Performance Evaluation of Stereo for Tele-presence

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

Abstract

In an immersive tele-presence environment a 3D remote real scene is projected from the viewpoint of the local user. This 3D world is acquired through stereo reconstruction at the remote site. In this paper, we start a performance analysis of stereo algorithms with respect to the task of immersive visualization. As opposed to usual monocular image based rendering, we are also interested in the depth error in novel views because our rendering is stereoscopic. We describe an evaluation test-bed which provides a world-wide first available set of registered dense "ground-truth" laser data and image data from multiple views. We establish metrics for novel depth views that reflect discrepancies both in the image and in 3D-space. It is well known that stereo performance is affected by both erroneous matching as well as incorrect depth triangulation. We experimentally study the effects of occlusion and low texture on the distributions of the error metrics. Then, we algebraically predict the behavior of depth and novel projection error as a function of the camera set-up and the error in the disparity. These are first steps towards building a laboratory for psychophysical judgement of depth estimates which is the ultimate performance test of tele-presence stereo.

1 Introduction

In a tele-presence environment, a user receives sufficient information about the task environment so that s/he feels physically present at the remote location. The key feature of a visually compelling tele-presence environment is that the scene is displayed stereoscopically and the rendered view changes according to the viewpoint of the user's head. An example of a tele-presence system [17] illustrated in Fig. 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 the he is 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. View synthesis methods such as [4, 15] are also view-dependent and less suitable for tele-presence applications because they hinder close interaction with virtual 3D-objects. In this paper, we address *view-independent* reconstruction from stereo in the context of tele-presence as described above.

What does performance analysis mean in the context of tele-presence and how does it differ from error analysis in classical stereo? In the related work section we will summarize multiple results on the evaluation of stereo algorithms. Most of these are based on the comparison between the estimated depth and ground-truth measurements. In an im-

*This work has been supported by NSF IIS-0083209, ARO/MURI DAAH04-96-1-0007, NSF CDS-97-03220, DARPA-ITO-DABT63-99-1-0017, Penn Research Foundation, and Advanced Network and Services. We highly appreciate the use of the TCS-1 Supercomputer at the Pittsburgh Supercomputing Center.

mersive tele-presence environment ideally –and this is our future research plan– we would first study the human perception of depth with stereoscopic projection. Since this is a long-term project and we need an evaluation now, we have set out to establish metrics which we are going to study analytically, experimentally, and with a new ground-truth data-set.

The first metric is still the classical view-independent world-centered depth difference, which might seem irrelevant in the sense of image based rendering, but will still affect 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 metric refers to novel virtual views (as perceived by the local user) and is the discrepancy between projected ground-truth and projected reconstructed points, referred to as residual flow in [23]. It captures error in rendered scenes resulting from depth error in reconstructed points.

The third metric also refers to novel views but is related to the fact that rendering in our system is stereoscopic. Even if the depth errors are along the viewing rays the user will still perceive the depth error with her polarized glasses. For example, when the user happens to view from the reference camera viewpoints the second metric is minimal but the third metric is significant.

To perform the experiments we established an experimental set-up by integrating a laser scanner and a cluster of cameras. Cameras are calibrated and registered with respect to the reference frame of the laser scanner. The range measurements obtained from the laser scanner serve as ground-truth data. To our knowledge, together with the University of Tsukuba manually measured ground-truth, our set-up is the first **dense** ground-truth data-set **registered** with respect to the camera cluster. The data used in this experiment are from multiple views of the face of a mannequin.

The above metrics are affected by all possible sources of error in stereo which we briefly discuss here. During the **correspondence** process, false disparities are due to noise in the images as well as due to errors in calibration which affect the rectification. Correspondence is not a well-defined problem since imaged world features may not be visible to be matched in all images, and because insufficient intensity variation results in infinite solutions. Depending on the choice of thresholds, we can include or exclude matched image points giving different output densities (numbers of matched points). We will vary this threshold-dependent output disparity density as a means of analyzing the proposed error metrics.

