Trinocular Stereo: a Real-Time Algorithm and its Evaluation

Jane Mulligan
Dept. of Computer Science
University of Colorado at Boulder
janem@cs.colorado.edu

Volkan Isler and Kostas Daniilidis
University of Pennsylvania*
GRASP Laboratory
{isleri,kostas}@grasp.cis.upenn.edu

Abstract

In telepresence applications each user is immersed in a rendered 3D-world composed from representations transmitted from remote sites. The challenge is to compute dense range data at high frame rates, since participants cannot easily communicate if the processing cycle or network latencies are long. Moreover, errors in new stereoscopic views of the remote 3D-world should be hardly perceptible. To achieve the required speed and accuracy, we use trinocular stereo, a matching algorithm based on the sum of modified normalized cross-correlations, and subpixel disparity interpolation. To increase speed we use Intel IPL functions in the pre-processing steps of background subtraction and image rectification as well as a four-processor parallelization. To evaluate our system we have developed a testbed which provides a set of registered dense "ground-truth" laser data and image data from multiple views.

1 Introduction

The power of today's general purpose and graphics processors and the high bandwidth of the recent Internet generations provide the necessary infrastructure for tele-presence systems. In this paper we describe the computer vision part of the realization of a new medium called tele-immersion. Tele-immersion enables users in physically remote spaces to collaborate in a shared space that mixes the local with the remote realities [10, 22]. An example of a tele-presence system [16] illustrated in Figure 1 brings two users from remote places to the "same" table. A real-time multiple view stereo reconstruction of a remote person is transmitted to the local site, combined with a stored off-line 3D-background and projected with stereoscopic projectors. The user wears polarized glasses and a 6-DOF head-tracker. The remote scene is always projected from the viewpoint of the local user as if he were looking through a window into the remote scene.

Figure 1. A local user on the left shares the same environment with a remote user on the right. A 3D description of the remote environment is projected stereoscopically on the screen from the viewpoint of the local user.

First attempts to realize immersive tele-presence involved slave stereo cameras that moved according to the local master's head and obtained a stereo-pair from the correct viewpoint. This *view-dependent* solution is impossible in a multi-user networked environment subject to latencies. In this paper, we address *view-independent* reconstruction from stereo in the context of tele-presence as described above. Having acquired a scene snapshot at a remote site we transmit it represented with respect to a world coordinate system. Displaying the 3D scene snapshot from a new point of view involves only primitive transformations hardwired in every graphics processor. In addition to real time response, the user should not experience depth distortion or outliers through her polarized stereo glasses. The basic question is how to achieve a perceptually best reconstruction in real-time.

The dense trinocular stereo algorithm we propose here is based on the maximization of a computationally expen-

*This work has been supported by NSF IIS-0083209, NSF IIS-0099201, NSF CDS-97-03220, ARO/MURI DAAH04-96-1-0007, Penn Research Foundation, and Advanced Network and Services.

sive correlation measure summed over the centre-right and the centre-left rectified image pairs. For the sake of speed, no ordering constraint is considered and there is no special handling of occlusions or specularities. Integer disparities are interpolated to obtain a subpixel estimate. Median filtering of the disparity map eliminates most of the outliers. Disparities can be filtered subject to the correlation value (goodness of fit) or the image gradient (matching feasibility). Two trinocular camera configurations are supported: an inline non-parallel triple and an L-shaped triple.

The second contribution of this paper is the evaluation of our results. We first introduced our performance metrics in [17]. Here, we present results on a new data-set of trinocular imagery and registered laser range data. Two metrics are introduced for evaluation: The first metric is still the classical view-independent world-centred nearest neighbour depth difference, which is critical to the performance of tele-collaboration systems where users interact with virtual 3D-objects whose visibility and collisions with the "real" scene must be monitored. The second error metric uses the imaging relationship to associate stereo and ground truth data. We compute the distance between a ground truth point which projects to an image pixel and the stereo depth point computed from the same pixel. Such depth errors along the viewing rays become obvious as new views are rendered according to the user's head motion. We use these metrics to compare the performance of our two trinocular configurations as well as examining the effects of correlation score, spatial gradient and median filtering as match quality metrics for our system.

In the next section we review the related work. Then we present a system overview and finally we describe the performance evaluation.

2 Related Work

We will not review the huge number of existing papers (see the annual bibliographies by Azriel Rosenfeld) on all aspects of stereo (the reader is referred to a standard review [6]). Application of stereo to image based rendering is very well discussed and reviewed in the paper by Narayanan and Kanade [18]. Stereo approaches may be classified with respect to the matching as well as with respect to the reconstruction scheme. Regarding matching we differentiate between sparse feature based reconstructions (see treatise in [7]) and dense depth reconstructions [20, 18]. Approaches such as [4, 28] address the probabilistic nature of matching with particular emphasis on the occlusion problem. Area-based approaches [13] are based on correlation and emphasize real-time responsiveness as we do. An approach with emphasis on virtualized reality is [18]. This system captures the action of a person from a dome of 51 cameras. Surround camera clusters are also very suitable for voxel-based techniques like space-carving [9, 24, 14, 5, 25]. The processing is off-line and in this sense there is no indication of how it could be used in telepresence beyond the off-line reconstruction of static structures.

