Multibaseline Stereo System for Evaluation of Binocular Stereo

Adnan Ansar* and Larry Matthies
Jet Propulsion Laboratory, Calif. Inst. of Tech., 4800 Oak Grove Drive, Pasadena, CA USA 91109

ABSTRACT

Binocular, correlation based stereo has been a key component in many efforts at autonomous vehicle navigation. However, estimation of ground truth range data, especially in field conditions, remains a challenge. We present a 5 camera, multibaseline stereo system and demonstrate its use as a passive ground truthing mechanism for binocular stereo. In this paper, we provide both a system description and a detailed overview of a novel depth-based multibaseline stereo algorithm. Our new algorithm avoids the need for pairwise camera rectification. We conclude with several simulations and real world experiments to verify our results.
Keywords: Multibaseline, stereo, navigation

1. INTRODUCTION

Binocular stereoscopic vision has been an important component in autonomous navigation. We list only a few of countless references on the subject.[1,2,3,4] Stereo provides a real-time, low cost method to infer 3D structure for use in hazard avoidance, path planning and other decision-making processes. Where such factors are a consideration, it is also a low visibility passive sensor. Key to the real-time aspect of stereo is the use of algorithms based on sum of absolute difference (SAD) correlation or basic variants. An overview and extensive bibliographies can be found in many places.[5,6] Such algorithms are amenable to substantial optimization on existing hardware.[7] In low to mid complexity environments, stereo by SAD correlation has proven itself a robust and reliable tool. However, as more demanding tasks are required of autonomous systems, the need to operate in more complex environments increases. While more exotic approaches[8], which often produce better results, have recently been proposed in the literature, they are not yet at a stage to be used in real-time systems. Thus, any improvements to existing stereo algorithms for use in navigation are likely to be based on correlation approaches. In order to evaluate any such improvement, we need a mechanism to establish ground truth. One possibility is to use active sensors such as ladar to establish a highly accurate 3D model of a scene. Even ignoring issues of quality and accuracy of 3D data, there are a number of complications involved in such an approach. Primary among them are spatial and temporal registration of the stereo data with the active sensor data. Spatial registration involves a potentially complex calibration that must be repeated whenever the active and passive systems are used together. It amounts to registering 3D point clouds[9] acquired by different sensors with different resolutions and geometries. By temporal registration, we refer not to a simple timestamp between datasets, but the actual mode of acquisition. A stereo pair is effectively an instantaneous snapshot of the screen. Ladar, on the other hand, is typically a single axis device which produces a raster scan of the scene, usually with a limited vertical field of view. It becomes necessary to compensate for both scene and vehicle motion in the later case, but not in the former. In complex terrain, particularly in the presence of foliage which can move in a breeze, misregistration is a possibility even if spatial calibration is perfect. Another factor is cost. CCD cameras are relatively cheap and easily obtained, whereas more specialized equipment may not be.

An alternative approach is to accept as "ground truth" the best possible result one can expect from correlation based stereo. We describe in this paper one attempt at achieving this best possible result. Note that real-time performance is not an issue for this algorithm, since it is intended primarily for offline evaluation of real-time algorithms. Our approach involves the use of multiple cameras involving multiple stereo baselines to estimate 3D structure. Since any subset of two cameras from our collection can be considered a binocular stereo pair, the end result is a direct comparison mechanism between ground truth and standard stereo, and since identical sensors and data are used for our comparison, the registration issues are simply non-existent. Before describing in detail our multibaseline evaluation algorithm, we

* Adnan.I.Ansar@jp.nasa.gov; phone 1 818 393-7242; fax 1 818 393-4085

briefly summarize the steps of our binocular stereo implementation. We will revisit these steps and show differences when we describe the multibaseline algorithm.

Figure 1: Left camera image after rectification and associated range map for 3 different scenes, increasing in complexity from top to bottom.

