Modeling Errors in Small Baseline Stereo for SLAM

Damith C. Herath, K. R. S. Kodagoda and Gamini Dissanayake
ARC Centre of Excellence in Autonomous Systems (CAS)
Faculty of Engineering
University of Technology, Sydney
Broadway, NSW, Australia
Email: {d.herath, s.kodagoda, g.dissanayake}@cas.edu.au

Abstract – **In the past few years, there has been significant advancement in localization and mapping using stereo cameras. Despite the recent successes, reliably generating an accurate geometric map of a large indoor area using stereo vision still poses significant challenges due to the accuracy and reliability of depth information especially with small baselines. Most stereo vision based applications presented to date have used medium to large baseline stereo cameras with Gaussian error models. Here we make an attempt to analyze the significance of errors in small baseline (usually <0.1m) stereo cameras and the validity of the Gaussian assumption used in the implementation of Kalman Filter based SLAM algorithms. Sensor errors are analyzed through experimentations carried out in the form of a robotic mapping. Then we show that SLAM solutions based on the Extended Kalman Filter (EKF) could become inconsistent due to the nature of the observation models used.**

Index Terms – **SLAM, Stereo vision, sensor modeling**

I. Introduction

There is a rich collection of literature addressing various aspects of stereo vision. Multitude of them deals with error analysis and modeling of errors. For instance Matthies and colleagues[1, 2] have studied stereo errors extensively from a robotic navigation perspective. In [1], it is shown that an ellipsoidal error model for stereo measurements provide considerably good results for robotic navigation. In this exposition our main intension is to shed light on stereo errors empirically from a SLAM perspective in indoor environments where small baseline (<0.1 m) cameras are used. The explicit mentioning of the small baseline stems from the fact that such cameras tend to violate linearization assumptions made in standard EKF implementations. This work could be thought of as a precursor to a set of algorithms being developed currently for improving consistency and performance for small baseline stereo based SLAM.

In section II we present the stereo sensor module as implemented in an Extended Kalman Filter (EKF) based SLAM algorithm. Section III presents an analysis of stereo errors for a commercially available stereo camera through a set of data collected in a mapping experiment. Results from this section provide the basis for the sensor model to be used in the next section. In Section IV we discuss instances in which a traditional SLAM implementation based on EKF becomes inconsistent and show linearization as the main cause of error through an empirical study. Section V concludes the paper with pointers to our current and future work.

II. Stereo Vision System

Our approach to stereo vision is from a robotic navigation perspective. We consider a robot moving on the X-Y plane while observing features in 3D space for its localization with respect to a global reference coordinate system while simultaneously mapping these observed features. Thus in this scenario a stereo vision sensor is a combination of stereo hardware and several other algorithms including a stereo correlation algorithm and a feature initialization and tracking module as depicted in Fig. 1. Live images are streamed through the stereo hardware through to an image pre-processing module where the raw images are rectified using the calibration data. At each time step these rectified image pairs are passed on to the stereo algorithm while the rectified reference image (left image in our case) is passed to the feature tracking/initialization algorithm. The feature initialization and tracking algorithm provides the stereo algorithm with a set of feature coordinates in the image plane (u,v) that are selected or tracked by the algorithm. Then using stereo information, disparity (d) and locations of the features in camera centric Cartesian coordinate frame (x_c,y_c,z_c) are calculated. Each component is discussed in detail below.

A. Stereo Vision and Algorithms

The stereo camera used is a *Videre Design* camera model STH-MDCS/C which has an approximate baseline of 0.09 *m* and the lenses used have an effective field of view of 84.9° x 68.9°. Automated calibration routines are provided with the hardware for camera intrinsic and extrinsic parameter estimations. A software API is provided with the stereo head for generating dense stereo range images called the 'Small Vision System (SVS)' [3]. The API provides access to a fast

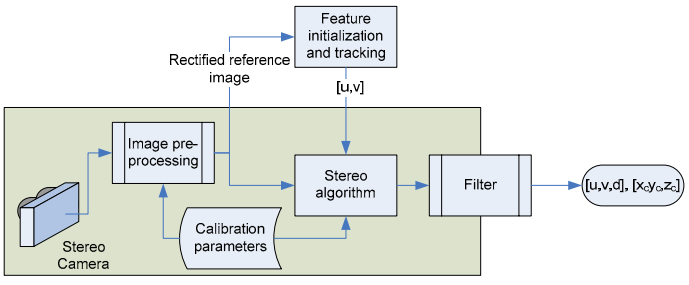


Fig. 1. The stereo vision system.

area correlation method from which a disparity image can be derived. The shaded area in Fig. 1 represents the components pertaining to the stereo system. Depending on the image composition it is possible to generate mismatches occasionally. SVS algorithm provides texture confidence threshold and left-right check as post filtering process for removing such high frequency noise components from the disparity image. However, even after implementation of afore mentioned filtering mechanisms occasional mismatches do exist (see Fig. 2). These are generally without support from surrounding regions and can be detected using simple first order statistics [4].