Recently a number of authors have taken up the task of rigorous evaluation and comparison of stereo systems. Szeliski [26, 27] has proposed an evaluation method based on the discrepancy in intensities between a novel view and the reference view warped according to the computed stereo depth. The method is mainly applied to motion sequences where the novel view is a real image. In our case the novel views are arbitrary and for this reason we need ground truth to evaluate the warped reference appearance.

Leclerc et al. [11] introduced the notion of self-consistency. Again, the views checked for consistency are from the set used for computation and they can definitely not cover the viewing volume of a user in a tele-presence environment. However, like [26] it is a truthful measure if we do not have access to any ground-truth.

Banks and Corke [2] evaluate several dense correlation and nonparametric similarity measures as well as match validity measures commonly used to select valid correspondence. In the absence of ground truth, similarity measures are compared based on the percentage of computed disparities which pass the left-right consistency check, which generally identifies unreliable disparities due to half occlusions in the scene. Other validity measures addressed include image gradient (texture), match score, and locally anomalous matches. These measures are shown to have considerable overlap in the matches rejected.

Scharstein et al. [23] also evaluate dense binocular stereo systems, but they use a collection of image datasets with ground truth, as we do. They propose two quality measures: RMS disparity error with respect to the ground truth data and a percentage of bad matches, based on a disparity error threshold of 1 pixel. These are computed over the entire image as well as in regions identified as textureless, occluded or in the neighbourhood of a discontinuity.

3 System's Overview and Algorithm

For depth reconstruction, a cluster of 5 firewire cameras (Fig. 2) are arranged on an arc at 10° separation to 'surround' the user and prevent any break of presence due to a hard edge where the reconstruction stops. These cameras are used to calculate trinocular stereo depth maps from overlapping triples. For example the combined trinocular reconstruction illustrated in Figure 5, was computed from 3 triples $\langle C_0, C_1, C_2\rangle$, $\langle C_1, C_2, C_3\rangle$, and $\langle C_2, C_3, C_4\rangle$.

Both responsiveness and quality of depth data are critical for immersive applications. In order improve the frame rate of our system we have applied a number of techniques to reduce the weight of calculation, particularly in the expen-

Figure 2. Camera configuration, user view.

sive correlation matching required to generate dense depth maps. The simplest technique for the developer of course, is to purchase more and faster computers. We have built our system on 5 quad PIII 550 MHz servers (one for each reconstructed view) and parallelized our code accordingly.

One of the servers acts as a trigger server for the firewire acquisition. When all of the reconstructors are ready for the next frame the trigger server triggers all of the cameras simultaneously. Each computer grabs the image from 1 camera and transmits and receives the images needed by its neighbours and itself. Within each quad machine the images are divided into 4 equal bands and each processor is devoted to a particular band. The thread for each processor rectifies, background subtracts, matches, median filters the disparities and reconstructs points in its band of the image. When all processors have completed processing the texture and depth map are transmitted via TCP/IP to a remote renderer. This data is encoded as 3-(320×240) unsigned char image planes (RGB) of texture, plus one unsigned short image plane where $1/z$ values have been scaled into unsigned short, and background and unmatched foreground pixels are flagged. The total is about 3 Mbits per view per frame.

3.1 Background Subtraction

Our expectation for tele-immersion is that the workspace will contain a person in the foreground interacting with remote users, and a background scene which will remain more or less constant for the duration of a session. To obtain the speed and quality of depth points our application requires, we reconstruct the background scene in advance of the session and transmit it once to the remote sites. While the user moves in the foreground during a session, we need a method to segment out the static parts of the scene. We have chosen to implement a background subtraction method similar to that proposed by Martins et al. [12].

A sequence of N (2 or more) background images B_i are acquired in advance of each session. From this set we compute a pixelwise average background image $\overline{B} = \frac{1}{N}\sum_i B_i$. We then compute the average pixelwise difference between $\overline{B}$ and B_i, $\overline{D} = \frac{1}{N}\sum_i(\overline{B} - B_i)$.

During a tele-immersion session each primary image I is subtracted from the static mean background $I_D = \overline{B} - I$, a binary image is formed via the comparison $I_B = I_D > T \times \overline{D}$ where T is a configurable threshold (generally we use $T = 7$). These thresholded difference images are quite noisy. A series of erosions and dilations is performed on I_B in order to sharpen the background mask. The morphological operations are implemented by IPL separable convolutions. Typical results are illustrated in Figure 3.

