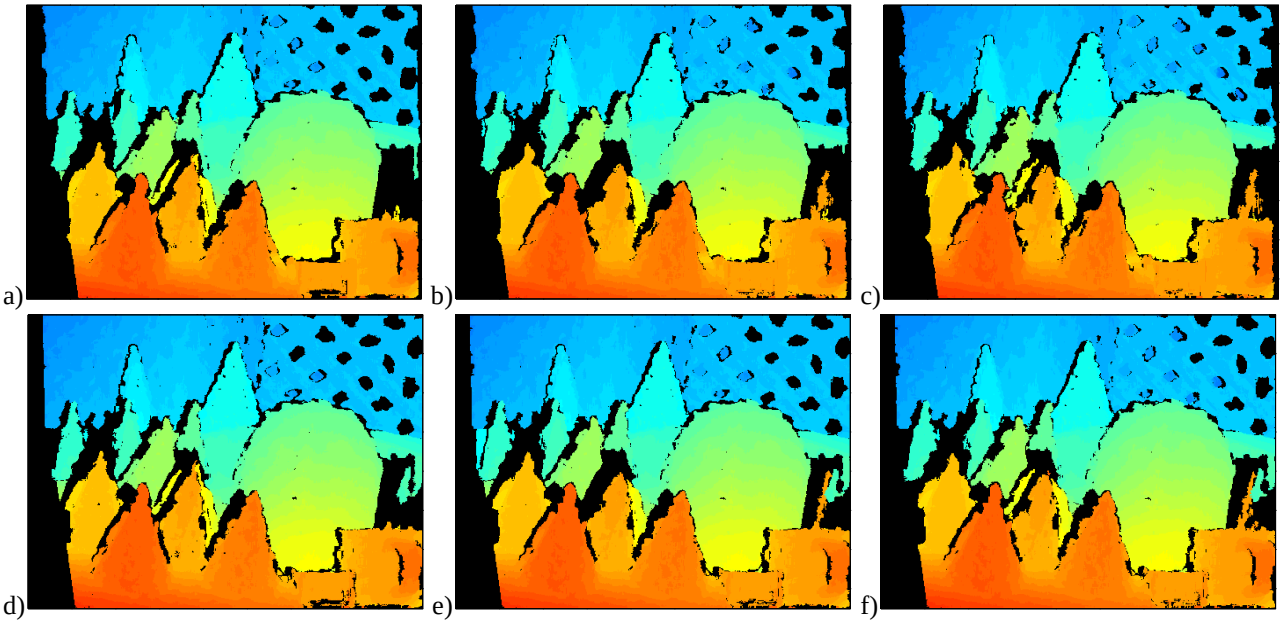


Figure 9: Stereo results on Fig. 6 using background subtraction and a) 7x7, b) 11x11 and c) 15x15 kernels. Same image and stereo parameters using bilateral filtering with e) 7x7, e) 11x11 and e) 15x15 filter. In each case, the bilateral results are better near discontinuities and for fine objects.

conventional SAD stereo. Unlike the cone example, ground truth is not available. However, because of the larger disparity differences involved and the sharper intensity variations between foreground and background, the advantages of the bilateral filter are, nevertheless, apparent. We show in Fig. 10 that bilateral filtering and normalized cross correlation reproduce the trees more accurately, but as with the cone example, cross correlation suffers from greater loss of texture in uniform regions. We use 15x15 kernels for the relevant filters to match our real-time system in this scenario. The bilateral filter and especially the normalized cross correlation lose some disparity data on the ground and in the background trees. In both cases, this is due to a weakening of the correlation match with respect to standard SAD. We illustrate this in Fig. 10 by showing histograms of the absolute subpixel curvature for each algorithm. While raw correlation scores for the different algorithms are not comparable, the curvatures of quadratic fits to the correlation scores are. They represent the sharpness of the fit and can be used directly as a confidence measure on correlation. We see that of the three variants pictured, normalized cross correlation has the weakest correlation peaks, and SAD with background subtraction the strongest. The bilateral filter represents a trade-off. It reproduces the near trees at least as faithfully as normalized cross correlation while minimizing the loss of texture on the ground. Note that the result pictured is typical of the whole sequence from which the current image is taken. A portion of this processed sequence is available at http://robotics.jpl.nasa.gov/∼aiansar/bifilt.

Finally, we prove the claim made in the introduction that bilateral pre-processing can benefit not only simple SAD correlation but modified correlators as well. We illustrate this fact by using shiftable windows with a 3 pixel horizontal shift in conjunction with both background subtraction and bilateral filtering. The results are pictured in Fig. 11. As with the standard SAD correlator, the shiftable window correlator also shows better definition of the near trees using bilateral filtering.