B. Feature Initialization and Tracking

Algorithm described in a series of papers by Kanade, Lucas and Tomasi [5-7] (collectively called the KLT algorithm) provides with a simple and efficient tracker and an interest operator pair which can be implemented in real-time. Here we use an implementation of this algorithm for selecting features and tracking them across images, which essentially solve the data association problem in SLAM.

III. Stereo Sensor Modeling

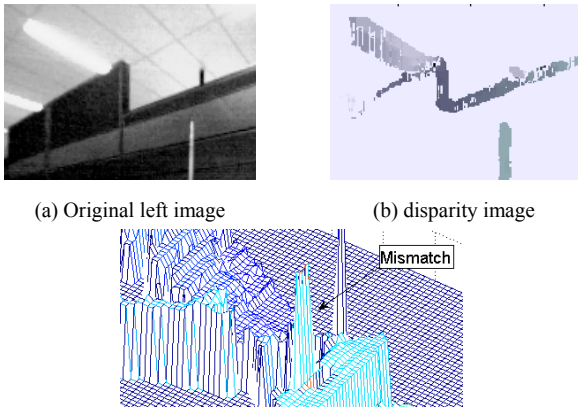
A robotic mapping experiment was carried out in order to understand the behavior of sensor noise which we discuss next along with an interpretation of the analyzed data.

A. Mapping Experiment

A pioneer robot equipped with a SICK laser aligned with the stereo camera was moved on a controlled path while observing artificial features laid on a large vertical planar surface. Artificial features were laid out on the planar surface so as to cover the whole field of view of the cameras. (Fig. 3) The SICK laser was used to maintain parallel alignment between the camera and the surface and to measure the nominal distance between the robot and the surface. Lateral and vertical distance to a particular feature in the camera coordinates were hand measured. Robot was moved in 0.05m increments from a distance of 6m to 1m capturing 30 image frames in each step.

B. Errors in u and v

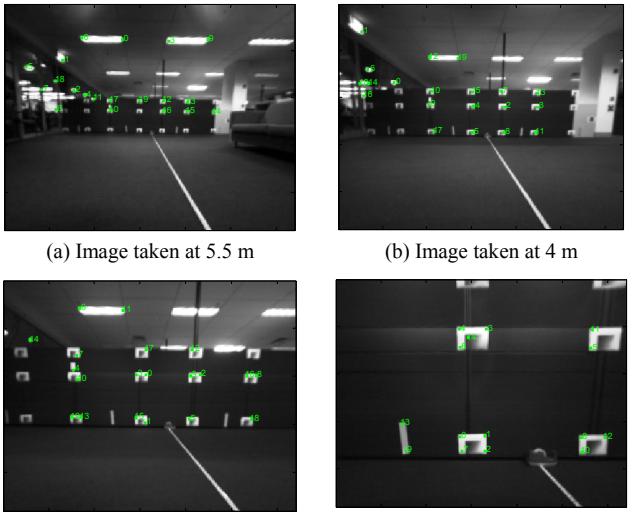
For this analysis only a single image is considered at each depth. These images are then assembled from a depth of 1m to 6m. 16 features covering the entire image plane are then initialised in the image corresponding to 1m depth and are then consecutively tracked through to image at 6m depth. This while tracking a set of features at fixed locations in space will map to varying *u, v* in left camera image coordinates. This captures the overall behaviour of *u,v* in the entire image plane. Fig. 4 (a) and (b) show the histograms of both *u* and *v* parameters with subtracted expected values. The expected values of *u* and *v* were estimated based on the measured three dimensional locations with respect to the camera. Qualitatively these distributions resemble Gaussians. It is to be noted that the distributions shown in Fig. 4 (a) and (b) has captured not only the errors in the camera, but also the errors in KLT tracking as well.