During **depth triangulation**, the error in the disparity can be amplified by the error in the projection matrices due to calibration. In our study we analytically and experimentally analyze the effect of varying depth, vergence, and combined vergence and baseline. We provide algebraic predictions for the first and second metrics, and verify their behaviour under simulation. The analytic error formulae are based on classical covariance propagation [26, 7].

We compare two algorithms for the matching step: The first algorithm is a near-real time trinocular stereo algorithm based on a modified normalized cross-correlation measure [16, 17]. The second is a publicly available C-implementation of the Roy and Cox algorithm [21].

In all comparisons, we show the histogram of an error metric. Error in stereo differs significantly from a Gaussian distribution regarding both shape of the probability density around the mode as well as outliers at the tail of the distribution. We strongly believe that neither RMS or median can give a fair representation of the error distribution.

In the next section we review the related work on evaluation of stereo-algorithms. In Section 3 we characterize the possible errors due to matching. In Section 4 we calculate the relative depth error as well as the discrepancy in a predicted view as a function of the camera poses and verify our formulae with synthetic simulations. Section 5 details the experimental set-up and the stereo algorithm used in the evaluation. Finally we apply the quality metrics on an extensive real data-set and compare the results with the theoretical prediction.

2. Related Evaluation Work

Since this is a paper on evaluation and not stereo algorithms per se we will refer only to standard textbooks [25, 6] and to the closest system to the description above (virtualized reality [18] and multi-baseline stereo [19]).

The closest evaluation approach is by Szeliski in [23, 24] and is based on the discrepancy in predicted intensities. This evaluation involves mainly motion sequences where the novel view is a real image. In our case the novel views are arbitrary and for this reason we need the ground truth to predict the reference appearance.

Our image-based metric intentionally does not measure the difference in intensities because this is really a perceptual question. We understand that the errors in low-textured areas are not as easily perceived as depth errors in textured neighborhoods, but modeling the perceivable intensity or the relation of a test image to the latter is still a matter of research.

The next closest approach is by Leclerc et al. [11] who 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 [23] it is a truthful measure if we do not have access to any ground-truth.

Analytical studies of the error in stereo have a long his-

tory [14, 9] and researchers have established the probability distribution function for a reconstructed point in space [2, 20]. The capability of controlling the relative pose on active binocular heads gave rise to very interesting studies [22, 5, 8, 13] on the role of fixation, vergence angle, and baseline length. Fundamentals on error propagation and modeling can be found in [26, 10, 7].

3 Correspondence Errors

As mentioned in the introduction, errors during the **correspondence step** arise because of insufficient information in the image intensity (or inability of an algorithm to deal with it) as well as because of noise in the intensities and the calibration-dependent rectification.

A unique solution for matching is impossible:

- if regions in one image do not correspond to any region in another image due to occlusion. This is a discrete optimization problem optimally solved with energy minimization or maximum flow - as is done in one of the algorithms we use for comparison [21].
- if the assumption about the same intensity or same filtered intensity response is not valid due to illumination changes or specular reflections.
- if the assumption about constant disparity in the considered local window is violated because of a depth discontinuity (for example at occlusion boundaries) or extreme perspective foreshortening or uncorrected radial distortion.

Matching is locally ambiguous with a "finite" ambiguity in case of periodic patterns and an "infinite" ambiguity in case of textureless areas. The latter is the most severe problem in stereo and cannot be tackled by any algorithm unless a global model or regularization assumptions are made.

The error classification above is data-driven and given the input images for a stereo algorithm, image areas can be pre-classified and results can be evaluated in separate areas as in the JISCT experiment [3] and in [12]. All of the above cases appear in our data and reflect in all three error metrics we use. We perform a boolean classification of image areas regarding occlusions and a continuous classification of image areas regarding intensity variation based on an image-gradient threshold.

4 Depth triangulation errors

In this section we investigate the error in depth produced during the triangulation step. As correctly pointed out by the reviewers, we study here the binocular case though our algorithm tested is trinocular. This is not a conflict because the trinocular algorithm yields only one correlation profile combined over all three images and one disparity value. This value can be mapped and used with any of the views. We use two views (central-right pair or left-right pair) and this difference is reflected in our study by the variation in the baseline. Given two viewpoints any additional viewpoint between the two can not constrain the error in the depth.

