January 1997

Precision in 3-D Points Reconstructed From Stereo

Gerda Kamberova
University of Pennsylvania

Ruzena Bajcsy
University of Pennsylvania

University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-97-20.

Precision in 3-D Points Reconstructed From Stereo

Abstract

We characterize the precision of a 3-D reconstruction from stereo: we derive confidence intervals for the components (X,Y,Z) of the reconstructed 3-D points. The precision assessment can be used in data rejection, data reduction, and data fusion of the 3-D points. Also, based on the confidence intervals a bad/failing stereo camera pair can be detected, and discarded from a polynocular stereo system. Experimentally, we have evaluated the performance of the confidence intervals for Z in terms of empirical capture frequencies vs. theoretical probability of capture for a test, ground truth, scene. We have tested the interval estimation procedure on more complex scenes (for example, human faces), but since we do not have ground truth models, we have evaluated the performance in such cases only quantitatively. Currently we are developing "ground truth" models for more complex (such as general indoor) scenes, and will evaluate quantitatively the performance of the confidence intervals for the depth of the reconstructed points in the "automatic" rejection of 3-D points which have high degree of uncertainty.

Comments

University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-97-20.

Precision in 3-D Points Reconstructed from Stereo

MS-CIS-97-20 (GRASP LAB 417)

Gerda Kamberova, Ruzena Bajcsy

University of Pennsylvania
School of Engineering and Applied Science
Computer and Information Science Department
Philadelphia, PA 19104-6389

1997

Precision in 3-D Points Reconstructed from Stereo

Gerda Kamberova and Ruzena Bajcsy
E-mail: kamberov@cis.upenn.edu, bajcsy@cis.upenn.edu
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104

Abstract

We characterize the precision of a 3-D reconstruction from stereo: we derive confidence intervals for the components (X, Y, Z) of the reconstructed 3-D points. The precision assessment can be used in data rejection, data reduction, and data fusion of the 3-D points. Also, based on the confidence intervals a bad/failing stereo camera pair can be detected, and discarded from a polynocular stereo system. Experimentally, we have evaluated the performance of the confidence intervals for Z in terms of empirical capture frequencies vs theoretical probability of capture for a test, ground truth, scene. We have tested the interval estimation procedure on more complex scenes (for example human faces), but since we do not have ground truth models, we have evaluated the performance in such cases only qualitatively. Currently we are developing "ground truth" models for more complex (such as general indoor) scenes, and will evaluate quantitatively the performance of the confidence intervals for these more complex scenes. We give preliminary results demonstrating the use of the confidence intervals for the depth of the reconstructed points in the "automatic" rejection of 3-D points which have high degree of uncertainty.

1 Introduction

The motivation for our research is aiding the data reduction and fusion in the 3-D reconstruction from polynocular stereo [26].

Consider the stereo algorithm pictorially represented on Figure 1 (page 1). The input is a pair of digital images, Im_L and Im_R. First, matching of the two images is performed, and correspondence between pixels in left and right images is established, an *integer disparity map*, $\boldsymbol{D}$ is computed. Second, a subpixel disparity correction is calculated, and a *subpixel disparity map*, $\tilde{\boldsymbol{D}}$, obtained. Third, using the subpixel disparity map and calibration projection matrices for the camera pair, the 3-D points, (X, Y, Z) are reconstructed. Ideally, (X, Y, Z) should be the exact coordinates of points from the 3-D scene sensed by the cameras. In reality, due to errors, mostly due to matching, (X, Y, Z) may be inaccurate, and there may be gross errors far from the true point coordinates. This may create artifacts in the 3-D reconstruction that are not present in the real scene. Also, due to random noise in the digital images [11] , there will be random variation in the 3-D reconstruction. Our goal is to characterize the random errors in the final reconstruction, and account appropriately for these errors. Random errors in stereo result from random fluctuations in sensor measurements (i.e. left and right digital images, respectively). The point coordinates, (X, Y, Z), are calculated from the disparity through a deterministic mapping. Thus confidence intervals for the disparity are of primary interest. We propagate errors from digital images to subpixel disparity, and obtain estimates for the variance of the recovered subpixel disparity at an image position. The propagation of variance at this level is done using an error propagation technique described in [10]. We obtain confidence

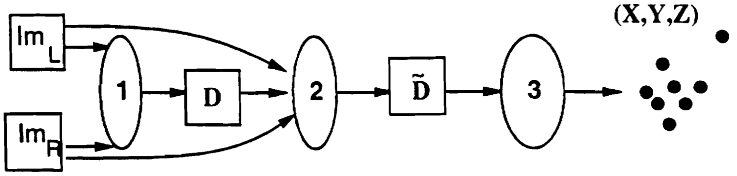

Figure 1: 3-D reconstruction from stereo flow diagram: **1** – matching; **2** – subpixel disparity computation; **3** – 3-D reconstruction

intervals for the disparity, and map them to confidence intervals for the coordinates of the 3-D reconstructed points.

The matching algorithm uses normalized cross-correlation [22],

$$c(\boldsymbol{I}_L, \boldsymbol{I}_R) = \frac{2\mathrm{Cov}(\boldsymbol{I}_L, \boldsymbol{I}_R)}{\mathrm{Var}(\boldsymbol{I}_L) + \mathrm{Var}(\boldsymbol{I}_R)}, \tag{1}$$

where $\boldsymbol{I}_L$ and $\boldsymbol{I}_R$ denote the pixel intensities over windows in the left and right images respectively, and Var and Cov denote the spatial sample variances and covariance over the windows. The matching procedure is responsible for outliers in integer disparity. Our experiments clearly show that these outliers are not affected by the random noise in the digital images. The cause of such outliers is the primary scene: occlusions, highlights, geometry, repetitive texture, etc. The correlation coefficient (1), is insensitive to small variations in image intensities. In particular, matching in the presence of strong texture signal is very stable, and not affected by the digital image noise present in our system. The integer disparity maps are stable. In order to investigate the effect of random errors on the stereo reconstruction in such cases, we must analyze the effect of random errors on the subpixel disparity correction[1].

Remark 1.1 We observe random fluctuations in the subpixel disparity map only. The empirical distributions of the subpixel disparities vary from pixel to pixel: we observe unimodal as well as a fraction of bimodal pixel distributions, with clearly separated modes. We are investigating these distributions. For the preliminary experiments reported in this paper, we have ignored the bimodal distributions, and have assumed normally distributed subpixel disparity noise.

Unless stated otherwise, the experiments presented here were performed on scenes with strong random texture (we have projected the texture). The reason for the use of the random texture is two-fold: (i) we use the procedures developed here for the data reduction and fusion in a large data set representing an office scene [27] which uses random texture projections[2]; and (ii) we want to isolate the random noise errors from gross matching errors. From our previous studies on camera/framegrabber noise [15], and the nature of the normalized cross correlation used for matching, we have established that the level of random noise in the camera/framegrabber system has no effect on the integer disparity in the case of high signal to noise ratio.