(a) Original left image (b) disparity image

(c) Close up view of the depth profile

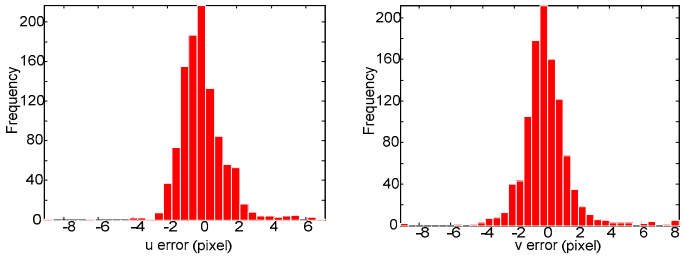
Fig. 2 Mismatches in stereo



(a) Image taken at 5.5 m (b) Image taken at 4 m

(c) Image taken at 2.5 m (d) Image taken at 1 m

Fig. 3 Rectified images of the planar surface covered with artificial features. Images are overlaid with KLT features.



(a) u (b) v

Fig. 4 Error distribution

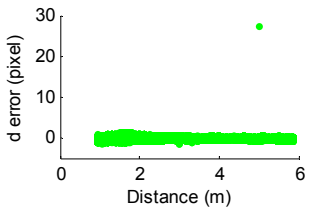
C. Errors in Disparity, d

This analysis differs from general stereo error analysis, where only a static camera is used. It is our intension to provide an encapsulating overview of how the error statistics vary as the camera moves. Thus the mapping data presents a unique perspective on the variance in disparity as observations

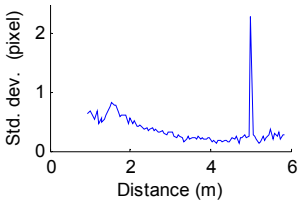
are made at varying distances, in this case an approximate range between 6m and 1m. This depth range translates to an effective disparity range approximately between 2 and 15 pixels with a focal length of 150 pixels and baseline of 0.09m.

Data for this analysis was extracted in the following manner. At each stop point at least 9 features were selected covering the whole planar surface visible in the initial image. These were then tracked along the 29 images captured at the same depth. Essentially this gives the disparity profile for the corresponding depth. Fig. 5(a) shows the variation of this disparity profile (zero mean) as the camera moves away from the planar surface from 1m to 6m. Fig. 5(b) shows the variation in estimated standard deviation corresponding to 5(a). Fig 5(c) shows the collective variation of disparity for the entire data set.

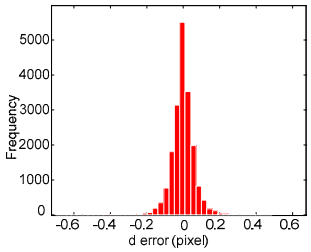
Several observations can be made in Fig. 5. Firstly, data still contains few visible outliers that could not be eliminated by various filtering operations. Secondly, a rather intuitive observation is the correlation in the variance of the disparity distribution with the distance to the observation. As would be expected, variance is smaller for features seen from afar and it increases gradually with nearby features. For far away features the disparity is small and also the discriminatory information contained within the correlation area is higher compared to a closer observation. This gives a higher confidence to disparity values estimated for features afar to ones closer. This leads to a correlated standard deviation with the observation distance, as opposed to the conventional approach of assuming a constant disparity standard deviation irrespective of depth. The observation standard deviation shown in Fig. 5 (b) could be approximated by a piece wise linear fit.



(a) Zero mean error distribution



(b) Zero mean std deviation: The peak is due to a stereo mismatch that was not detected by any of the filtering methods employed.



(c) Disparity error distribution

Fig. 5 Disparity error.

IV. Stereo Vision Based SLAM

In the Simultaneous Localization and Mapping (SLAM) problem, a vehicle with known kinematics model is utilized to navigate in an environment containing "features", starting at an unknown location building a map while estimating the robot pose with respect to that map. The vehicle is equipped with a sensor (stereo vision system), which can make relative observations between the feature and the vehicle. Although the vehicle evolution model and feature observation models are nonlinear, in most applications, a linearization based on the Taylor series expansion applied in the extended Kalman filter framework is used to solve the SLAM problem [8,9].

The SLAM frame work based on Kalman filtering is well established [8] and hence a general discussion on SLAM formulation is beyond the scope of this paper (refer [8, 10] for more information about SLAM). However, a concise formulation of the SLAM problem is presented here for completeness. The motion of the vehicle can be modeled as,

$$\mathbf{x}_v(k+1) = \mathbf{f}_v\left(\mathbf{x}_v(k), \mathbf{u}_v(k+1), \mathbf{v}_v(k+1)\right) \quad (1)$$

where, $\mathbf{f}_v(.)$ process model,

$\mathbf{u}_v(k)$ vector of control inputs

$\mathbf{v}_v(k)$ vector of uncorrelated process noise errors with zero mean Gaussian with covariance $\mathbf{Q}_v$.