1.1 Outline of binocular stereo

Starting with a pair of images acquired by a binocular stereo system, processing proceeds as follows. (1) Non-linear distortion is removed and the images are rectified to a plane containing the two optical centers so that image scanlines correspond to epipolar lines. This reduces the correlation problem from a 2D search to a 1D search. (2) The images are pre-processed in several ways. These include Gaussian smoothing, reduction (if desired), and bandpass filtering to accommodate differences in photometric properties between the cameras. (3) For each pixel p in the left rectified image at coordinates *(x,y)*, we take a candidate horizontal disparity value for the corresponding pixel in the right image, now assumed to be at *(x+d,y)*. (4) We take the sum of absolute differences of a rectangular region centered at *(x,y)* in the left image and *(x+d,y)* in the right image. This value is considered the correlation score for pixel p at disparity *d*. (5) We minimize the correlation score for each pixel *p* as a function of *d*, resulting in a disparity map co-registered with the left image. (6) We perform several heuristic algorithms including left-right line of sight check and blob coloring/filtering on the disparity image. (7) We perform subpixel refinement using the disparity values and correlation scores of the filtered disparity image. From the intrinsic geometry of the camera, we translate subpixel disparity to a range for each pixel.

With the exception of implementation details and optimizations[6] the above outline or some basic permutations of it applies to any real-time correlation based stereo algorithm. Before proceeding to our evaluation algorithm, we show in Figure 1 some stereo results on three scenes of increasing complexity. Data was acquired using 1024 x 768 Dragonfly firewire board cameras with 60° field of view and with a stereo baseline of 10 cm. In the examples below, and in all examples to follow, we use 11 pixel x 11 pixel correlation windows on 512 x 384 images (reduced by a factor of 2 from the original 1024x768 images). Observe that in the low and mid complexity scenes, stereo does quite well. In the high complexity scene, the results are more difficult to judge.

2. MULTIBASELINE SYSTEM

2.1 System description

Our multibaseline camera system consists of 5 Dragonfly firewire cameras from Point Grey Research. The cameras are mounted on a rigid metallic support in a cross configuration. The support is mounted on a pan-tilt unit attached to a cart for easy mobility. The center-left pair of cameras has a 16 cm baseline, the center-right a 10 cm baseline, the center-top a 10 cm baseline and the center-bottom a 16 cm baseline. Our intuition in developing this configuration was to have stereo pairs oriented in the two principle directions with several (3 each) useable baselines. In addition to the ground truthing mechanism to be described, this allows us great flexibility data collection of binocular pairs. The system is pictured in Figure 2.

Figure 2: Five camera Multibaseline stereo system

2.2 Algorithm Introduction

We now describe in detail our multibaseline stereo algorithm. While the algorithm was produced primarily for use with the 5 camera rig pictured in Figure 2, it should be noted that it will accommodate any number of cameras in any configuration. The primary difference between our approach and many other trinocular and multibaseline approaches is that we iterate directly over depth rather than over disparity. Simlar depth based approaches have been proposed [10], but our implementation is simpler and makes no assumption about piecewise planarity of the scene. Indeed, for the types of scenes in which we are interested, this is far from the case. It is known [11] that minimization over inverse depth with three or more cameras is inherently less ambiguous than minimization over disparity in a comparable setup. While this is beneficial, our motivation in using a depth-based search is different. We see two primary benefits. First, it avoids the need to rectify individual stereo pairs, since there is no consistent epipolar geometry across multiple cameras. If individual stereo pairs are rectified, there is either an issue of using a pre-selected stereo pair and referencing all other cameras to it or of solving a non-trivial accounting problem in relating relevant pixels in any rectified pair with those in any other pair.[12] A second important benefit to our method is that it allows depth samplings to be chosen so that a 3D region of interest can be explored in greater detail. This is not dependent on the inverse integer disparity values of any camera pair. In other words, we can explore a

sparse depth sampling in some region of space and a denser sampling in another, where objects of interest lie. This is not necessarily a benefit when operating a stereo system in an unknown environment, but will be useful in evaluation, when the environment and important depth regions are well known. While not directly relevant to our experimental setup, our algorithm will also accommodate as a natural side effect the perspective warping associated with cameras with non-parallel optical axes. Note that while the rectification problem is no longer relevant if we relax standard stereo algorithms to search over a 2D disparity space, the warping problem remains.