We will use the confidence intervals in processing 3-D points resulting from polynocular stereo as follows: (i) to remove some inconsistent, error points, i.e. *improving accuracy* (although this question has to be answered at a fundamentally different level, i.e. detecting outliers at disparity, matching level); (ii) to remove redundant consistent points, i.e. *data reduction*; and (iii) combining the data, i.e. *data fusion.*

Related Work

Our work is inspired by the call in the computer vision community for a systematic performance evaluation of vision algorithms [7, 9]. We study the precision of a stereo system, not the accuracy of the matching. For a review on matching see [4], or more recently [29, 16, 2]. Most of the parametric methods used in computer vision for dealing with random errors can be grouped into the following categories: (i) propagation of covariance [10]; (ii) maximum likelihood (MLE) type of methods (including maximum a posteriori (MAP), Bayesian, and markov random field (MRF) methods) [2, 25, 28]; (iii) Least square (LS), and mean square error (MSE) methods (including robust mean square estimators) [3, 18, 16, 21]; and (iv) confidence intervals based methods [14]. There is no clear cut division between some of these methods, for example, MRF methods can be viewed in Bayesian formalism, LS, MSE and MLE coincide in particular for normal models. A non-statistical method used is direct (min-max) interval propagation when feasible [19]. The authors underline that the method cannot substitute for the careful error evaluation, but may be used as a first order approximation.

MLE type methods, MAP, MRF and Bayesian, explicitly model expectations about the input scenes, and the output of the algorithms (given the input). Estimates are obtained optimizing on a criterion that the results should be faithful to the input model and to the observed data. Computationally these methods are very expensive (iterations, Monte-Carlo simulations, annealing). LS and MSE are quite popular, but often implicit assumptions about normal models are not stated. Very clear presentation is given in optical-flow estimation procedure by [23]. Robust estimators for optical-flow estimation are given in [3]. Unfortunately, these are computationally expensive.

What we consider here is yet another way of looking at the uncertainty and error evaluation process — through confidence intervals (confidence sets). The advantage of using confidence intervals is that together with a result

[1] The computation of the subpixel disparity correction in [26] is based on [6].

[2] The projection of random texture is not a very restrictive requirement. Currently there are commercially available devices for fast rate projection (60 projections per second), which is not obtrusive to the human eye.

we obtain also a measure of the reliability and the uncertainty in that result. To our knowledge, there are no reports in the literature on evaluation of 3-D reconstruction from stereo based on sound theoretically derived confidence intervals for the coordinates of reconstructed points. In [14] a fixed size confidence interval for depth from stereo is given. Three major differences between tease results and what we present here exist. In [14]: (i) confidence intervals only for the gross distance to a plane are derived (this is motivated by the particular application considered). (ii) only intervals Z are obtained (because, here, we do go into the effort of deriving confidence intervals for the disparity, we are able to derive intervals for X, Y, and Z); (iii) the confidence intervals are only for the particular, planar case; on the other hand for this case, and the range of depth covered, they are optimal in the minimax sense (they guarantee minimum maximum probability of capture). Currently we are working on deriving optimal intervals in disparity (a necessary step in this process is the derivation of the "current, non-optimal" intervals).

The interval estimators we propose are not robust (in the statistical sense [5]). We realize that although the assumptions we made are backed up by experimental data, this is not enough to guard against changes in the sensor behavior. Currently we are exploring the use of robust minima decision rules for the estimation of the disparity based on the theory developed in [12].

Organization of the paper: In Section 2 we present the image model. Section 3 gives the theory regarding the disparity model , the subpixel disparity correction solution, and the estimation confidence intervals in disparity (based on assumption of uncorrelated, locally constant digital image noise levels). Section 4 considers the 3-D reconstruction problem and the propagation of the disparity confidence intervals to confidence intervals for X, Y and Z. In Section 5 we show results for two types of experiments: (i) the evaluation of the confidence intervals for the depth Z by comparison of empirical capture frequency (ECF) and theoretical probability of capture (TPC); (ii) preliminary results showing the use of statistics of the confidence intervals in Z for rejection of points in the reconstruction (4% to 50% data reduction is achieved with no significant degradation).

2 The Image Model

We work with a simple additive noise model. We *assume* that the pixel random noise is signal-independent, uncorrelated in time and spatially[3], and with locally constant spatial variance (in a small neighborhood, for example 5x5, the pixel variance is almost constant). For a fixed image position $(u, v), 1 \leq u \leq M, 1 \leq v \leq N$, where MN is the size of the image, the intensity is

$$I(u, v) = \lambda(u, v) + \Delta I(u, v), \tag{2}$$

where $\Delta I(u, v)$ is zero-mean random noise, and $\lambda(u, v)$ is the unobservable (ideal) pixel intensity. $\text{Var}\, I(u, v) = \text{Var}\, \Delta I(u, v)$ (here Var is used in its precise statistical meaning) By using the camera properly (not overexposing), the pixel noise is independent and identically distributed (iid) over time. Under (2), the variance of the pixel noise can be estimated using the sample variance of a sequence of image differences,

$$\text{Var}(I_n^1(u, v) - I_n^2(u, v)) = 2\text{Var} I(u, v) = 2\text{Var}\Delta I(u, v),$$

where (u, v) varies over the pixel positions. To estimate $\text{Var}\Delta I(u, v)$: (i) the images of a flat field[4] are taken: (ii) pair-wise subtracted; and (iii) the sample variance is computed pixelwise over the differences; (iv) the noise variance $\text{Var}\Delta I(u, v)$ is estimated by half the sample variance of the differences.

3 From Pair of Images to a Disparity Map

An intrinsic geometric property of stereo images is that projections of the same 3-D point in left and right images lie on corresponding *epipolar lines* (resulting from intersection of the two image planes with the plane defined by the 3-D point and the two camera's projection centers),[20]. We assume that the cameras are in normal configuration (parallel optical axes, and with base line parallel to the image planes), this results in epipolar lines

[3] There are two issues to be considered with about pixel noise: (i) what can be stated about the noise level of a given pixel in different images (this is what we refer to as noise dependence/independence "over time"); and (ii) what can be stated about the noise level of different pixels in the same image (that is what we refer to as "spatial" dependence/independence).