The features are assumed to be stationary and can be modeled as,

$$\mathbf{x}_f(k+1) = \mathbf{x}_f(k) \quad (2)$$

The vehicle is fitted with a sensor, which can observe the features with respect to the vehicle.

$$\mathbf{z}_i(k) = \mathbf{h}_i\left(\mathbf{x}_v(k), \mathbf{x}_{fi}(k), \mathbf{w}_i(k)\right) \quad (3)$$

where, $\mathbf{h}_i(.)$ is the observation model of the i^{th} feature, $\mathbf{w}_i(k)$ is the observation noise associated with i^{th} feature that is assumed to be zero mean Gaussian with covariance $\mathbf{R}_i$.

A. Gaussian Distributed x,y,z as Observations

The most commonly used observation model in the literature is the range, bearing $[r,\theta]^T$ [8] and elevation $[r,\theta,\varphi]^T$ [11] with three dimensional observations. Although this is the natural observation model for most traditional sensors like laser and sonar, stereo vision provides a set of different observation models which up to now has received only limited attention. The natural tendency in stereo vision is to use $\mathbf{z}_c = [x_c, y_c, z_c]^T$ as the observation model [4, 12]. First we assume the observations to be with zero mean uncorrelated Gaussian noise. The observation prediction becomes nonlinear due to the transformations from world coordinates to the

camera coordinates. However, it can successfully be handled through EKF framework providing consistent SLAM results as shown in the following simulation.

The robot's simulated path is shown in Fig. 6 along with noise corrupted odometric path and a set of randomly generated 3D features scattered in the environment. Observations to these features are corrupted from a zero mean Gaussian distribution with ($\sigma_x = \sigma_y = \sigma_z = 0.05$ m). When a standard EKF is used to solve this SLAM problem consistent results are achieved as shown in Fig. 7. The consistency is verified by the robot's pose estimates being well bounded by the 2σ error bounds.

B. Projected x,y,z as Observations

$[u, v, d]^T$ are the primary observations in a stereo vision system and $[x_c, y_c, z_c]^T$ are derived observations. Those are derived from a projective mapping of feature locations from the image plane to the camera centric coordinate frame. The underlying relationship between a feature location in the image coordinates and the camera coordinates are given by, (note the coordinate frame used relates to the one commonly used in SLAM [4])

$$x_c = \frac{Bf}{d}; y_c = \frac{-Bu}{d}; z_c = \frac{-Bv}{d} \qquad (4)$$

In the previous section it was shown that *u, v* and *d* can safely be assumed as independent Gaussian distributions. Thus assuming $\sigma_u, \sigma_v, \sigma_d$ represents the pixel uncertainties in image *u,v* location and the disparity *d* respectively, the covariance matrix in the image plane is given by,

$$\mathbf{\Sigma}_I = \begin{bmatrix} \sigma_d^2 & 0 & 0 \\ 0 & \sigma_u^2 & 0 \\ 0 & 0 & \sigma_v^2 \end{bmatrix} \qquad (5)$$

This can now be transformed in to the camera coordinate frame using first order linearization method as,

$$\mathbf{R} = \mathbf{J}\,\mathbf{\Sigma}_I\,\mathbf{J}^T = \frac{B^2}{d^2}\begin{bmatrix} \frac{f^2\sigma_d^2}{d^2} & \frac{-uf\sigma_d^2}{d^2} & \frac{-vf\sigma_d^2}{d^2} \\ \frac{-uf\sigma_d^2}{d^2} & \sigma_u^2 + \frac{u^2\sigma_d^2}{d^2} & \frac{uv\sigma_d^2}{d^2} \\ \frac{-vf\sigma_d^2}{d^2} & \frac{uv\sigma_d^2}{d^2} & \sigma_v^2 + \frac{v^2\sigma_d^2}{d^2} \end{bmatrix} \qquad (6)$$

where **J** is the Jacobean of (x_c, y_c, z_c) with respect to (d, u, v). In order to analyze the effect on SLAM, a simulation study based on the previously simulated environment (Fig. 6) was carried out. Gaussian noise was added to $[u, v, d]^T$ and $[x_c, y_c, z_c]^T$ observations were generated using (4). The baseline, *B,* and focal length, *f,* were assumed to be 0.09m and 150 pixels respectively. This data was then passed through the previously used SLAM algorithm, with modifications to accommodate the error model in (6).