2.3 Algorithm Overview

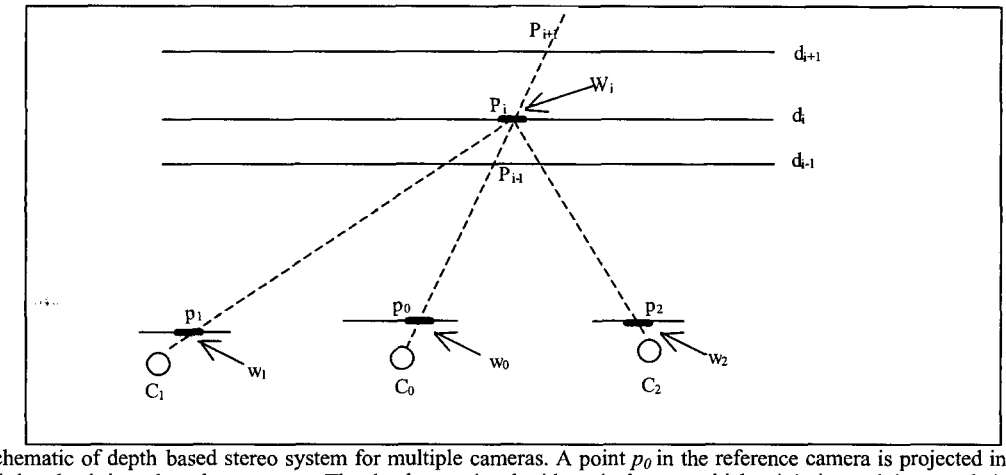

Figure 3: Schematic of depth based stereo system for multiple cameras. A point p_0 in the reference camera is projected into space at depth d_i and then back into the other cameras. The depth associated with p_0 is the one which minimizes a joint correlation function over all the cameras.

In Figure 3, we present a simple overview of the algorithm using 3 cameras viewed from overhead. First a reference camera is chosen. Call it C_0. For each pixel p_0 in C_0, we imagine that it originated from some 3D point Pi at depth d_i. We then project a correlation window w_0 of chosen size and centered at pixel p_0 to a window W_i at depth d_i in space. This window is then reprojected to windows w_1 and w_2 in the remaining cameras, C_1 and C_2 respectively. We can now compute correlation scores for pixel p_0 and depth d_i by combining the sum of absolute differences of windows over pairs of cameras.

$$\sum_{q_l \in C_l, q_k \in C_k} |q_l - q_k|.$$

The most straightforward approach is to take either the product or sum of correlation scores over pairs.

$$S_i = \prod_{l,k} \sum_{q_l \in C_l, q_k \in C_k} |q_l - q_k|, \quad S_i = \sum_{l,k} \sum_{q_l \in C_l, q_k \in C_k} |q_l - q_k|. \tag{1}$$

It is easy to show that the product or sum of this collection of functions at their (nearly) joint minima has higher curvature and is thus better defined than any of the component functions. Observe that the correlation windows need no longer be rectangular. We describe in the next subsection, how this is handled in practice. If S_i is the correlation score as computed above for d_i, we find the index i which minimizes S_i. Then the depth $D(p_0)$ of pixel p_0 in camera C_0

$$D(p_0) = \arg\min_i S_i.$$

Where Si is computed as in Equation (1). For further refinement, we perform the functional equivalent of subpixel interpolation by fitting to a quadratic the sets $\{d_{i-1}, d_i, d_{i+1}\}$ and $\{S_{i-1}, S_i, S_{i+1}\}$. $D(p_0)$ is then set equal to depth value which minimizes the resulting parabola. The above description and schematic convey the idea behind our algorithm, but the implementation differs in several ways. We now describe this in greater detail.