[4] Fixed scene of uniform, high enough, illumination. The flat field should be taken at highest possible illumination level. The reason being that the actual signal-dependent component of the noise has Poisson distribution. Thus the higher the signal level, the higher the noise variance, and an estimate for the highest possible level of camera noise be obtained [24].

coinciding with the horizontal (u) coordinate lines[5]. Thus for a given pixel in the left image, the 2-D search for corresponding[6] pixel in the right image is reduced to 1-D search along the same number line in the right image. The result of the matching procedure is an *integer disparity map*: for every pixel in the left (rectified) image the amount by which its corresponding pixel in the right (rectified) image is offset. For many applications the integer disparity map has too coarse a resolution. Some means of evaluating subpixel, real number, disparity are attempted. The most common approach for recovering the subpixel disparity is to interpolate the integer disparity between neighboring pixels, [9, 18]. The subpixel disparity correction procedure used in [26] (described later) does not interpolate, it relies on model in [6]. It takes as input the integer disparity map and the original images and gives a subpixel disparity correction. The integer disparity map based on normalized cross correlation coefficient (1) is stable, not affected by the image noise. In order to obtain the effect of the random noise on disparity, we propagate the errors in the input images to that of the subpixel disparity errors. In this section, (i) we present the disparity model; (ii) give the subpixel disparity correction solution; (iii) estimate the variance in the subpixel disparity correction; and (iv) under a normal model for the subpixel disparity correction, we derive confidence intervals.

The Disparity Model

The computational problem is: given the left and right images, Im_L and Im_R, and an integer disparity map, $\boldsymbol{D}$, to compute the subpixel disparity $\tilde{\boldsymbol{D}} = \boldsymbol{D} + \boldsymbol{D}_\varepsilon$ where $\boldsymbol{D}_\varepsilon$ is the subpixel disparity correction, and also to obtain an estimate of the variance of $\boldsymbol{D}_\varepsilon$.

Assumptions made in [26], are: (i) local surface planarity over window of size $m \times m$; (ii) spatially uncorrelated pixel noise within an image and between left and right images; (iii) zero-mean pixel noise, and locally constant pixel noise variance (over the window); (iv) first order linear approximations are acceptable in the following local disparity model

$$I_R(u,v) = \beta I_L(A(u,v)(v - \tilde{\boldsymbol{D}}(u,v))) + \delta, \tag{3}$$

where I_R and I_L denote right and left (rectified) images; δ and β are parameters accounting for different offset and gain in the two sensors; $A(u,v)$ is a transformation between the neighborhoods in left and right images which preserves the center on the neighborhood. The equation is applied locally, over the window, generating a system the solution of which leads to the recovery of the subpixel disparity correction for the center pixel of the neighborhood.

Let (u,v) be integer local coordinates in the window, where u is an offset in row from the center, and v is an offset in column. The center pixel has coordinates $(0,0)$. Up to a first order approximation, the above nonlinear system (3) becomes a linear one. In local coordinates, the subpixel disparity correction system is as in [26]:

$$I_R(u,v) = (d_\varepsilon\ \beta, \gamma,\ \delta)\Big(\tfrac{\partial I_R}{\partial v}(u,v),\ I_L(u,v),\ \tfrac{\partial I_L}{\partial v}(u,v)v,\ 1\Big)^t, \text{ for } (u,v) \text{ in the window.} \tag{4}$$

Derivatives with respect to v denote image derivatives in the horizontal direction, d_ε is the error in the subpixel disparity correction for the center of the window in the left image, and γ is a parameter approximating the action of the map A, t denotes transpose of a vector. The derivatives $\frac{\partial I_R}{\partial v}(u,v)$ and $\frac{\partial I_L}{\partial v}(u,v)$ are approximated by finite differences.

Solving for the Subpixel Disparity Correction

The current implementation of the subpixel disparity estimation employs a linear least squares method. In matrix notation, (4) becomes

$$\boldsymbol{I_R} = \boldsymbol{\theta}\boldsymbol{\mathcal{I}}, \tag{5}$$

where $\boldsymbol{I_R}$ is $1 \times m^2$ observation vector of right image window gray values, $\boldsymbol{\mathcal{I}}$ is $4 \times m^2$ measurement matrix, and $\boldsymbol{\theta}$ is 1×4 parameter to be estimated,

$$\begin{aligned} \boldsymbol{\theta} &= (d_\varepsilon,\ \beta, \gamma,\ \delta) \\ \boldsymbol{I_R} &= [I_R(u,v)]_{(u,v)\in Window} \\ \boldsymbol{\mathcal{I}} &= [(\tfrac{\partial I_R}{\partial v}(u,v),\ I_L(u,v),\ \tfrac{\partial I_L}{\partial v}(u,v)v,\ 1)^t]_{(u,v)\in Window}. \end{aligned}$$

The function to be minimized is

$$F(\boldsymbol{I},\boldsymbol{\theta}) = (\boldsymbol{I_R} - \boldsymbol{\theta}\boldsymbol{\mathcal{I}})(\boldsymbol{I_R} - \boldsymbol{\theta}\boldsymbol{\mathcal{I}})^t, \tag{6}$$

[5] All experiments we did use verging cameras, so the images were rectified [1], and brought in the normal camera configuration context.

[6] Two pixels are corresponding if they are images of the same 3-D point.

and the solution to the least squares problem is

$$\boldsymbol{\theta} = -\boldsymbol{I_R}(\mathcal{I}\mathcal{I}^t)^{-1}, \tag{7}$$

where M^{-1} denotes the inverse matrix of M. The subpixel disparity at the center pixel in the left image window equals the integer disparity at the pixel plus the recovered d_ε (which is the first component of $\boldsymbol{\theta}$, (7)). We denote the $1 \times 2m^2$ vector of (rectified) image data in left and right images from the selected window with $\boldsymbol{I}$,

$$\boldsymbol{I} = \left(\{I_L(u,v)\}_{(u,v)\in Window},\ \{I_R(u,v)\}_{(u,v)\in Window}\right).$$

Next we propagate the variances of the random fluctuations in image windows in left and right images through the subpixel disparity correction procedure [10]. Careful attention must be paid here since the observation vector $\boldsymbol{I_R}$ and the measurement matrix $\mathcal{I}$ are not independent. Still, using the explicit form of the way the discrete image derivatives are computed, we are able to express $\mathcal{I}$ in terms of the primary image data (not the derivatives in horizontal direction themselves), and break the dependence (conditioned on the primary image data $\boldsymbol{I}$). Let $\boldsymbol{\Delta\theta}$ denotes the random error in the parameter estimation procedure. Up to a first order approximation,

$$\boldsymbol{\Delta\theta} = (\tfrac{\partial \boldsymbol{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{I},\boldsymbol{\theta}))^{-1}\tfrac{\partial \boldsymbol{g}}{\partial \boldsymbol{I}}(\boldsymbol{I},\boldsymbol{\theta})\boldsymbol{\Delta I}, \tag{8}$$

where

$$\begin{aligned} \boldsymbol{g}(\boldsymbol{I},\boldsymbol{\theta}) &= \tfrac{\partial F}{\partial \boldsymbol{\theta}}(\boldsymbol{I},\boldsymbol{\theta}) = -2(\boldsymbol{I_R} - \boldsymbol{\theta}\mathcal{I})\mathcal{I}^t \\ \tfrac{\partial \boldsymbol{g}}{\partial \boldsymbol{\theta}} &= 2\mathcal{I}\mathcal{I}^t \\ \tfrac{\partial \boldsymbol{g}}{\partial \boldsymbol{I}} &= -2\frac{\partial \boldsymbol{I_R}\mathcal{I}^t}{\partial \boldsymbol{I}} + 2\frac{\partial \boldsymbol{\theta}\mathcal{I}\mathcal{I}^t}{\partial \boldsymbol{I}} \end{aligned}$$