5. Summary and Conclusions

Real-time stereo algorithms typically rely on a simple correlation mechanism applied to imagery processed in some way to account for photometric variations between cameras. Improvements to such algorithms address either the correlation step or, as with our work, the preprocessing step. As we have shown, those modifications which target the latter are likely to also benefit the former. We have presented in this paper an improvement to the filtering step employed by most SAD based correlation algorithms which replaces the conventional bandpass filters with background subtraction of a bilaterally filtered image. The result suppresses photometric variation between cameras, much like other bandpass filters, while maintaining much greater fidelity of data at discontinuities in intensity, hence in most discontinuities in range. This produces better stereo at these range discontinuities. We have also shown that our solution has some advantages over the alternative cross correlation approach in that it has less loss of range in uniform regions. Further-

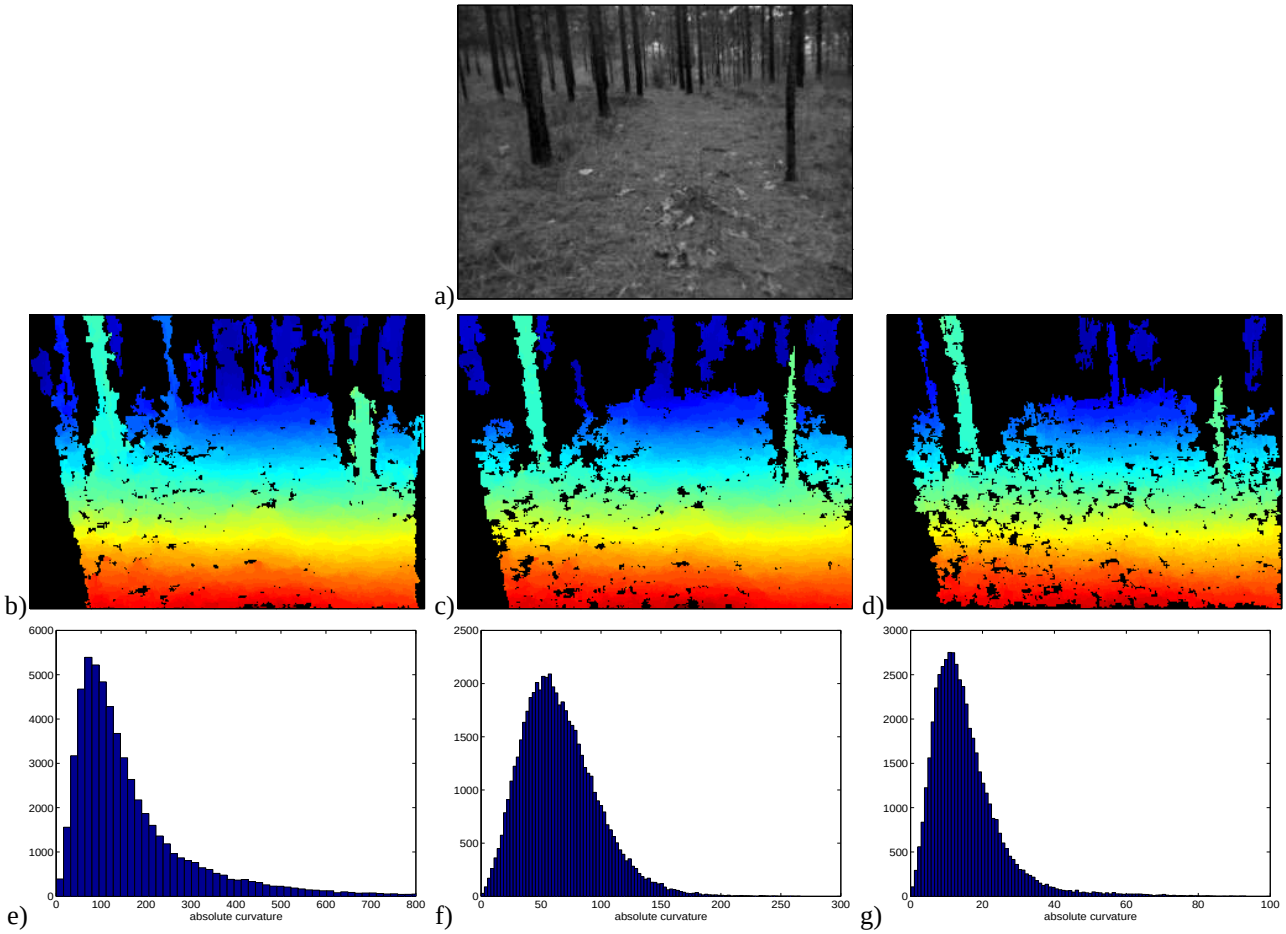


Figure 10: Stereo result on scene in a) using b) background subtraction c) bilateral filtering and d) normalized cross correlation. Tree edges are more faithfully reproduced using both bilateral filtering and normalized cross correlation. However, there is greater loss of valid range, particularly on the ground, using cross correlation. This is explained by examining the curvatures of the subpixel fit. We see histograms of these for e) background subtraction f) bilateral filtering and g) normalized cross correlation. Observe that the mode is highest for e) and lowest for g).