Consider the stereo setup in Figure 2 where two cameras (left and right) are verging with angles α_l and α_r. The origin of the world coincides with the image center of the right camera. The positive X axis of the world is the ray $\vec{O_r O_l}$. We assume that the vertical disparity is exactly zero and given by the calibration parameters used for the image rectification. This assumption is only approximate and we plan to investigate this issue in the future. For now, we constrain ourselves to the "flatland" ($X - Z$ plane) and consider only variations in vergence and baseline. Given a world point P,

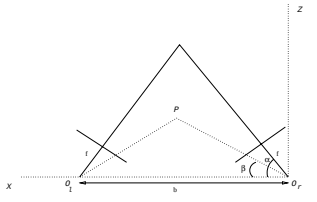

Figure 2. Stereo set-up for depth triangulation given matches

our measurement of depth can be expressed as a function of baseline b, vergence angles α_l, α_r, focal length f and the world coordinates (X,Z):

$$x_r = \frac{f\ (X \tan\alpha_r - Z)}{X + Z \tan\alpha_r}$$
$$x_l = \frac{f\ (Z + (b - x)\tan\alpha_l)}{b - X - Z \tan\alpha_l}$$

The actual triangulation algorithm uses the projection matrices from calibration but since the vertical disparity is assumed to be zero we can solve the system above instead.

If we replace x_l with $f \tan\beta_l$ and x_r with $f \tan\beta_r$ the triangulated point reads

$$X = \frac{b(\sin(\beta_r - \beta_l + \alpha_l - \alpha_r) - \sin(\beta_r + \beta_l - \alpha_l - \alpha_r)}{\sin(-\beta_l + \beta_r - \alpha_r + \alpha_l)}$$
$$Z = \frac{b(\cos(\beta_r - \beta_l + \alpha_l - \alpha_r) - \cos(\beta_r + \beta_l - \alpha_l - \alpha_r)}{\sin(-\beta_l + \beta_r - \alpha_r + \alpha_l)}$$

In dense area based approaches we can assume that x_l is exact and that matching errors reflect only in the disparity $d = x_r - x_l$. The variance of the pixel disparity can be approximately propagated into the depth variance as follows:

$$\sigma_Z^2 = (\frac{\partial Z}{\partial d}\sigma_d)^2. \quad (1)$$

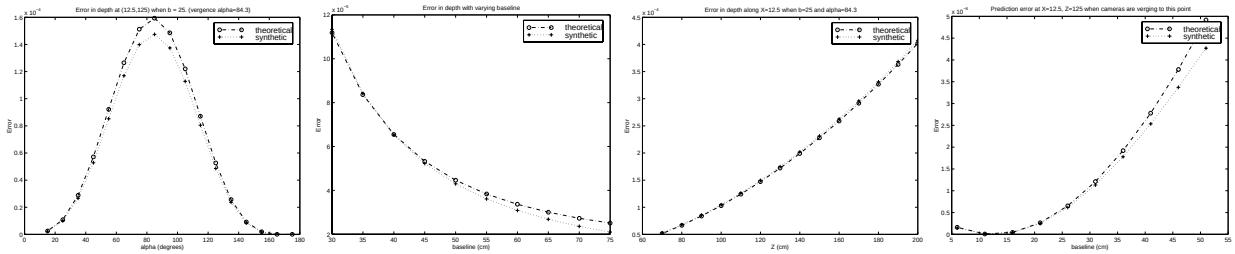

Figure 3. **Left:** Error at X=12.5cm and Z=125cm as a function of vergence angle α ($b = 25, \sigma_d = 14.8\mu, f = 0.6cm$). Cameras verge at this point when $\alpha = 84.2894^o$. **Middle-left:** Error at X=12.5, Z=125 when cameras are verging at this point with changing baseline.$\alpha_l = 180 - tan^{-1}[\frac{Z}{b-X}], \alpha_r = tan^{-1}[\frac{Z}{X}], \sigma_d = 14.8\mu, f = 0.6cm, b \in [30, 100]$. **Middle-right:** Prediction Error for the point X=12.5cm and Z=125cm. Left camera is centered at X=25cm, Z=0cm. The predicted view is at X=baseline, Z=0cm. All cameras verge at the point X=12.5 Z=125cm. $\sigma_d = 14.8\mu, f = 0.6cm$. **Right:** Prediction error along the line X=12.5cm as a function of Z. $\alpha = 84.289^o, b = 25, \sigma_d = 14.8\mu, f = 0.6cm, Z \in [60, 200]$. Cameras verge at Z=125. Each point in the synthetic plot represents 1000 trials.