2.4 Algorithm implementation

First, some preprocessing steps are performed on individual images. All images are assumed to result from perspective projection. Thus, we begin as with most stereo implementations by removing any nonlinear distortions using camera models obtained by careful prior calibration. Images and associated camera models are then scaled if desired and a difference of Gaussian or Laplacian of Gaussian filter is applied both to smooth noise and to eliminate variations due to photometric differences between the cameras.

We choose a reference camera C_0 as described above. In our 5 camera evaluation rig, this is normally the center camera. We segment the 3D space into planes parallel to the image plane of C_0. The placement of these planes is arbitrary and depends on the scene involved. If maximum and minimum relevant range values are known, the planes can be distributed more densely within this region. If a one-one map between plane depths and the various integer disparities arising from camera pairs within the system is desired, for example to compare to other algorithms in a meaningful way in disparity space rather than range space, the appropriate planes can be chosen easily.

Normally, the next step in most stereo algorithms is rectification. We show now how we avoid rectification without the need to introduce a 2D search over all images. Let I be an integer indexing of the planes and J be an indexing of the cameras. Then for each $i \in I$, we project all pixels in camera C_0 to the plane P_i. Let T_{ij} be the plane homography from P_i to the image plane of camera C_j. Let

$$\tau_{i,kl} = T_{il}^{-1} T_{ik}, \qquad i \in I; k,l \in J. \tag{2}$$

Equation 2 then provides a homography between camera planes which respects any chosen depth plane. Thus for any region of image C_0 which has a uniform depth approximated by plane P_i, the transformation described by $\tau_{i,kl}$ is in fact a homography for that region. It is essential to note that we perform linear or cubic interpolation to warp the pixilated image data. If this is not done (i.e. if nearest neighbor approximations are used) there is no benefit to choosing trial depths other than those which correspond to fixed pixel disparities among the various cameras. For each camera C_j with $j \neq 0$, Let

$$E_{i,j} = \tau_{i,j0} C_j. \tag{3}$$

It follows that $E_{i,j}$ and C_0 match over any pixels arising from objects on depth plane P_i, or more realistically, that they match closely for objects at approximately the depth of P_i. Thus, for each depth plane, we compute matrices $M_{i,j}$ and $S_{i,j}$

$$M_{i,j} = \left| E_{i,j} - C_0 \right|, \qquad S_i = \bigotimes_j K * M_{i.j}. \tag{4}$$

where in Equation 4, $\otimes$ represents either product or sum as in Equation 1, the * operator is convolution, and K is an averaging filter of size equal to the desired correlation window. For each pixel p at coordinates (m,n) in C_0, we find the depth D of p as

$$D(p) = \arg\min_{P_i} S_i(m,n). \tag{5}$$

Following this, we perform quadratic interpolation at each pixel as described in Section 2.3 to refine our results. The final step of the algorithm is a filtering procedure which eliminates any spurious, isolated patches of range.

In addition to the missing rectification step, observe that there is no left-right line of sight check. This is generally unnecessary for any multibaseline system since the ambiguous correlations matches are automatically reduced because of the additional cameras. Furthermore, in our algorithm, 3D points are being unambiguously projected into image planes, minimizing the problem to an even greater extent.