more, our method does not sacrifice the real-time performance which drives current correlation based algorithms. We are currently working to compensate for the loss of data in the background seen in Fig. 10. We believe that a hybrid approach, using bilateral subtraction only in certain regions dictated by image statistics and normal background subtraction elsewhere, will solve this problem.

References

[1] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. IEEE Intl. Conf. Computer Vision*, pp. 836–846, 1998.

[2] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions via graph cuts," in *Proc. IEEE Intl. Conf. Conputer Vision*, (Vancouver, Canada), pp. 508–515, 2001.

[3] Y. Ohta and T. Kanade, "Stereo by intra- and inter-scanline search using dynamic programming," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 7, no. 2, pp. 139–154, 1985.

[4] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1/2/3, pp. 7–42, 2002.

[5] K. Konolige, "Small vision systmes: Hardware and implementation," in *Proc. Intl. Symposium of Robotics Research*, (Hayama, Japan), 1997.

[6] J. Woodfill and B. V. Herzen, "Real-time stereo vision on the PARTS reconfigurable computer," in *Proc.*

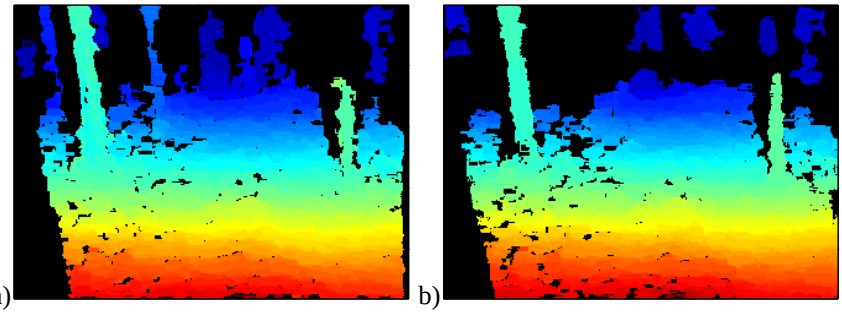


Figure 11: Shiftable window stereo using a) background subtraction and b) bilateral filtering. As with standard SAD, the bilateral pre-filter produces sharper results on the near trees.

IEEE Symp. FPGAs for Custom Computing Machines, (Napa, CA), pp. 242–250, 1997.

[7] L. Mathies, "Stereo vision for planetary rovers: Stochastic modeling to near real-time implementation," *International Journal of Computer Vision*, vol. 8, pp. 71–91, 1992.

[8] O. Faugeras, B. Hotz, H. Mathieu, T. Viéville, Z. Zhang, P. Fua, E. Théron, L. Moll, G. Berry, J. Vuillemin, P. Bertin, and C. Proy, "Real time correlation based stereo: algorithm implementations and applications," Tech. Rep. 2013, INRIA, 1993.

[9] K. Muhlmann, D. Maier, J. Hesser, and R. Manner, "Calculating dense disparity maps from color stereo images, an efficient implementation," *IJCV*, vol. 47, pp. 79–88, April 2002.

[10] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. IEEE European Conf. Computer Vision*, (Stockholm, Sweden), pp. 150–158, 1994.

[11] A. Bobick and S. Intille, "Large occlusion stereo," *International Journal of Computer Vision*, vol. 33, no. 3, pp. 181–200, 1999.

[12] H. Hirschmüller, "Improvements in real-time correlation-based stereo vision," in *IEEE Conf. Computer Vision and Pattern Recognition*, (IEEE Conf. Computer Vision and Pattern Recognition), pp. 141–148, 2001.

[13] M. Okutomi and T. Kanade, "A locally adaptive window for signal matching," *International Journal of Computer Vision*, vol. 7, no. 2, pp. 143–162, 1992.

[14] S. Smith and J. Brady, "SUSAN - a new approach to low level image processing," *Int. Journal of Computer Vision*, vol. 23, pp. 45–78, May 1997.

[15] D. Marr, *Vision*. New York: W.H.Freeman and Company, 1982.

[16] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *CVPR*, pp. 195–202, 2003.