According to Cramer-Rao if our estimator is unbiased the above expression gives the lower bound on the Z-variance. Throughout this section we are going to vary the ground truth depth, the vergence angles, and the baseline. For the non-varying parameters each time we assume values equal to the calibration parameters of our set-up. The focal length is 6 mm and consequently the pixel size is 15μm. The average depth point for a typical scene is Z=125cm. We consider only the cases where the cameras are verging symmetrically($\alpha_r = \alpha$, $\alpha_l = 180^o - \alpha$). All angles are shown in degrees.

We compute $(\frac{\partial Z}{\partial d})^2$ and plot the algebraically predicted **relative** depth error σ_Z^2/Z^2 which we call the theoretical error. To verify our formulae we also run a synthetic experiment 1000 times adding Gaussian noise $\mathcal{N}(0, \sigma_d)$ and taking the ensemble average for the relative error for the same depth point $E[\|\Delta Z\|^2]/Z^2$ which we call synthetic error.

In the first experiment we study the error in reconstructing a fixed world point when the fixation point of the cameras changes. This means we ask the question, given a person fixed at a position, should the cameras verge in front or behind the person. Regarding correspondence we know that the cameras should fixate on the person so that the disparity range is limited. However, regarding the triangulation step we observe both theoretically and synthetically that the error in depth is maximized when the cameras verge to the point of interest (Figure 3-left). Basu [22] showed that we could obtain the opposite effect if the resolution were not constant over the image but foveating.

In the second experiment, we keep the cameras fixating at the same point but increase the base-line. As intuitively expected (Fig. 3-middle-left), the relative depth error at the fixated point decreases with increasing baseline.

In the third experiment, we assume that the stereo set-up is fixed and we just let the considered point in the scene move along the symmetry axis $X = \frac{b}{2}$. As expected (Fig. 3-middle-right) the relative depth error increases as we move away from the cameras.

For the algebraic prediction of the second error metric, we assume that the novel view is in the plane XZ with center at the point $(X = b_v, Z = c_v)$ and vergence angle α_v. We compute the projection x_v of the ground-truth point (X, Z) on the new image plane and the projection $\hat{x}_v$ of the reconstructed point $(\hat{X}, \hat{Z})$. We then take the derivative of the difference $x_v - \hat{x}_v$ with respect to the disparity. The variance of this error metric is then

$$E[|x_v - \hat{x}_v|^2] = (\frac{\partial(x_v - \hat{x}_v)}{\partial d}\sigma_d)^2$$

To study the discrepancy in the novel view at a point on the symmetry axis (Fig. 3-right), we let the virtual camera move along the baseline and we observe that the minimum of the error metric for the novel views occurs when the virtual camera lies in the middle of the baseline.

5 Experimental Setup

It is difficult to make any metric statement about the quality of a stereo reconstruction, or the camera configuration from which it was generated without ground truth data to use for comparison. True ground truth is almost impossible to acquire, 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 4. We chose to use a mannequin as the subject of all of our dataset images, in part because we are particularly interested in reconstructing humans for communication in the context of tele-immersion. Since the capture process requires a completely static scene through one or two image

a. b.

Figure 4. Experimental Setup. a) mannequin, scanner head and camera cluster. b) 3D target for coordinate frame registration.

grab cycles and a laser scan (about 1 minute) no live subject was suitable. We have also removed the mannequin's wig, because neither the laser nor the stereo setup can extract reliable depth from hair. Unfortunately, although meeting the above conditions, the mannequin's face turned out to have non-negligible highlights which affect the matching quality.

The Cyberware Head and Face 3D Color Scanner (Model 3030) has a motorized scanner head which 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 mannequin only, although our camera images have a much larger field of view.