2.5 A simple variation

As a result of the way our algorithm is structured, there is one variation in addition to using the product or sum to compute a joint correlation score as described above. Instead of a joint score, we have the flexibility to compute a binocular stereo range using different stereo pair on a pixel by pixel basis. This is accomplished by replacing Equation 5 with

$$D(p) = \underset{P_i, C_j}{\operatorname{argmin}} M_{i,j}(m,n). \tag{6}$$

In other words, at each pixel we take the minimum correlation score arising from any of the individual binocular pairs. Note that in this case, we get range data wherever any two cameras have overlap. In the combined approach, we require a point to be visible in all cameras. In addition to greater visibility, we expect this approach to produce results better than binocular stereo. The neighborhood of any 3D point will have an optimal pairwise match, depending on baseline and visibility. We simply allow this pair to be self-selected for each pixel.

3. RESULTS

Below we present results from simulated data as well as from the mid and high complexity scenes discussed previously. We show that our multibaseline algorithm can produce more accurate range results over a greater dynamic range than binocular stereo. While this involves a much more exhaustive search of the 3D space than the comparable disparity search for binocular stereo, we feel that the increased computational complexity is not a factor for an evaluation tool. As mentioned previously, the work presented in this paper is intended for offline evaluation of real-time algorithms. In the following the algorithms described in Section 2.4 are referred to as "combined" since they use the product or sum of multiple correlation scores. The variant described in Section 2.5 is referred to as "minimum pair", since the best correlation score from each stereo pair is chosen at each pixel.

3.1 Simulation

We wish to validate two points in this section. First, we show that the joint correlation function described above produces better results than correlation from a single stereo pair. This is well known for trinocular stereo for example, but we demonstrate it in the context of our depth based search algorithms. Secondly, we show that depth based search can yield more accurate range estimates than disparity based stereo, since there is no *a priori* coupling between the choice of depth planes and integer disparity values of any stereo pair. For binocular stereo, depth estimates are typically biased towards integer disparity even after subpixel interpolation.

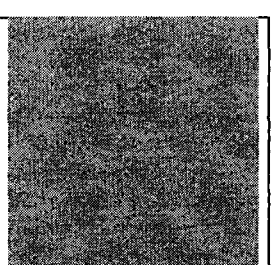

Figure 4: Image of textured plane

Our simulation setup consists of a virtual 5 camera rig with identical geometry to the one described above. The only difference is that our virtual cameras are 300 x 300 resolution for convenience of processing. We place 3 textured planes in space at uniform depths of 4, 8 and 16 meters from our camera rig. Image formation consists of projecting the texture into each camera according to predefined camera models. Note that we are careful to accommodate mixed pixel effects. The intensity registered at any pixel is the average intensity over all 3D points which project to that pixel. Figure 4 is an image taken by the central camera of the virtual

plane at 8 meters. Note that we do not add signal noise, since our goal in this exercise is not to evaluate the robustness of the correlation scheme.

In Figure 5 we show the correlation scores at the central pixel of the center camera for depths ranging from 5.5 to 10 meters. Shown are the scores for the individual binocular pairs (center-left, center-right, center-top, center-bottom) as well as the combined scores using both the product and sum approaches described in Section 2. Observe that the minima for both the product and sum are much sharper than that of any individual camera pair. It follows from the high curvature at the minima that subsequent quadratic interpolation produces better results than in the case of a single camera pair.