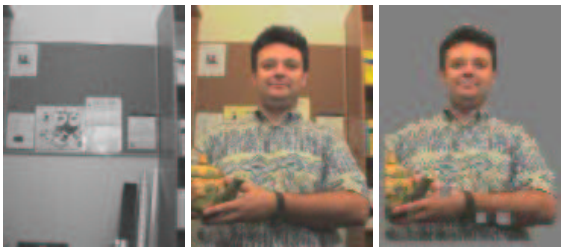


Figure 3. Background image, foreground image and subtracted result.

3.2 Matching Metric

In our efforts to maintain speed and quality in dense stereo depth maps we have examined a number of correlation correspondence techniques. We have concluded that the depth quality of trinocular Modified Normalized Cross Correlation (MNCC) is necessary to our application.

The reconstruction algorithm begins by grabbing images from 3 strongly calibrated cameras. The system rectifies the images so that their epipolar lines lie along the horizontal image rows so that corresponding points lie on the same image lines, thus simplifying the search for correspondences.

The modified normalized cross-correlation (MNCC) correspondence metric is:

$$corr_{MNCC}(I_L, I_R) = \frac{2\ \mathrm{cov}(I_L, I_R)}{\sigma^2(I_L)\ +\ \sigma^2(I_R)}. \qquad (1)$$

where I_L and I_R are the left and right rectified images over the selected correlation windows.

For each pixel (u, v) in the left image, MNCC produces a correlation profile $c(u, v, d)$ where disparity d ranges over acceptable integer values. Selected matches are maxima in this profile, which satisfy various 'peak' characteristics. Parabola fitting on the correlation profile is used to identify the subpixel peak location and calculate the subpixel disparity adjustment.

3.3 Trinocular Stereo

The trifocal constraint is a well known technique to refine or verify correspondences and improve the quality of stereo range data. It is based on the fact that for a hypothesized match $[u, v, d]$ in a pair of images, there is a unique location we can predict in the third camera image where we

Figure 4. Five camera views.

expect to find evidence of the same world point [6]. A hypothesis is correct if the epipolar lines for the original point $[u, v]$ and the hypothesized match $[u - d, v]$, intersect in the third camera image. The most common scheme for exploiting this constraint is to arrange the camera triple in a right angle (or L-shape), allowing matching along the rows and columns of the reference image [19, 1, 8].

Our initial telecubicle configuration, illustrated in Figure 2, placed cameras on an arc surrounding the user at the same level. This does not allow us to arrange or rectify triples of camera image planes such that they are coplanar, and therefore it is more expensive for us to exploit the trinocular constraint.

Following Okutomi and Kanade's observation [21], we optimize over the sum of correlation values with respect to the true depth value rather than disparity. Essentially we treat the camera triple $\langle L, C, R\rangle$ as two independent stereo pairs $\langle L, C_L\rangle$ and $\langle C_R, R\rangle$.

When revising our system design to parallelize and improve its speed, we discovered that by using foreground segmentation we need consider only one half to one third of the pixels in the reference image C_R. This makes it feasible to calculate the entire correlation profile for each pixel one at a time. To calculate the sum of correlation scores we precompute a lookup table of the location (u_{C_L}, v_{C_L}) in C_L corresponding the current pixel in C_R (based on the right-left rectification relationship). We also compute a linear approximation for the disparity $\widehat{d_L} = M(u_{C_R}, v_{C_R}) \times d_R + b(u_{C_R}, v_{C_R})$ at $[u_{C_L}, v_{C_L}]$ which arises from the same depth point as $[u_{C_R}, v_{C_R}, d_R]$. The maximum error in $\widehat{d_L}$ for our surround configurations and disparity ranges of $d_R = [-100, 100]$ is on the order of 10^{-13}. As we calculate the correlation score $corr_R(u_{C_R}, v_{C_R}, d_R)$, we look up the corresponding $[u_{C_L}, v_{C_L}]$ and compute $\widehat{d_L}$, then calculate the correlation score $corr_L(u_{C_L}, v_{C_L}, \widehat{d_L})$. We select the disparity d_R which optimizes

$$corr_T = corr_L(u_{C_L}, v_{C_L}, \widehat{d_L}) + \text{corr}_R(u_{C_R}, v_{C_R}, d_R)$$

The method can be summarized as follows:

Pixelwise Trinocular Stereo

Step 1: Precompute lookup table for C_L locations corresponding to C_R locations, and d_L approximation lookup tables M and b

Step 2: Acquire image triple $\langle L, C, R\rangle$

Step 3: Rectify $\langle L, C_L\rangle$ and $\langle C_R, R\rangle$ independently.