Our stereo rig is a heavy duty tripod with a 36×4 in aluminum base plate mounted on it. Images were captured using Sony DFW-V500 Firewire cameras connected to a Matrox Meteor II/1394 capture card. The limited size of the camera platform allowed 5 camera clusters for the 5° and 10° camera separations, but only 3 cameras for 15° and 20° configurations. The Meteor has 3 firewire ports and thus only 3 images could be captured at a time, requiring a cabling switch for 5 camera configurations.

Typically our tele-cubicle camera configurations "surround" the user on the arc of a circle about 1.3 metres in diameter. We use a trinocular stereo method to reduce ambiguities in hypothesized matches, and therefore outliers, with the trinocular epipolar constraint. Due to the surround configuration the camera triples are non-parallel and require some extra effort to calculate depth maps [16]: We treat the cameras as 2 independent pairs, left and centre, and centre and right. Correlation values from the two pairs are combined by precomputing correlation images for ranges of disparity in the left camera pair, then the computed correlation for each tested $[u_R, v_R, d_R]$ is added to that precomputed for the corresponding $[u_L, v_L, d_L]$. This results in large correlation lookup tables for the left image pair.

The second algorithm we use for comparison is Roy and Cox's [21]. Their publicly available implementation supports only binocular stereo. This algorithm transforms the correspondence problem into a maximum flow graph problem. The nodes of the graph are all possible (x,y,disparity) triples thus requiring full inter- and intra-scanline optimization. The algorithm uses a local coherence constraint.

We have run both algorithms on the same original resolution 640x480, however Roy and Cox's algorithm ran only on the portion including mannequin data which was 180x180. The trinocular algorithm ran with a disparity range of 64 pixels and Roy and Cox's with a disparity range of 120. The discrete subpixel disparity step for Roy and Cox's algorithm is 0.25 pixels. In our trinocular algorithm we vary the thresholds on correlation score and image gradient to select 'good' matches. There was no corresponding threshold in Roy and Cox's algorithm. The disparity smoothness parameter in Roy and Cox's algorithm was set to 2 (it was the only value from a range from 1-10 giving at least visually appealing results).

We measured the exact runtime for matching after rectification and before triangulation for both algorithms on one of the ES67 Alpha processors of an ES40 Alpha node of the TSC-1 at PSC. We normalized the measured time of the trinocular algorithm to the dimensions and disparity range of the settings we ran Roy and Cox's algorithm. Our algorithm took on the average 0.490 seconds versus 308 seconds of Roy and Cox's algorithm. Our algorithm is 600 times faster but it does not perform any global optimization.

Data was collected for a number of conditions of interest in telepresence configuration design. We varied the baseline of the stereo rig by changing the angular separation of the cameras on our nominal 1.3 m circle from 5° to 10°, 15°, and 20° (baseline approximately 11, 23, 34 and 45 cm respectively). We had no precise way to set the vergence points. We attempted to centre a calibration object in each image, and align the object in various views 3 at a time, by superimposing their luminance images as a single colour display. We positioned the calibration object at approximately 75 cm, 100 cm, 125 cm and 150 cm distance in front of the centre camera. Finally for each angular separation and vergence we positioned our mannequin at 5 rotations (-45°, -15°, 0°, 15°, and 45°) with the 0° position directly face on to the centre camera.

To achieve registration of the laser and stereo coordinate frames we developed a 3D target with 3 planar surfaces (illustrated in Figure 4b). 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. For each vergence and separation 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 visi-

ble 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 $t_{SC} = (n_{C1}n_{C2}n_{C3})^{-1}(d_{C1} - d_{S1}, d_{C2} - d_{S2}, d_{C3} - d_{S3})^T$.

The data set acquisition proceeded as follows:

- the camera rig was configured for a particular angular separation and vergence.
- a sequence of camera calibration images was captured
- the 3D calibration images and laser scan were captured
- for each of 5 rotations:
  - the mannequin was positioned in the workspace
  - the images were captured
  - the laser scan was captured.