The estimate of the covariance in the subpixel disparity error correction, $\Sigma_{\boldsymbol{\Delta\theta}}$ is estimated by

$$\Sigma_{\boldsymbol{\Delta\theta}} = \mathrm{E}(\boldsymbol{\Delta\theta}^t\boldsymbol{\Delta\theta}) = (\tfrac{\partial \boldsymbol{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{I},\boldsymbol{\theta})^{-1})^t\,\tfrac{\partial \boldsymbol{g}}{\partial \boldsymbol{I}}(\boldsymbol{I},\boldsymbol{\theta})\,\Sigma_{\Delta \boldsymbol{I}}(\tfrac{\partial \boldsymbol{g}}{\partial \boldsymbol{I}}(\boldsymbol{I},\boldsymbol{\theta}))^t\,\tfrac{\partial \boldsymbol{g}}{\partial \boldsymbol{\theta}}(\boldsymbol{I},\boldsymbol{\theta})^{-1}.$$

Under the assumptions made about the noise in the digital images (zero-mean, uncorrelated, constant variance in window), $\Sigma_{\Delta \boldsymbol{I}} = s^2\boldsymbol{E}$, where $\boldsymbol{E}$ is the identity matrix, and s^2 is the common variance for the pixel noise in the window. The variance σ^2 in the subpixel disparity correction $\boldsymbol{D}_\varepsilon$ for the center pixel in left image window(i.e. d_ε) is the element of $\Sigma_{\boldsymbol{\Delta\theta}}$ with index $(1,1)$.

Confidence Intervals for Subpixel Disparity

A confidence interval for the disparity is a random interval (i.e. an interval whose bounds are functions of the random noise) [5]. As a random interval it is characterized by the probability (confidence level) with which it captures the disparity.

We assume that the subpixel disparity correction error is normally distributed (see Remark 1.1). From $\tilde{\boldsymbol{D}} = \boldsymbol{D} + \boldsymbol{D}_\varepsilon$, the subpixel disparity at (u,v) is normally distributed with variance $\sigma^2(u,v)$ estimated by the subpixel disparity correction variance at (u,v) (which has been computed).

For a given $0 < \alpha < 1$, we obtain $(1-\alpha)$-level confidence interval, $C(\tilde{\boldsymbol{D}}(u,v))$, for the subpixel disparity at position (u,v):

$$C(\tilde{\boldsymbol{D}}(u,v)) = [\,\tilde{\boldsymbol{D}}(u,v) - \zeta(\alpha/2)\sigma(u,v)\,,\ \tilde{\boldsymbol{D}}(u,v) + \zeta(\alpha/2)\sigma(u,v)\,], \tag{9}$$

where $\zeta(\alpha/2)$ denotes the $\alpha/2$ quantile of the standard normal distribution[7]. $C(\tilde{\boldsymbol{D}}(u,v))$ is a confidence interval for the disparity with probability of capture $(1-\alpha)$ and size $2\zeta(\alpha/2)\sigma(u,v)$. Before actually computing the disparity, we know that the probability that the disparity is in the interval equals $(1-\alpha)$, i.e. $\Pr[\tilde{\boldsymbol{D}}(u,v) \in C(\tilde{\boldsymbol{D}}(u,v))] = (1-\alpha)$. Note that there is inverse dependence between the size of the interval and its probability of capture. The probability of capture gives the reliability of the interval, while the size of the interval is related to the uncertainty in the value of the true disparity (larger interval size means higher reliability of the interval, but more uncertainty in the true disparity).

[7] The number $\zeta(\alpha/2)$ is such that the probability of a random variable with standard normal distribution to be less than $\zeta(\alpha/2)$ is $\alpha/2$.

4 From Disparity and Images to 3-D Points

The 3-D Reconstruction

Once the disparity is computed, the correspondence between pixels in left and right images is established. Theoretically, from here, the 3-D reconstruction is a simple geometric problem [8]: for a pair of corresponding points, (u_L, v_L) and (u_R, v_R), the 3-D point (X, Y, Z) which is projected onto (u_L, v_L) and (u_R, v_R), respectively, is the intersection of the two rays r_L and r_R, where r_L passes through the left camera optical center and (u_L, v_L), and r_R passes through the right camera optical center and (u_R, v_R). To get this intersection we need the camera calibration (rectified) projection matrices M_L and M_R[8]. The input to the reconstruction algorithm is a pair of (rectified) camera projection matrices, and the recovered disparity map $\tilde{\boldsymbol{D}}$. Computationally, the homogeneous coordinates of the 3-D point (kX, kY, kZ, k) are obtained from the calibration matrices and the disparity map through a linear map. Because of the random noise in disparity (also, imprecision in the calibration and rectification procedures), there will be imprecision in the position of the recovered 3-D points. We propagate the confidence intervals in disparity through that linear map, and from there by going to non-homogeneous coordinates, we obtain the confidence intervals for X, Y, and Z.

Confidence Intervals in 3-D

For a pixel (u, v) in the left image, the corresponding pixel in the right image is $(u, v - \tilde{\boldsymbol{D}}(u, v))$. These pixel positions relate to the 3-D point by

$$\begin{aligned} (u, v, 1)^t &= M_L(kX, kY, kZ. k)^t & (10) \\ (u, v - \tilde{\boldsymbol{D}}(u, v), 1)^t &= M_R(kX, kY, kZ, k)^t. & (11) \end{aligned}$$

Solving (10-11) for X, Y, and Z we obtain,

$$(kX, kY, kZ, k)^t = Q(u, v, v - \tilde{\boldsymbol{D}}(u, v), 1)^t, \quad (12)$$

where Q^{-1} is a matrix with first, second, and forth rows equal to the first, second, and forth row of M_L, respectively, and third row equal to the second row of M_R.