Step 4: Calculate foreground mask for C_R and R

Step 5: For every foreground pixel $C_{Rmask}(u, v)$

Step I: For every disparity $d_R \in D_r$
If $R_{\text{mask}}(u + d_R, v) \in$ foreground

Step i: compute $corr_R(u_{C_R}, v_{C_R}, d_R)$

Step ii: lookup $[u_{C_L}, v_{C_L}]$

Step iii: compute $\widehat{d_L} = M(u_{C_R}, v_{C_R}) \times d_R + b(u_{C_R}, v_{C_R})$

Step iv: compute $corr_L(u_{C_L}, v_{C_L}, \widehat{d_L})$

Step v: $corr_T = corr_L + corr_R$

Step vi: If $corr_T$ is a peak

Step 1: Fit parabola to find subpixel correlation peak and disparity adjustment d_{adj}

Step 2: Update $corr_{best} = corr_T$, $d_{best} = d_R + d_{adj}$

Step 6: Goto 2

3.3.1 L-Shape

We have implemented an algorithm for L-configurations to test its properties versus our existing system. We rectify the triple such that the upper (U) and lower-left (L) images are column aligned and simultaneously the left and right (R) images are row aligned. No explicit relationship is enforced between the upper and right images as in [1] because it introduces too much distortion for dense correlation stereo methods. The immediate advantage is that only 3 rectifications are required. Further, in the pixelwise approach, there is no need to lookup the centre-left index. However traversing the left image columnwise is less efficient in terms of memory access than row traversal. The algorithm otherwise proceeds as above, computing the maximum sum of MNCC

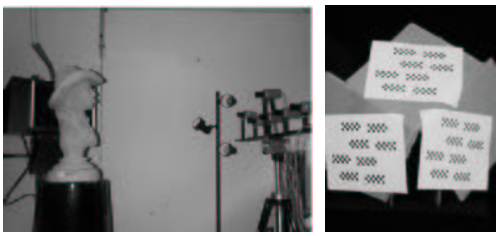
correlations over the disparity range for each pixel. Again a linear approximation for the corresponding upper disparity is calculated, given the lower disparity and the current pixel location.

An added challenge with our five camera cluster is the combination of multiple reconstructions into a single rendered view. We currently depend on the accuracy of our calibration to a common reference frame for all cameras. Figure 4 shows a set of camera views for a single frame in the current telecubicle camera cluster. From this image set 3 reconstructed views are calculated for overlapping triples. Figure 5 shows a profile rotation of the total set of 104,350 depth points calculated using trinocular MNCC for the frame in Figure 4.

Figure 5. Three trinocular reconstructions combined and rendered, rotated view.

4 Experiments

A number of recent papers [23, 2, 26, 27] have addressed the problem of evaluating and comparing dense stereo techniques. These authors propose metrics for evaluation, and emphasize problem areas for stereo, including occlusion and lack of texture. They also examine various match validity measures such as correlation score, spatial gradient and left-right checks. Most agree that comparison to ground truth is the gold standard of such evaluations. True ground truth is very difficult to obtain, but we have devised a method to acquire registered dense depth data of the same scenes we reconstruct by using a CyberWare Laser Scanner (http://cyberware.com/). The experimental setup is pictured in Figure 6. The acquired object is a concrete statue of Buffalo Bill smoking a cigar. Since the capture process requires a completely static scene through one or two image grab cycles and a laser scan (about 1 minute) no live subject was suitable. The Cyberware Head and Face 3D Color Scanner (Model 3030) has a motorized scanner head which



a. b.

Figure 6. Experimental Setup: a) Buffalo Bill statue, scanner head and camera cluster, b) 3D target for coordinate frame registration.

travels around the subject to be scanned in a 360° circle. It captures a cylinder of range values about 30 cm in height and 40-50 cm in diameter (sampling pitch $\theta \sim 1$ mm, y $\sim$ 700 um, z $\sim$ 100 um). We have therefore been limited to ground truth for the head of the statue only, although our camera images have a somewhat larger field of view. Images were captured using Sony DFW-V500 Firewire cameras connected to a Matrox Meteor II/1394 capture card.

To achieve registration of the laser and stereo coordinate frames we developed a 3D target with 3 planar surfaces (illustrated in Figure 6b). Calibration patterns with distinct coded targets are attached to each plane. The planes are not orthogonal because our calibration algorithm cannot extract the visible targets if they are too distorted. Each time the cameras were reconfigured the 3D target was placed in the workspace and a laser scan performed. Without moving the target, a set of images was captured. A separate calibration process was performed for the intrinsics and extrinsics of the cameras only. To register the 3D frames the visible targets were extracted for all camera views. The corresponding target points were reconstructed in the stereo frame from all pairs of cameras. The target points associated with each 3D target plane were used to estimate the equation of the plane in camera space $n_{Ci}\vec{x} - d_{Ci} = 0$. Similarly a subset of points belonging to each plane was extracted (by hand) from the scanner data, and the plane equations estimated ($n_{Si}\vec{x} - d_{Si} = 0$). We compose the matrices $N_C = [n_{C1}\ n_{C2}\ n_{C3}]$ and $N_S = [n_{S1}\ n_{S2}\ n_{S3}]$. We can then calculate the laser to camera transformation $T_{SC} = [R_{SC}\ t_{SC}]$ by estimating the closest rotation matrix R_{SC} satisfying $N_C = R_{SC}N_S$. This is given by UV^T where U, V are the left and right singular vector matrices of $N_C N_S^{-1}$. The translation can then be computed $t_{SC} = N_C^{-1}[d_{C1} - d_{S1}, d_{C2} - d_{S2}, d_{C3} - d_{S3}]^T$.