5.1 Comparisons with Ground Truth

To illustrate the effects of various paramaters 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 [1]. By n% disparity density we denote the highest n% of image points sorted according to a figure of merit. Such a figure of merit can be the goodness of matching or the uncertainty in matching. Goodness of matching is given by the value of the normalized cross-correlation. Matching uncertainty is given by the the image gradient. The image gradient is proportional to the curvature of the local correlation profile except for positions of intensity maxima.

Throughout this section we will be showing histograms of errors instead of RMS or median of the error distribution. In the first group of plots we include only areas of the mannequin's face which are visible in all images. We study three error metrics as explained in the introduction and two selection criteria (correlation and image gradient) which give the output disparity density. The three error metrics are: 1. The 3D-discrepancy DIFF3D between reconstructed and closest ground-truth points; 2. The pixel discrepancy DIFFPIXNOV between reconstructed and ground-truth data projected on novel views; 3. The 3D discrepancy DIFF3DNOV between the reconstructed and the ground truth surface along rays in novel views.

We start with a comparison of 3D-discrepancies between our trinocular and Roy and Cox's algorithm in Fig. 5. We observe that the trinocular algorithm produces a higher number of outliers. However, the contribution of low error levels to the distribution is much higher in the trinocular algorithm which except the outliers resembles an one-sided gaussian with small standard deviation.

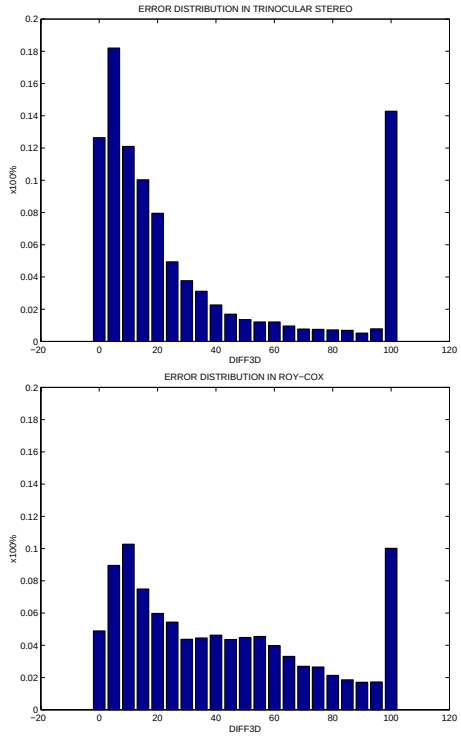

Figure 5. Histograms of the 3D-discrepancy for the trinocular algorithm (top) and Roy and Cox's algorithm (bottom).

We continue with a figure showing how the trinocular algorithm performs in areas we know from the ground-truth that they are occluded in one of the images and visible in the other. In Fig. 6 (left) we see that for densities up to 90% at least 80% of the occluded points are detected as occluded. However, we see a negative jump when we do not apply any thresholding where we expect that all points filling the occlusion area have wrong depth values. The same is true for Cox and Roy's algorithm for which we just present the histogram of 3D-discrepancies and see how erroneous they are.

We next show histograms of the first error metric (3D-discrepancy) produced for output densities from 50% to 100% - meaning that 50% (100% respectively) are left after thresholding with a correlation threshold (Fig. 6-middle)