In order to propagate the uncertainty in disparity through the reconstruction algorithm, we rewrite (12) as

$$(kX, kY, kZ, k)^t = S(u, v, 1, \tilde{\boldsymbol{D}}(u, v))^t. \quad (13)$$

where the matrix S is defined as follows

$$S(i, 1) = (Q(i, 1),\ Q(i, 2) + Q(i, 3),\ Q(i, 4),\ -Q(i, 3)),\ i = 1, 2, 3, 4.$$

We propagate the intervals for disparity to confidence intervals for Z, $C_Z(u, v) = [L_Z(u, v), U_Z(u, v)]$,

$$\begin{aligned} L_Z(u, v) &= \frac{S(3, :)\left(u, v, 1, C_l(\tilde{\boldsymbol{D}}(u. v))\right)^t}{S(4, :)\left(u, v, 1, C_l(\tilde{\boldsymbol{D}}(u, v))\right)^t} & (14) \\ U_Z(u, v) &= \frac{S(3, :)\left(u, v, 1, C_u(\tilde{\boldsymbol{D}}(u. v))\right)^t}{S(4, :)\left(u, v, 1, C_u(\tilde{\boldsymbol{D}}(u, v))\right)^t}, & (15) \end{aligned}$$

where, $C(\tilde{\boldsymbol{D}}(u, v)) = [C_l(\tilde{\boldsymbol{D}}(u, v)), C_u(\tilde{\boldsymbol{D}}(u, v))]$ is the confidence interval in disparity, and $S(i, :)$ is the ith row of the matrix S, $i = 1..4$. The lower and upper confidence bounds in X and Y are obtained by substituting $S(3, :)$ with $S(1, :)$ and $S(2, :)$,.respectively, in (14-15).

Remark 4.1 In this propagation, it is important to pay attention to the Jacobian of the map. The sign of the Jacobian does not depend on the image values and the disparity, but only on the calibration projection matrices and the image positions. The sign of the Jacobian can be precomputed. We do not propagate intervals for positions (u, v) for which the Jacobian is 0; and for positions for which the Jacobian is negative, we switch the order of the bounds (lower and upper) after propagation.

The confidence intervals for disparity and for each of X, Y and Z have the same probabilities of capture (when well defined), i.e.

$$\Pr[D \in C(\tilde{\boldsymbol{D}}(u, v))] = \Pr[X \in C_X(u, v)] = \Pr[Y \in C_Y(u, v)] = \Pr[Z \in C_Z(u, v)].$$

Although $C(\tilde{\boldsymbol{D}}(u, v))$ is symmetric around $\tilde{\boldsymbol{D}}(u, v)$, the intervals for the 3-D coordinates are not symmetric around the reconstructed coordinates.

[8] These 3 x 4 matrices describe the projection of 3-D points onto the left and right image planes, respectively. They are recovered through a calibration process prior to the stereo reconstruction.

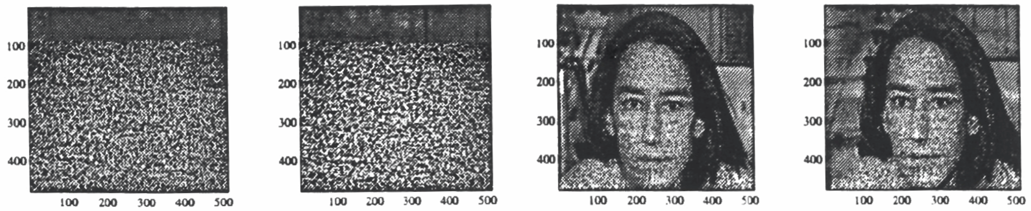

Figure 2: Original stereo pair images. Left: Planar scene. Right: A face.

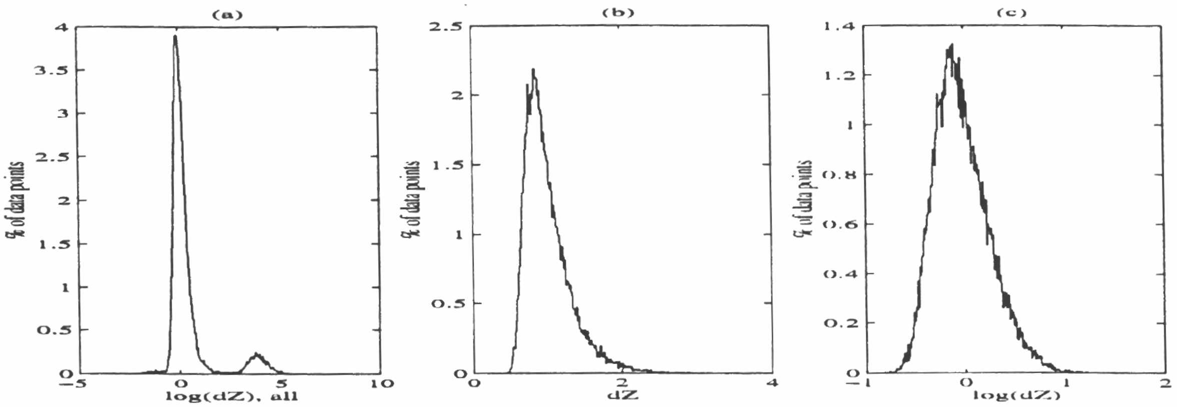

Figure 3: Pair(1,4): Histograms of confidence interval sizes, dZ, in mm at 0.68 TPC. (a) a histogram over all intervals, log scale; (b) a histogram, only over intervals of reconstructed points, corresponding to image projection of the plane; (c) same as (b), but in log scale. We differentiate between (a) and (b)-(c) since the image of the plane does not coincide with the whole digital image.

5 Experiments

Here we present results for two particular scenes: a planar surface, and a human face (see Figure 2). For the planar surface we present quantitative data, for the human face we show only qualitative results. In our experiments we consider only the confidence intervals for Z. The reason for that is that we have a ground truth model for the depth Z of the test target (pointwise), but not in X and Y.

No post-processing or manual adjustment has been done on any of the results and figures presented in this paper.

Hardware
We have tested the theory using four video cameras SONY XC77-RR (pairing the cameras in different ways, changing vergence, base line, recalibrating), and variety of scenes. The camera pixel clock frequency is 14.318MHz. We used a single Data Translation DT1541 framegrabber with sampling frequency about 10MHz. We are aware of the aliasing artifacts in this camera-frame grabber system but they are negligible in our scenes with high-contrast texture. A pair of cameras were usually 6-10cm apart, non-verging, or verging angle of approximately 40°. The cameras view a volume about 80cm away from the base line of a pair. Each time a camera pair is reconfigured, it is strongly calibrated [26]. We used a slide projector to project random texture onto the scenes (for that purpose we have printed a random texture pattern on slide).