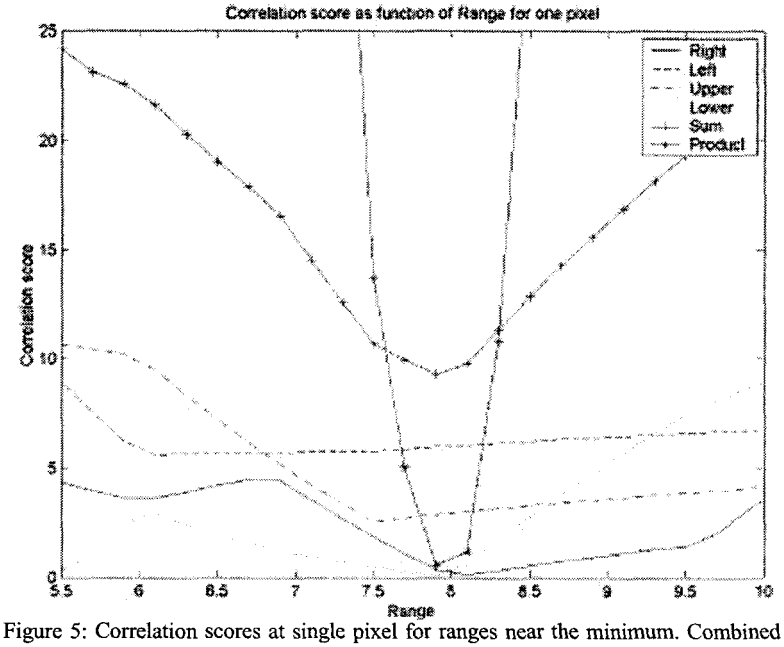

Figure 5: Correlation scores at single pixel for ranges near the minimum. Combined curves have much sharper minima than the binocular scores.

We computed depth maps using our binocular stereo algorithm with the 3 available lateral baselines of 10, 16 and 26 cm for each of the 3 textured planes mentioned above. For the multibaseline algorithm, we space the depth search planes at intervals of 10 cm. To prevent bias in the estimates, we arrange these planes so that they were not coincident with any of the 3D textured planes being recovered. In Figure 6 are tabulated the mean and standard deviation for recovered depths for each of the planes. Observe that the results for binocular stereo are highly variable. This is an artifact of good depth recovery at inverse integer disparity values. For our multibaseline algorithms, the results are consistently good and consistently better than for binocular stereo. While further evaluation with more sophisticated models is warranted, we feel this first step has been successful in validating our approach as an adequate ground truth mechanism.

		Target plane at 4 m	Target plane at 8 m	Target plane at 16 m
Binocular algorithm	10 cm baseline	3.9663 ± 0.2638 m	8.0265 ± 0.0376 m	16.0683 ± 0.1808 m
	16 cm baseline	4.0071 ± 0.1006 m	7.8589 ± 0.3284 m	15.5545 ± 0.3667 m
	26 cm baseline	3.9976 ± 0.0370 m	7.9243 ± 0.0491 m	15.8352 ± 0.1389 m
Multibaseline algorithm	Combined (sum)	4.0056 ± 0.0124 m	8.0130 ± 0.0185 m	16.0190 ± 0.0239 m
	Combined (product)	4.0024 ± 0.0051 m	8.0021 ± 0.0131 m	15.9997 ± 0.0243 m
	Minimum pair	4.0079 ± 0.0166 m	7.9999 ± 0.0107 m	16.0018 ± 0.0130 m

Figure 6: Recovered depths of target planes for the algorithms evaluated. Results are reported as mean ± standard deviation in m

3.2 Results with real imagery

We show the result of applying our multibaseline algorithms to the mid and high complexity scenes pictured in Figure 1. In Figures 7 and 8, we present the combined and minimum pair versions of our algorithm along with the original binocular depth maps. As with the binocular results described in Section 1, we used 11 pixel x 11 pixel correlation windows on 512 x 384 images.

The results of the two combined approaches (product and sum) are indistinguishable to the naked eye, so we only show the sum result. Note that despite the dense range maps for this variant, the sides of the image are truncated. As mentioned earlier, this is a consequence of requiring all 5 images to contain each point. In subsequent implementations, we plan to relax this constraint and use only the available data where all images fail to overlap.