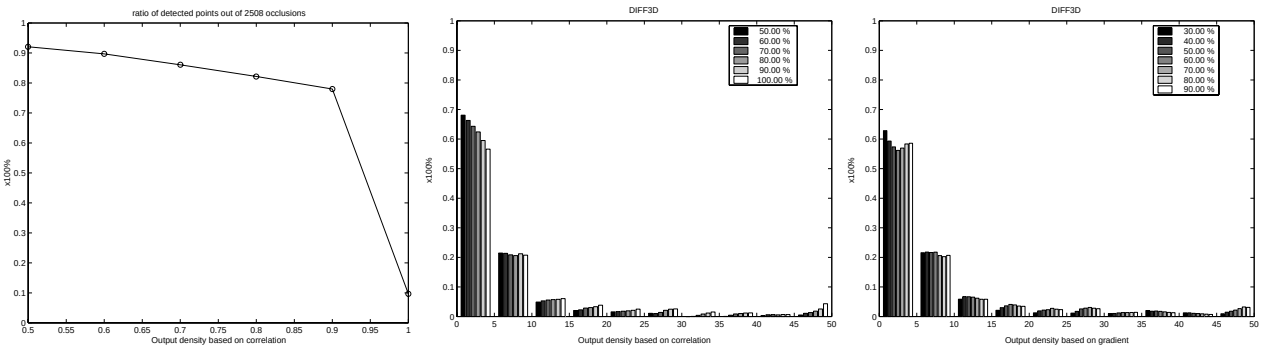

Figure 6. Left: Number of detected occlusion points inside the veridical occlusion area as a function of the output density deduced from a correlation threshold. Histograms of 3D-discrepancy (DIFF3D-metric) for several output densities (see legend) deduced from a correlation threshold (middle) and image gradient threshold (right).

and image gradient (Fig. 6-right). We consistently see that the higher is the output density the higher are contributions in the significant error levels. Throughout all comparisons we apply an image gradient threshold of 0.5 when we run over correlation thresholds and a threshold of 1.0 on the correlation (max=2.0) when we run over image gradient thresholds.

Last, we study the performance of the trinocular algorithm in novel views (Fig. 7). The three figure columns correspond to viewpoints at 100, 200, and 300mm, respectively. In all three columns we observe a shift of the concentration to the higher error levels. This shift is most prominent for the pixel discrepancy (first row) than for the 3D-discrepancies along rays (second row). There is not significant difference between correlation and gradient dependent densities, so due to space limitations we omit the histograms with the gradient dependent densities.

6 Conclusion

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,
- defining three quality metrics relevant to tele-presence,
- and showing experimentally and in comparison with another algorithm the performance of our trinocular algorithm with respect to several output disparity densities.

We plan an extensive psychophysical study how humans perceive distortions in depth using stereoscopic projection. As opposed to simpler psychology experiments we can provide controllable real dynamic data. We hope that such a study will result to metrics that reflect tele-presence performance and help in designing stereo algorithms for such tasks.

References

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

[2] S. Blostein and T. Huang. Error analysis in stereo determination of 3-d point positions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 9:752–765, 1987.

[3] R. Bolles, H. Baker, and M. Hannah. The jisct stereo evaluation. In *DARPA Image Understanding Workshop*, pages 263–274, 1993.

[4] E. Chen and L. Williams. View interpolation for image synthesis. In *ACM SIGGRAPH*, 1993.

[5] S. Das and N. Ahuja. Performance analysis of stereo, vergence, and focus as cues for active vision. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17:1213–1219, 1995.

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

[7] W. Foerstner. 10 pros and cons against performance characterization of vision algorithms. In *Workshop on Performance Characteristics of Vision Algorithms, Cambridge, UK, April 19–20*, 1996.

[8] N. Georgis, M. Petrou, and J. Kittler. Error guided design of a 3d vision system. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20:366–379, 1998.

[9] B. Kamgar-Parsi and B. Kamgar-Parsi. Evaluation of quantization error in computer vision. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 11:929–940, 1989.

[10] K. Kanatani. *Geometric Computation for Machine Vision*. Oxford University Press, Oxford, UK, 1993.

[11] Y. Leclerc, Q. 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.