5.1 Evaluation of the Confidence Intervals in Depth

Experimental setup
The goal of this series of experiments was to assess the precision of the reconstruction algorithm in terms of the confidence intervals in depth, Z, and to evaluate the performance of the confidence intervals (for varieties of stereo pairs). The test scene was a vertical planar surface coinciding with the coordinate plane $(X-Y)$. The plane was positioned at precise, and known position for Z. A random texture was projected on the planar surface (Figure 2), and a sequence of 100 stereo pair images (for each pair under study) was taken.
Ground test reconstruction: The reconstruction algorithm, [26], augmented with the confidence intervals estimation procedure described in this paper was run for one stereo pair images. Confidence intervals, C_X, C_Y, C_Z,

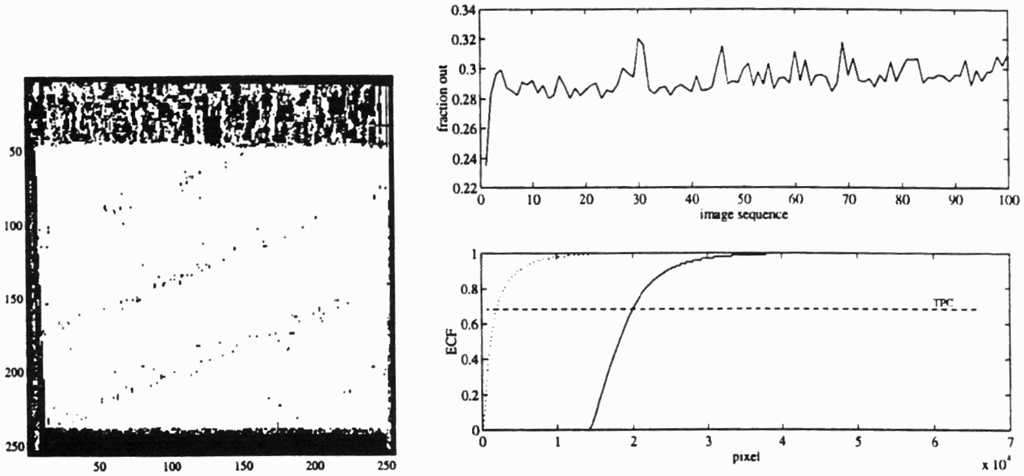

Figure 4: Pair(1,4): Data for ECF, planar scene. The image on the left shows in white the pixels for which the intervals for the corresponding 3-D depth have ECF$\geq$TPC= 0.68. The top graph on the left shows for each of the 100 reconstructions, the fraction of pixels for which the corresponding reconstructed Z is not captured by C_Z. The bottom graph shows sorted ECF of the intervals: the solid line is for the intervals corresponding to all reconstructed points; the doted line is for those corresponding to pixels projections of the plane; and the dashed line represent the TPC=0.68.

at specified level of confidence, $(1-\alpha)$ (i.e theoretical probability of capture(TPC) $(1-\alpha)$) for disparity, and the 3-D coordinates X, Y, Z were computed.

Computation of ECF: Next, the reconstruction algorithm was run 100 times, on each of the 100 stereo pairs. For each stereo pair and for each pixel position in the left image of the pair, we counted how many times the newly computed depth Z for that pixel position fell within C_Z. The final counts were divided by the total number of sample runs, thus the *empirical capture frequencies* (ECF) were computed.

Results for one stereo pair

Figure 3 shows histograms of the sizes, *dZ*, in mm for the confidence intervals with *theoretical probability of capture* (TPC) of 0.68, for one fixed stereo pair (labeled $Pair(1,4)$) and the planar scene Figure 2. Figure 4 shows various statistics for the empirical capture frequency for $Pair(1,4)$. The pixels which corresponds to the projections of the plane (with random texture) have better ECF. In the bottom graph the pixel positions for which the ECF is below TPC have non-satisfactory performance.

Results for four different stereo pairs

We performed the same evaluations for four other stereo pairs, labeled Pair(1,2)', Pair(1,4)', Pair(2,3)' and Pair(3,4)' (In, particular, Pair(1,4) and Pair(1,4)', used the same cameras but in a different configuration). This time the random texture plane was covering the complete view of all cameras, so comparison could be drawn. Confidence intervals, with probability of capture 0.68, were computed for the reconstructed points for each pair, and the ECF were calculated during 100 runs of the polynocular stereo reconstruction algorithm [26]. Statistics for ECF, and for the confidence interval sizes, *dZ*, in mm, are summarized in Table 5.1. Note that the TPC is less than the actual ECF. TPC underestimates the performance of the intervals. Under the same model, deriving

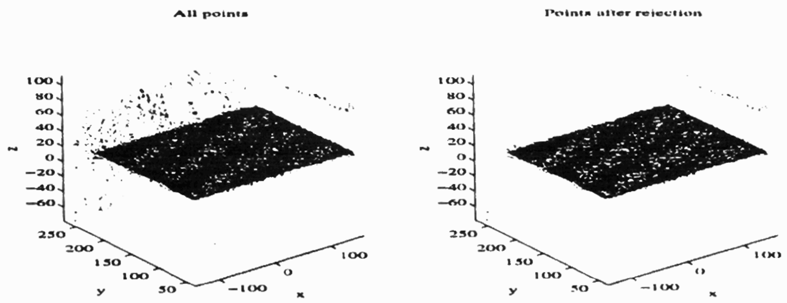

Figure 5: The reconstructed points (X, Y, Z) shown as 3-D plots. Left: all reconstructed points. Right: Only the points for which the confidence interval sizes are less than the mean size.

		ECF				*dZ*mm		
Pair	*mean*	*std*	*median*	*min*	*max*	*mean*	*std*	*median*
(1,2)'	0.6838	0.3214	0.8000	0.1598	4.2830	0.9234	0.3043	0.8722
(1,4)'	0.7549	0.2842	0.8700	0.1180	226.5228	0.7082	0.9669	0.6666
(2,3)'	0.7327	0.2912	0.8500	0.1122	80.9347	0.6886	0.9303	0.6180
(3,4)'	0.6931	0.3154	0.8100	0.1725	4.9167	0.8375	0.2730	0.8035

Table 1: Statistics of ECF, and of confidence interval sizes, *dZ* in mm, across all points, for confidence intervals with TPC of 0.68, for 4 different stereo pair cameras

intervals with higher TPC also leads to intervals with larger sizes, meaning worse resolution and more uncertainty in the true position. In building the intervals we have not taken into account that the parameter space for the disparity correction is bounded, i.e. the subpixel disparity correction is at most 1 in absolute values. Currently, we are addressing that.

5.2 Data Reduction Based on Confidence Intervals Statistics

Although the primary objective of the intervals is to establish the precision in the reconstruction algorithm, we have successfully used them in automatically rejecting unreliable points in the reconstruction. A simple criterion we have used to reject points was: reject reconstructed points for which the confidence interval size is larger than a statistics of the interval sizes in Z. In particular, we have used the median statistic, and also the mean of the interval sizes, *dZ*, over points corresponding to all image positions. Such a procedure removes 4-50% of the points in a reconstruction depending on the scene and the quality of that reconstruction for a given statistic. The better the reconstruction, the less points are removed (unless the statistic used is the median, in which case 50% are removed). The benefit of reducing the data set size is clear where multiple data from polynocular stereo must be fused (as in [27] where hundreds of reconstructions, based on 904 images, were used for the office scene recovery) or when the data set must be transferred over the Internet. Of course, some bad outlier points due to bad matches still survive and have high confidence. Such points cannot be removed without prior information on the scene, or without the use of some heuristics.