The data set acquisition proceeded as follows:

- the camera rig was configured.
- a sequence of camera calibration images was captured
- the 3D calibration images and laser scan were captured
- for each object data set:

Figure 7. Trinocular triple camera views and laser data.

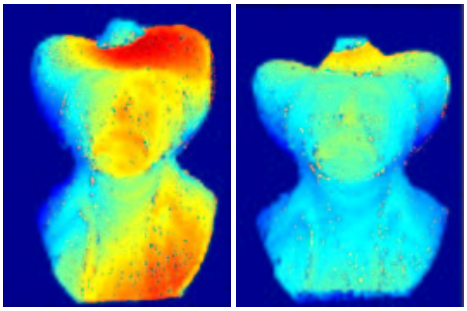


Figure 8. Disparity maps for inline and L-shape trinocular triples.

- the statue was positioned in the workspace
- the images were captured
- the laser scan was captured.

The registered data used in our experiments is illustrated in Figure 7. We computed the disparity maps illustrated in Figure 8 using our inline triple and L-shape trinocular stereo algorithms. The value of ground truth registered data is that it allows us to identify error sources and compare various instantiations of stereo reconstruction. In this paper we examine the errors arising due to occlusions in the scene and we compare the L-cluster to the inline cluster.

A somewhat subtle issue in looking at the ground truth data is how to identify "correspondences" between the laser and stereo data. One possibility is to associate each reconstructed point with the nearest laser point in 3D. This allows outliers to be associated with depth points that did not generate them, but all stereo points are accounted for. A second possibility is to project the laser points into the image and associate the stereo point arising from a pixel with the nearest laser point which also projects to the pixel. We illustrate both approaches in the plots below.

To illustrate the effects of various parameters and thresholds on the performance of algorithms with respect to ground truth error, we evaluate error at various levels of output density as proposed by Barron and Beauchemin [3]. By n% disparity density we denote the highest n% of image points sorted according to the goodness of matching given by a match validity metric such as MNCC correlation score. The dependence on the image gradient was studied in [17].

In Fig. 9 we show the reconstruction profiles for densities of 100% and 90% (1st-3rd and 2nd-4th respectively), and excluding, for the sake of visualization, fully or half-occluded points (1st-2nd and 3rd-4th image, respectively). By fully occluded points we mean the model points which when projected are occluded in both centre and right rectified images. By half-occluded points we mean the model points which when projected are visible only in the centre image. As we will also observe in later plots, the majority of the outliers lie in the 10% tail of the density distribution and therefore the 90%-density profiles are "cleaner". As expected, when we do not show the fully occluded points (1st/2nd images) we obtain reconstructions with more holes but less outliers.

Recent work on evaluation of stereo methods has mainly addressed binocular algorithms. Scharstein et al. [23] use percentage of points with disparity error less than 1, with respect to ground truth, as a quality metric for stereo. They also look at RMS disparity error. We use a trinocular stereo method in an effort to improve the quality of our depth maps. Figure 10 shows correlation score based density plots for a binocular MNCC stereo algorithm versus the trinocular system we describe. RMS absolute disparity difference for corresponding percentiles of points included is consistently lower for the trinocular system. Using Scharstein's metrics, the inline triple has 59.9% valid points (RMS disparity error 3.7 pix), while the L-shape has 69.7% valid disparities (RMS disparity error 5.4 pix). The binocular method had 59.8% valid disparities (RMS disparity error 2.3 pix), which is very similar to the trinocular method. The density plot illustrates this for the 100% included case, however it also demonstrates that trinocular sum of correlation values are more robust for eliminating bad matches than binocular.