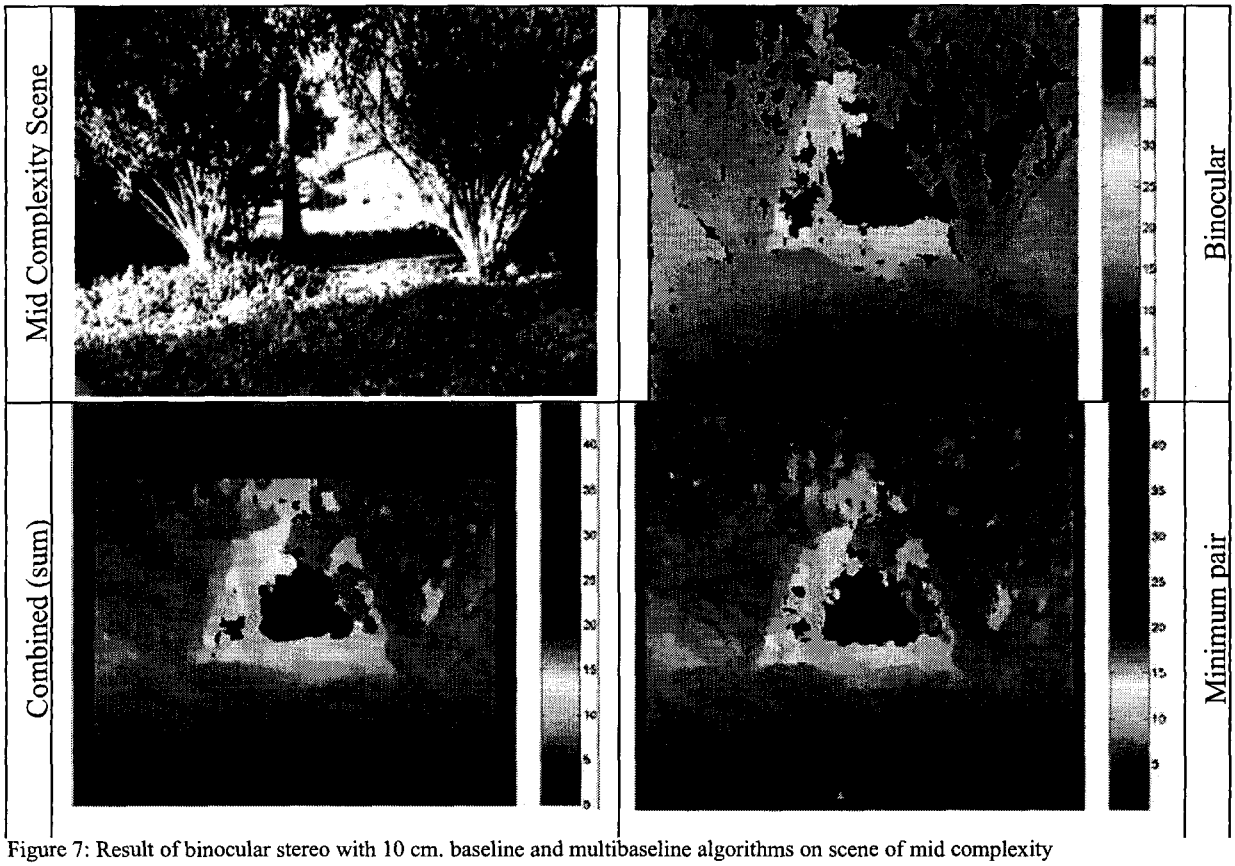

Figure 7: Result of binocular stereo with 10 cm. baseline and multibaseline algorithms on scene of mid complexity

In the mid complexity scene pictured in Figure 7, both multibaseline variants appear to produce not only denser but sharper results than the binocular algorithm. In the case of the combined approach, the sharpness results from the increased sharpness of correlation peaks. In the case of the minimum pair approach, there is a larger set of candidates for baseline (on a per pixel basis). Hence, more range values will have meaningful inverse disparity correspondents.

The high complexity scene is more difficult to interpret. The overhanging foliage makes the range map appear segmented. However, we have shown with 3D reconstruction of this data that the segmented components do, in fact, correspond to branches and overhanging leaves. This is in agreement with the recollection of those present during the data collection and can be seen more clearly when multiple frames are processed. In Figure 8, we present the high complexity scene in the same manner as in Figure 7. Notice that the combined result is much more uniform than either

the binocular or minimum pair result. Unlike the mid complexity case, there are regions of this image for which none of the available binocular pairs perform well. However, in combination, there is enough data to remove ambiguity and correlate correctly.