Here we show preliminary results regarding the rejection of data with higher degree of uncertainty, as well as reducing the data set by 50% (preserving the important features) based on the statistics of the confidence interval sizes for Z.

We remind the reader that our objective is not to report on the performance of the polynocular stereo algorithm itself, this has been reported already [26]. We report on the precision of this algorithm, and on the relative improvement in the reconstruction, or data reduction, by utilizing the confidence interval sizes to reject outlier points.

Results for the planar scene
Figure 5 shows data for the reconstruction based on Pair(1,4) and the already discussed planar scene. The statistic used for the rejection criterion was sample mean of the interval sizes, *dZ*, over all reconstructed points. Note that the criterion for rejection of outliers cannot be based on the statistic of the reconstructed points corresponding to the plane since we dó not want to assume, in general, that we have any prior information about the scene. 7% of all the points were rejected.

Results for the human face scene
Another set of experiments was done by reconstructing points for a human face. The person was seated in the view of the cameras, and random texture was projected to aid the matching (Figure 2). The goal was to reject points which do not belong to the face (with the random texture). Figure 6, top and bottom, shows two views of the reconstructed point set (top is looking straight down the Z axis). Figure 6, left column, shows two views of all the reconstructed points. Figure 6, right column, shows two views of the reconstructed points after rejecting points based on the median statistic of the confidence interval sizes (the median is taken over all points). Compared to left column, the data on the right are 50% less, still the important facial points are retained, most of the fuzz and artifacts around the face are rejected. Not all artifacts around the face are removed. For example, due to repitative texture (curtain folds) there are highly inaccurate matches (far from the true positions) with high precision (very stable). As already stated, these must be addressed during the matching procedure. The

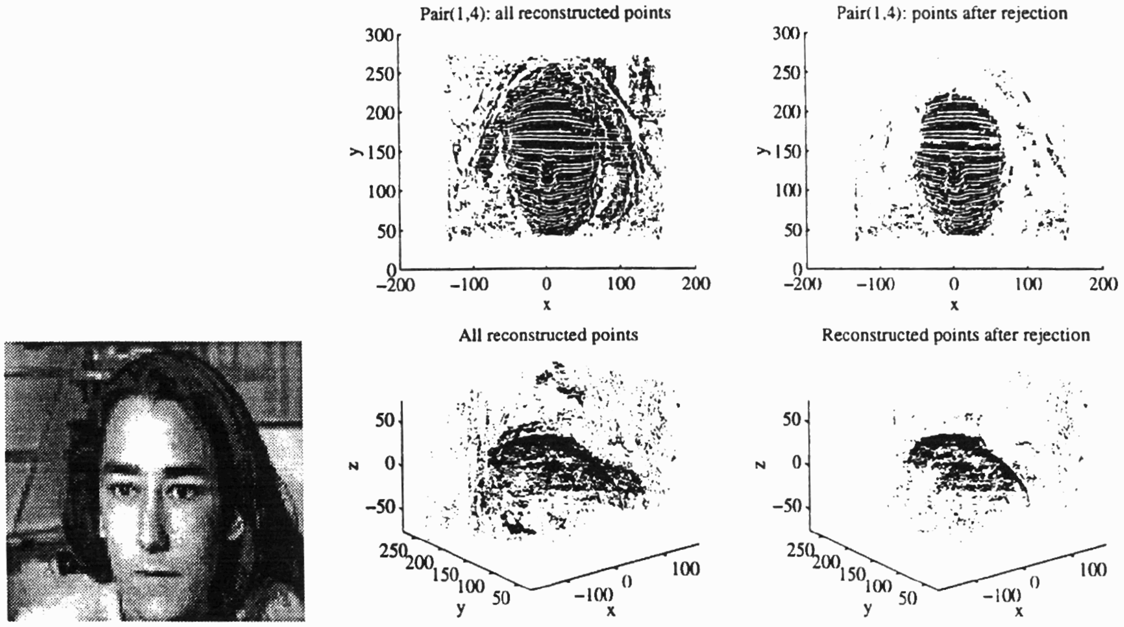

Figure 6: The original scene without the random texture, and the reconstructed points.

confidence intervals do not provide the explanation in order to remove them. For a general scene (not a special, singular case as with the plane), the median is a more stable statistic as a measure of the attributes of the underlying distributions.

Why Does It Work? The size of the confidence intervals is related to the scene attributes. The intervals are large when: (i) the signal-to-noise ratio of the image is locally low, as in the textureless areas (see the background in Figure 2 and 6), in the areas overlaid by highlights where the texture contrast is low, and in dark or shadowed areas (as in the hair in Figures 2 and 6); and (ii) when the pixel values in the left and right image windows used for assessing the match do not correspond well with each other (as in the case of highlights, in the hair and eyebrows which does not image as regular surface texture, and near occlusions where the matches are notoriously bad in all area-based stereo matching algorithms [29]).

6 Conclusions

We characterized the precision of a reconstruction from polynocular stereo in terms of confidence intervals for the components of the 3-D reconstructed points. These intervals were based on confidence intervals in subpixel disparity. Assumptions under which the derivations were made: (i) uncorrelated, zero-mean, locally constant (within window) variance of digital image noise; (ii) normally distributed subpixel disparity noise; and (iii) local planarity (within window) of the observed scene.

We will use the intervals for outlier rejection, in data reduction, and data fusion in the point data set resulting from a polynocular stereo. We showed results from some preliminary experiments related to the data rejection and data reduction issues. An attractive feature of the approach is that it does not rely on manually set thresholds.

We have evaluated the performance of the intervals in terms of ECF vs their TPC, and have showed that the TPC is actually a lower bound on the EFC. The confidence intervals in disparity are the keystone around which the 3-D point intervals were constructed. In order to improve the performance of the disparity intervals, to construct intervals with TPC closer to the ECF, we are working on the construction of fixed-size confidence intervals with guaranteed probability of capture based on minimax-decision theory [12][9]. In order to apply the the minimax-decision theory, we must specify the sampling distribution for the random noise per pixel in disparity. This is a very difficult modeling process for general scenes. The distributions can be derived theoretically (propagating the image noise distributions), or empirically. The advantage of deriving the distributions theoretically would be that the model would be independent on the scene (the Poisson component of the noise, [11], would capture the dependence). The difficulty is the recovery of the image noise parameters since the noise in digital images is

[9] The intervals we derived correspond to the maximum likelihood estimator which does not take into account the restricted parameter space.

cumulative of many factors [15]. One of the reasons we are using induced texture which has constant parameters over the image is to simplifies the task of recovering the distribution empirically. Thus we are decreasing the number of degrees of freedom that must be exhausted in order to create the empirical model. Initially, we simplify the modeling process by assuming normal distribution for the disparity noise. We will use the recovered disparity variance as a parameter for normal model for this sampling distributions. Also, we are exploring *uncertainty classes of distributions* suitable for modeling the uncertainty in disparity. This will allow us to derive robust fixed size confidence intervals with minimal maximal probability of capture [13]. The overhead in computing the fixed-sized confidence intervals is only a one-time off line computation. Details will be given in a forthcoming paper.