Fig. 11 shows the difference between an inline and an L-shaped triple reconstruction. The L-shaped set-up exhibits more holes due to the nature of the occlusions in the particular statue: The probability that a point becomes half-

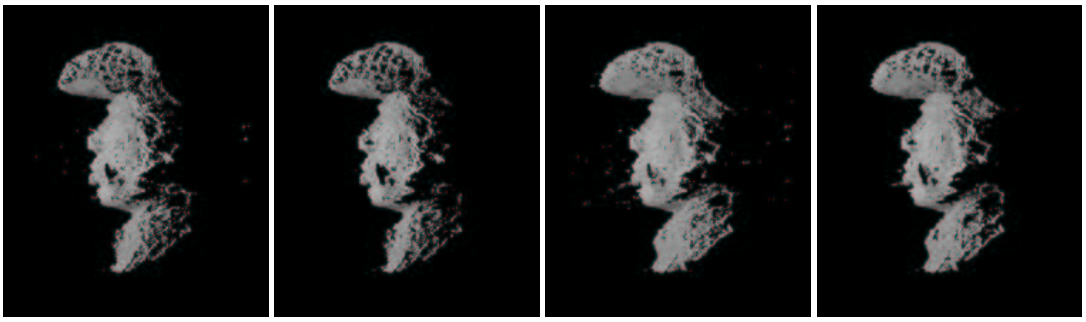


Figure 9. Reconstruction profiles: The *first* image shows a profile for 100% density without points occluded in the centre and the right original images, the *second image* shows a profile for 90% density without points occluded in the centre and the right original images, the *third* image shows a profile for 100% density without half-occluded points, and the *fourth* image shows a profile for 90% density without half-occluded points.

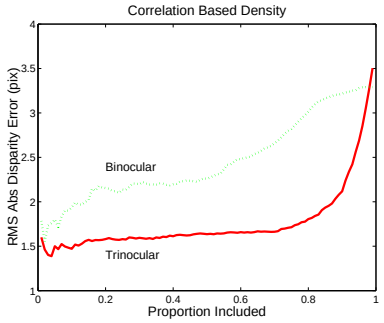


Figure 10. Correlation score based density plots for binocular and trinocular stereo.

occluded when adding a third camera in the vertical direction is higher than a when adding a camera in the horizontal direction.

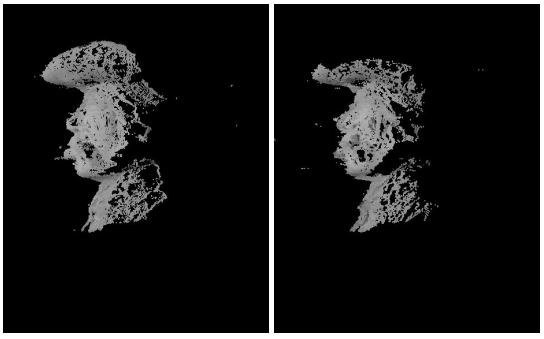


Figure 11. The reconstruction profiles for 90% depth density for an inline(left) and an L-shaped (right) configuration, respectively.

Figure 12 uses density plots to demonstrate the relevance of correlation scores and occluded points in reconstruction quality. We plot the proportion of points included by a correlation threshold against the root mean squared (Fig. 12 a,b) and median (Fig. 12 c,d) distance between corresponding laser and stereo points for both the inline and L configurations. The errors are calculated for the indicated proportion of points retained by fixing a threshold on the correlation score (ie. we calculate correlation thresholds which give us 20%, 30%, 40% etc of the data, then calculate the error metric for points which satisfy the threshold). Overall the median errors of 2–4mm are reasonable given the configuration of the rigs and the limits on ground truth registration. The L-shape reconstruction has consistently, if slightly lower median error, while its RMS error is higher than the inline configuration for higher percentages of included points. The RMS plots seem to suggest more outliers for the L-shape, but more systematic error for the inline configuration. The viewing ray (VR) correspondence method (12 b,d), gives higher error measures than Nearest Neighbour (NN) (12 a,c), probably because the ground truth registration was calculated using Euclidean distance.

Texture is crucial to correlation matching. In addition to correlation score, we can attempt to eliminate poor matches in low texture areas by examining the spatial gradient. Figure 13's density plots illustrate the effectiveness of using the spatial gradient for the RMS nearest neighbour distance metric. The effect on the RMS error is small until 90% of points are eliminated, when the error shoots up. This is probably the result of occlusion boundaries with high texture, and high uncertainty.

In their evaluation of stereo techniques, Banks and Corke [2] also look at means of identifying locally anomalous disparities. In our system we select matches based on the difference between the computed disparity and the median of its neighbours. Figure 14 illustrates density plots using the absolute difference between the disparity at each location and the median of its neighbours. Varying the me-