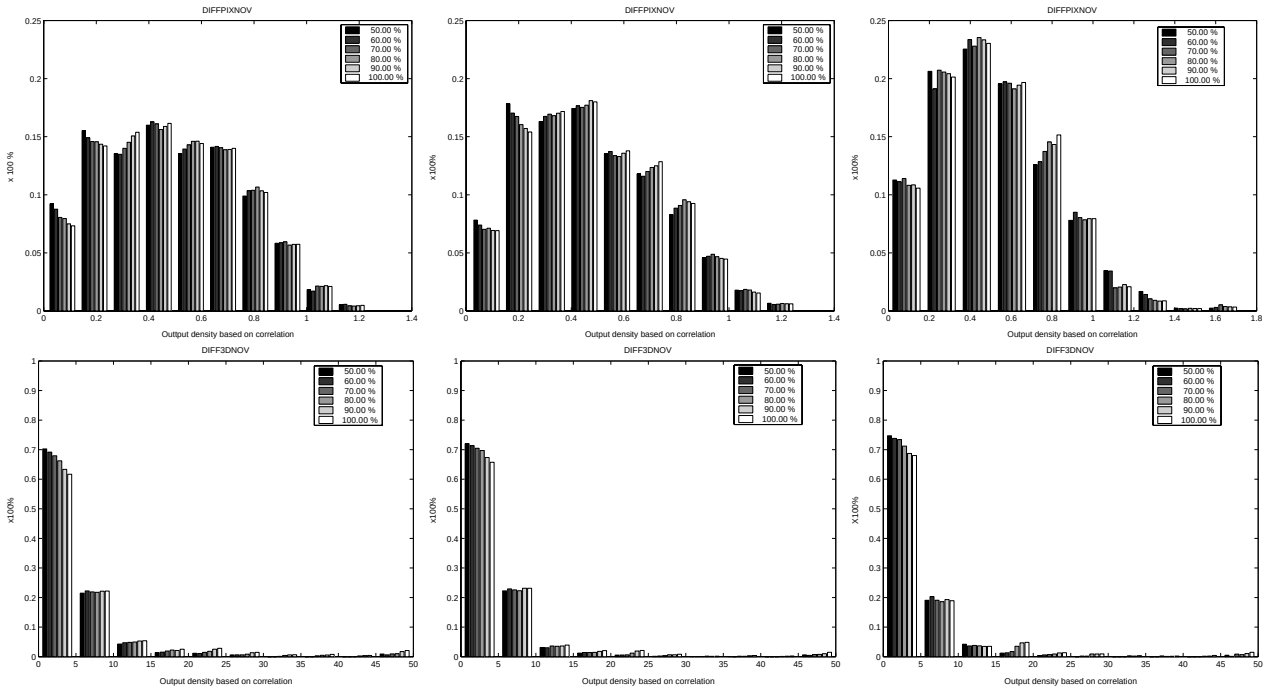

Figure 7. First row: Histograms of the pixel discrepancy (DIFFPIXNOV-metric) in three novel views for several output densities (see legend) deduced from a correlation threshold. Second row: Histograms of the pixel discrepancy (DIFF3DNOV-metric) three novel views for several output densities (see legend) deduced from a correlation threshold.

[12] M. Maimone and S. Shafer. A taxonomy of stereo computer vision experiments. In *Workshop on Performance Characteristics of Vision Algorithms, Cambridge, UK, April 19–20*, 1996.

[13] C. Y. M. Marefat and F. Ciarello. Error analysis and planning accuracy for dimensional measurement in active vision inspection. *IEEE Trans. Robotics and Automation*, 14:476–487, 1998.

[14] L. Matthies and S. Shafer. Error modeling in stereo navigation. *IEEE Journal of Robotics and Automation*, RA-3:239–248, 1987.

[15] W. Matusik, C. Buheler, R. Raskar, S. Gortler, and L. McMillan. Image-based visual hulls. In *ACM SIGGRAPH*, 2000. to appear.

[16] J. Mulligan and K. Daniilidis. Trinocular stereo for non-parallel configurations. In *Proc. Int. Conf. on Pattern Recognition*, pages 567–570, Barcelona, Spain, Sep. 1-3, 2000.

[17] 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.

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

[20] J. Rodriguez and J. Aggarwal. Stochastic analysis of stereo quantization error. *IEEE Trans. Pattern Analysis and Machine Intelligence*, pages 467–470, 1990.

[21] S. Roy and I. Cox. A maximum-flow formulation of the N-camera stereo correspondence problem. *Proc. Int. Conf. Computer Vision*, 1998.

[22] H. Sahabi and A. Basu. Analysis of error in depth perception with vergence and spatially varying sensing. *Computer Vision and Image Understanding*, 63:447–461, 1996.

[23] 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.

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

[25] E. Trucco and A. Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, Upper Saddle River, NJ, 1998.

[26] S. Yi, R. Haralick, and L. Shapiro. Error propagation in machine vision. *Machine Vision and Applications*, 7:93–114, 1994.