We do not assume normality in the distributions of X, Y, Z. These distributions are obtained from the distribution of D propagating it through the 3-D reconstruction, and are, definitely, not normal. Also X, Y, Z are not independent. The good news is that, *conditioned* on the disparity D (in the statistical meaning of the term) X, Y, Z are independent. We know the joint distribution of X, Y, Z. Jointly, we think of a 3-D confidence set, not an interval. When should we use confidence intervals, and when confidence sets? The answer depends on what are we going to use them for. For the removal of unreliable points, the 1-D intervals are enough. If we want to have a confidence of a certain level (probability of capture), that a reconstructed point is in a certain 3-D set (box or circle), then we need the confidence sets based on the joint (X, Y, Z) distribution, not intervals for the individual components. We can use the 1-D X, Y, Z distributions (which are obtained by propagating the distribution of D, independently, in each component through the reconstruction process) with conditioning on D to derive 3-D confidence sets for (X, Y, Z) from the product

$$\Pr[(X,Y,Z) \in C_X \times C_Y \times C_Z] \;=\; \Pr[X \in C_X \mid D] \Pr[Y \in C_Y \mid D] \Pr[Z \in C_Z \mid D] \Pr[D].$$

We will use the confidence intervals for data reduction, and the confidence sets for data fusion.

We have shown results here for a particular stereo algorithm, but the theory is general and could be used in many applications.

References

[1] N. Ayache and C. Hansen, "Rectification of Images for Binocular and Trinocular Stereovision", *Proc. of 9th International Conference on Pattern Recognition*, 1, pp 11-16, Italy, 1988.

[2] P. Belhumeur, "A Bayesian Approach to Binocular Stereopsis", *Intl. J. of Computer Vision*, 19(3), pp 237-260, 1996.

[3] M. Black and P. Anandan, "The Robust Estimation of Multiple Motions: Parametric and Piecewise Smooth Flow Fields", *Computer Vision and Image Understanding*, 63, pp 75-104, 1996.

[4] U. Dohond and J. Aggrawal, "Structure from Stereo: a Review", *IEEE Transactions on Systems, Man, and Cybernetics*, 19(6), pp 1489-1510, 1989.

[5] P. Bickel and K. Doksum, *Mathematical Statistics*, Holden-Day,Inc.,Oakland.CA, 1977.

[6] F. Devernay, "Computing Differential Properties of 3-D Shapes from Stereoscopic Images without 3-D Models", INRIA, RR-2304, Sophia Antipolis, 1994.

[7] W. Förstner, "10 Pros and cons against performance characterization of vision algorithms", *Workshop on performance characterization of vision algorithms*, Robin College, Cambridge, 1996.

[8] R. Haralick and L. Shapiro, *Computer and Robot Vision*, 2, Addison-Wesley, 1993.

[9] R. Haralick, "Performance characterization in computer vision", *Performance versus methodology in computer vision*, Haralick and Meer, editors, University of Washington, Seattle, 1994.

[10] R. Haralick, "Propagating Covariance in Computer Vision", *Workshop on performance characterization of vision algorithms*, Robin College, Cambridge, 1996.

[11] E. Healey and R. Kondepudy, "Radiometric CCD Camera Calibration and Noise Estimation, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(3), pp. 267-276, 1994.

[12] G. Kamberova and M. Mintz, "Minimax Rules Under Zero-one Loss for a Restricted Location Parameter", *Journal of Statistical Planning and Inference*, (accepted), 1997.

[13] G. Kamberova, "Robust Location Estimation for MLR and Non-MLR Distributions", *PhD Dissertation*, University of Pennsylvania, 1992.

[14] R. Mandelbaum, G. Kamberova and M. Mintz, "Stereo Depth Estimation: a Confidence Interval Approach", accepted *Intl. Conf. Computer Vision*, 1998.

[15] G. Kamberova, "The Effect of Radiometric Correction on Multicamera Algorithms", Technical Report.

[16] M. Okutomi and T. Kanade, "A Multiple-baseline Stereo", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(4), pp. 353-363, 1993.

[17] "A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiments", *Intl. J. of Computer Vision*, 7(2), pp 143-162, 1992.

[18] L.Mattines, R. Szeliski, T. Kanade, "Kalman Filter-based Algorithms for Estimating Depth from Image Sequences", *Intl. J. of Computer Vision*, 3, pp 209-236, 1989.

[19] R. Marik, J. Kittler and M. Petrou, "Error Sensitivity of Vision Algorithms Based on Direct Error-Propagation", *Workshop on performance characterization of vision algorithms*, Robin College, Cambridge, 1996.

[20] S. Maybank and O. Faugeras, "A Theory of Self-Calibration of a Moving Camera", *Intl. J. of Computer Vision*, 8(2), pp 123-151, 1992.

[21] P. Meer, D. Mintz, A. Rozenfeld, "Robust Regression Methods for Computer Vision: a Review", *Intl. J. of Computer Vision*, 6(1), pp. 59-70, 1991.

[22] H. Moravec, "Robot Rover Visual Navigation", *Computer Science:Artificial Intelligence*, pp. 13-15, 105-108, UMI Research Press 1980/1981.

[23] H.-H. Nagel, "Optical Flow Estimation and the Interaction Between Measuremnt Errors at Adjacent Pixels", *Intl. J. of Computer Vision*, 15, pp 271-288, 1995.

[24] Photometrics Homepage,
Photometrics High performance CCD Imaging, `http://www.photomet.com`, 1996.

[25] B. Ripley, *Statistical Inference for Spatial Processes*, Cambridge University Press, 1988.

[26] R. Šára, "Reconstruction of 3-D Geometry and Topology from Polynocular Stereo",
`http://www.cis.upenn.edu/ radim/Stereo/stereo.html`,
GRASP Laboratory, University of Pennsylvania, Philadelphia.

[27] R. Šára, "3-D Reconstruction of an Office",
`http://www.cis.upenn.edu/ radim/PennOffice/home.html`,
GRASP Laboratory, University of Pennsylvania.

[28] R. Szeliski, *Baysian Modeling of Uncertainty in Low-level Vision*, Kluwer Academic Press, 1989.

[29] R. Šára and R. Bajcsy, On Occluding Contour Artifacts in Stereo Vision", *Proc. Int. Conf. Computer Vision and Pattern Recognition*, IEEE Computer Society, Puerto Rico, 1997.