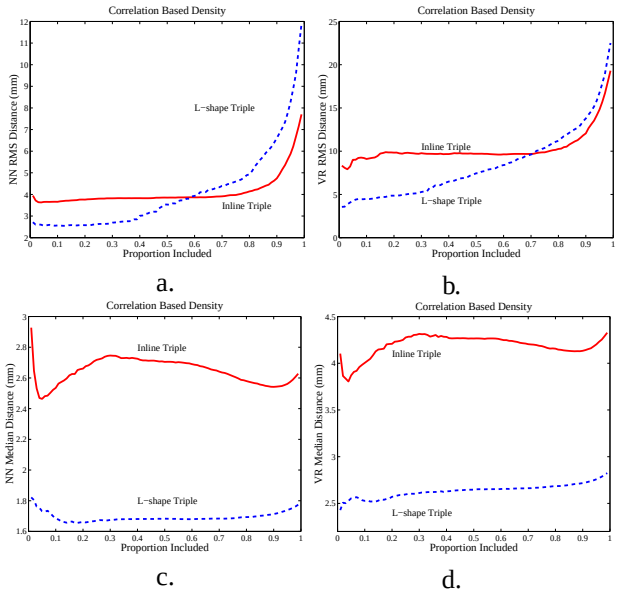


Figure 12. Root mean squared and median 3D distance between corresponding points vs. output densities obtained from descending correlation thresholds for inline triple and L-shape reconstructions. a) RMS 3D difference between nearest neighbours(NN), b) RMS 3D difference between points along the same viewing ray(VR), c) median 3D difference between nearest neighbours(NN), d) median 3D difference between points along the same viewing ray(VR).

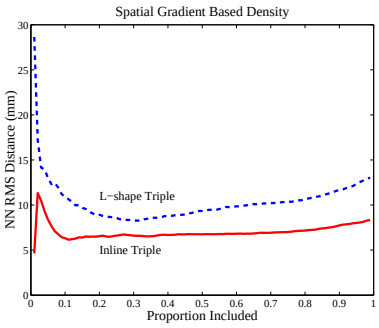


Figure 13. Spatial gradient based density plots for the inline and L-shape triple configurations.

dian filter difference threshold causes a sharp drop in error when the first 10% of points are eliminated, thereafter there is very little improvement.

In Figure 15 we reproduce Szeliski and Zabih's [27, 26] prediction error metric. Camera views -1, 0 and 1 represent the inline triple used to reconstruct the depth information.

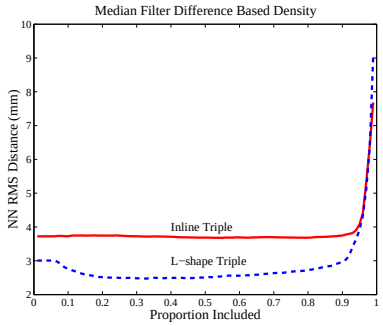


Figure 14. Median filter difference based density plots for the inline and L-shape triple configurations.

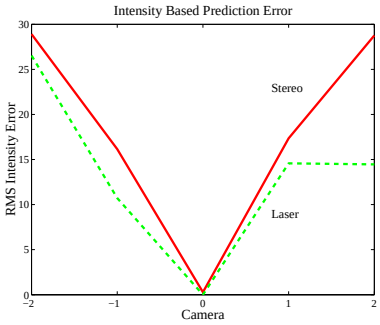


Figure 15. RMS intensity error for images from cameras -2 to 2, warped to the reference image from camera 0, according to the reconstructed depth data for the inline reconstruction.

We can see that the reference image (0) has essentially zero error. For the non-reference views, the error climbs to about 15 and for two unrelated views we see RMS error of about 28 intensity levels. The dashed curve plots the ground truth data used to warp the reference image into novel views. Again the error for the reference view is essentially zero, however the prediction error in novel views is still significant. It would appear that for our system, this metric reflects inconsistent image intensity and deficiencies in calibration almost as much as stereo accuracy.

Finally in terms of speed our system reconstructs 2-3 frames per second, depending on the contents of the scene and the size of the disparity search range. We run online at 320x240 pixel image size and 64 disparities. Typical timings for various algorithm stages are indicated in Table 1.

5 Conclusion

In this paper we presented a new rectification and matching algorithm for trinocular stereo. The algorithm works

Step	Tri-MNCC
Rectify	49
Background	31
Matching	350
Median Filter	8
Reconstruct	3
Total	456 ms

Table 1. Timings for online implementation of Trinocular MNCC.

for both inline non-parallel and L-shaped camera configurations. The emphasis was on optimizing the balance of speed vs accuracy required in tele-presence applications. Because of speed constraints we did not employ any global optimization like dynamic programming. To minimize matching ambiguities we employed three cameras and to maximize accuracy we used a computationally expensive similarity measure (Modified Normalized Cross-correlation) instead of simple measures like SAD or SSD. These results in a performance of 2-3fps (depending on the number of foreground points) on a quad-Pentium machine with a median 3D error of approximately 2mm.

We have contributed to the evaluation of stereo algorithms by building a unique experimental set-up with fully registered ground-truth laser data and image data, and by examining two possible 3D distance metrics (Nearest Neighbour and Viewing Ray). We studied the median and RMS of each error metric vs the depth density based on the correlation score. We also examined the value of spatial gradient and median filter metrics for selecting good matches using RMS Nearest Neighbour metric vs depth density. Because we know the ground-truth we could also observe the error behaviour in the non-occluded vs the half-occluded regions.