Figure 8: Result of binocular stereo with 10 cm baseline and multibaseline algorithms on scene of high complexity

4. CONCLUSION AND FUTURE WORK

We have presented a set of multibaseline, depth-based stereo algorithms and have demonstrated them with our 5 camera stereo evaluation rig. By adopting a looser definition of ground truth, we hope to use this system as a tool to evaluate improvements to our real-time, correlation based, binocular stereo algorithms. We have shown that the system is highly accurate when prior depth information is available. This indicates that high accuracy over a large range is to be expected if a dense collection of depth planes is chosen during evaluation. The additional computation cost is not a consideration for this system, since it is not intended for real-time use. Finally, we have shown with real imagery that denser data can be recovered. While this data may be difficult to judge from a 2D range map, it agrees well with the prior observations of the experimenters and holds up to 3D reconstruction.

In the future, we will undertake a more formal analysis of the depth vs. disparity issue. We will also conduct more detailed simulations involving more complex models. Finally, we intend to acquire more real data for testing. Our ultimate goal is to use this system on an ongoing basis to evaluate incremental improvements to our real-time stereo system. While our current configuration is adequate for this purpose, we will also study the optimal number and placement of cameras for the best depth recovery in anticipation of any future upgrade to our evaluation system.

REFERENCES

1. P. Bellutta, R. Manduchi, L. Matthies, K. Owens, A. Rankin, "Terrain perception for Demo III", *Proceedings of the Intelligent Vehicles Symposium*, Dearborn, Michigan, October 2000.

2. E. Krotkov, M. Hebert, and R. Simmons, Autonomous Robots, *Stereo Perception and Dead Reckoning for a Prototype Lunar Rover*, **Vol. 2, No. 4**, pp 313-331, December 1995.

3.C. Jennings and D. Murray, "Stereo vision based mapping and navigation for mo bile robots", *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 1694-1699, 1997.

4. C. Olson, L. Matthies, M. Schoppers, M. Maimone, "Stereo Ego-motion Improvements for Robust Rover Navigation", *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 1099-1104, Seoul, Korea, 2001.

5. O.D. Faugeras, *Three-Dimensional Computer Vision, A Geometric Viewpoint*, MIT Press, Cambride, Massachusetts, 1993.

6. E. Trucco and A. Verri, *Introductory Techniques for 3-D Computer Vision*, Prentice Hall, Upper Saddle River, New Jersey, 1998.

7. S. Goldberg, M. Maimone and L. Matthies, "Stereo Vision and Rover Navigation Software for Planetary Exploration," *Proceedings of the IEEE Aerospace Conference*, **vol. 5**, pp. 2025-2036, Big Sky, Montana, 2003.

8. V. Kolmogorov and R. Zabih, "Multi-camera Scene Reconstruction via Graph Cuts," *Proceedings of the European Conference on Computer Vision*, **vol. 3**, pp. 82-96, Copenhagen, Denmark, 2002.

9. K.S. Arun, T.S. Huang, and S.D. Blostein, "Least-Squares Fitting of Two 3-D point sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **vol. 14**, pp 239-256, February 1992.

10. S. Baker, R. Szeliski, and P. Anandan, "A layered approach to stereo reconstruction," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 434-441, Santa Barbara, California, June 1998.

11. M. Okutomi, T. Kanade, "A Multiple-Baseline Stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **vol. 15 No. 4**, pp. 353-363, April 1993.

12. J. Mulligan and K. Daniilidis, "Trinocular Stereo for Non-Parallel Configurations," *Proceedings of the IEEE International Conference on Pattern Recognition*, **vol. 1**, pp. 567-570, Barcelona, Spain, 2000.