Our future work involves the fusion of several trinocular views, occlusion handling, the integration of silhouettes and correspondences, and the integration of motion and stereo [15].

References

[1] Nicholas Ayache. *Artificial Vision for Mobile Robots: Stereo Vision and Multisensory Perception.* The MIT Press, Cambridge, MA, 1991.

[2] Jasmine Banks and Peter Corke. Quantitative evaluation of matching methods and validity measures. *The International Journal of Robotics Research*, 20(7):512–532, July 2001.

[3] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12:43–78, 1994.

[4] P. Belhumeur. A bayesian approach to binocular stereopsis. *Intl. J. of Computer Vision*, 19(3):237–260, 1996.

[5] G. Cheung, T. Kanade, J. Bouguet, and M. Holler. A real time system for robust 3d voxel reconstruction of human motions. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 714–720, Hilton Head Island, SC, June 13-15, 2000.

[6] U. Dhond and J. Aggrawal. Structure from stereo: a review. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(6):1489–1510, 1989.

[7] O. Faugeras. *Three-dimensional Computer Vision*. MIT-Press, Cambridge, MA, 1993.

[8] Olivier Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. The MIT Press, Cambridge, MA, 1993.

[9] K.N. Kutulakos and S.M. Seitz. A theory of shape by space carving. In *Proc. Int. Conf. on Computer Vision*, pages 307–314, 1999.

[10] J. Lanier. Virtually there. *Scientific American*, pages 66–75, April 2001.

[11] Y. Leclerc, Q.T. Luong, and P. Fua. Measuring the self-consistency of stereo algorithms. In *Proc. Sixth European Conference on Computer Vision*, pages 282–298, Dublin, Ireland, 2000.

[12] Fernando C. M. Martins, Brian R. Nickerson, Vareck Bostrom, and Rajeeb Hazra. Implementation of a real-time foreground/background segmentation system on the intel architecture. In *IEEE ICCV99 Frame Rate Workshop*, Kerkyra, Greece, Sept. 1999.

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

[14] S. Moezzi, L.-C. tai, and Ph. Gerard. Virtual view generation from 3d digital video. *IEEE Multimedia*, 4:18–26, 1997.

[15] J. Mulligan and K. Daniilidis. Predicting disparity windows for real-time stereo. In *Proc. Sixth European Conference on Computer Vision*, pages 220–235, Dublin, Ireland, 2000.

[16] J. Mulligan and K. Daniilidis. View-independent scene acquisition for tele-presence. In *Proc. Int. Symposium on Augmented Reality*, pages 105–110, Munich, Germany, Oct. 5-6, 2000.

[17] J. Mulligan, V. Isler, and K. Daniilidis. Performance evaluation of stereo for tele-presence. In *Proc. Int. Conf. on Computer Vision*, Vancouver, Canada, Jul. 9-12, 2001. accepted.

[18] P. Narayanan, P. Rander, and T. Kanade. Constructing virtual worlds using dense stereo. In *Proc. Int. Conf. on Computer Vision*, pages 3–10, 1998.

[19] Yuichi Ohta, Masaki Watanabe, and Katsuo Ikeda. Improving depth map by right-angled trinocular stereo. In *Proceedings of the 8th International Conference on Pattern Recognition (ICPR'86)*, volume I, pages 519–521, Paris, France, Oct. 1986.

[20] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(4):353–363, 1993.

[21] Masatoshi Okutomi and Takeo Kanade. A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 15(4):353–363, April 1993.

[22] R. Raskar, G.Welch, M.Cutts, A.Lake, L.Stesin, and H.Fuchs. The office of the future: A unified approach to image-based modeling and spatially immersive displays. In *ACM SIGGRAPH*, pages 179–188, 1998.

[23] Daniel Scharstein, Richard Szeliski, and Ramin Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *Proceedings of the IEEE Workshop on Stereo and Multi-Baseline Vision*, Kauai, HI, Dec. 2001.

[24] S.M. Seitz and C.R. Dyer. Photorealistic scene reconstruction by voxel coloring. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 1067–1073, Puerto Rico, June 17-19, 1997.

[25] D. Snow, P. Viola, and R. Zabih. Exact voxel occupancy with graph cuts. In *IEEE Conf. Computer Vision and Pattern Recognition*, Hilton Head Island, SC, June 13-15, 2000.

[26] R. Szeliski. Prediction error as a quality metric for motion and stereo. In *Proc. Int. Conf. on Computer Vision*, Kerkyra, Greece, Sep. 20-23, 1999.

[27] R. Szeliski and R. Zabih. An experimental comparison of stereo algorithms. In *Workshop on Visual Algorithms*, 1999.

[28] C. Tomasi and R. Manduchi. Stereo without search. *Proc. European Conf. Computer Vision*, 1996.