Pose estimate results are shown in Fig. 8 indicating the inconsistent filter behavior. The same simulation was repeated with *B* = 0.5m to resemble a wide baseline stereo camera. The error plots are shown in Fig. 9. It shows consistent SLAM results when compared with that in Fig. 8. This leads us to the logical conclusion that the linearization mechanism used in the EKF implementation is inadequate to handle the strong nonlinearities present in the small baseline stereo observations. To further illustrate this phenomenon, consider the Gaussian random variable $[d, u]^T$ (only two components used for clarity) representing the disparity and horizontal image coordinate for a given feature at $x_c = 10m$ and $y_c = 1m$. With *B* =0.09m and *f* = 150 pixels, this translates to mean disparity, *d* of 1.32 pixels and mean *u* of 15 pixels. A Monte Carlo simulation was carried out using (4) to transform Gaussian distributed $[d, u]^T$ into $[x_c, y_c]^T$. Fig. 10 (a) and (b) show the resulting distributions with 0.09m and 0.5m as baselines respectively. This clearly indicates the non Gaussian nature of the transformed observations when a small baseline camera is used (Fig. 10 (a)). The smaller the baseline is the shorter the range is at which the nonlinear effect manifest. Jung [12] suggests that with the ratio

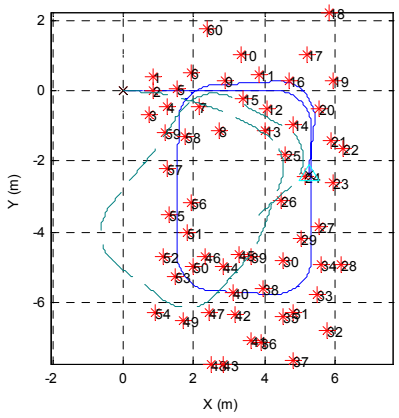


Fig. 6 Simulated environment: *solid line* – true path, *dashed line* – odometry path, * - features

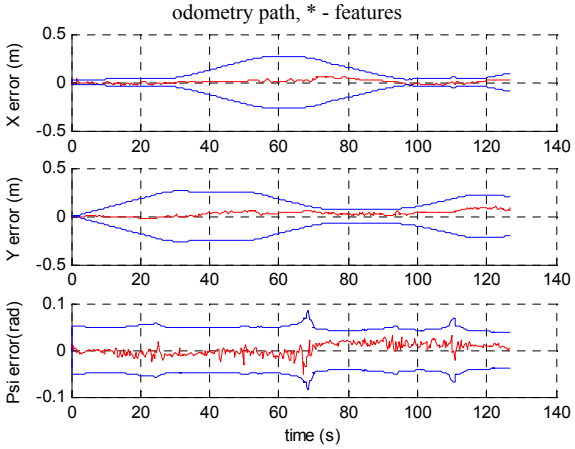


Fig. 7 Error plot: *x,y,z* as Gaussian observations

(*baseline/depth*) > (1/30), a more conducive zero mean approximation could be obtained. With a 0.09m baseline, this ratio translates to observing ranges less than 2.6m. It is far from acceptable given that for a successful implementation of SLAM features need to be observed for prolong periods of time, especially from a distance to improve the heading errors.

C. u,v,d as Observations

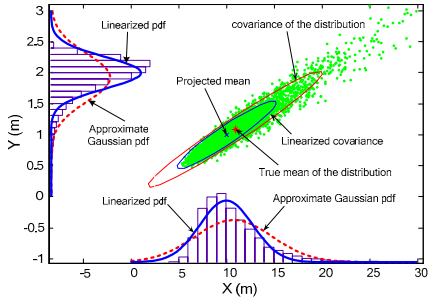
$[u,v,d]^T$ are the primary observations in a stereo vision system. Previously we showed that those observations can be considered as having independent Gaussian distributions with zero means which is more conducive in an EKF implementation. Using the same simulated data from the previous section, an EKF was implemented with a modified observation model consisting of $[u,v,d]^T$. Results from this implementation are presented in Fig. 11. Compared to the $[x,y,z]^T$ model this yield improved results, however still the filter remains inconsistent as can be seen from the pose error for most parts being over the 2σ bounds.

This inconsistency could be explained by noting that,

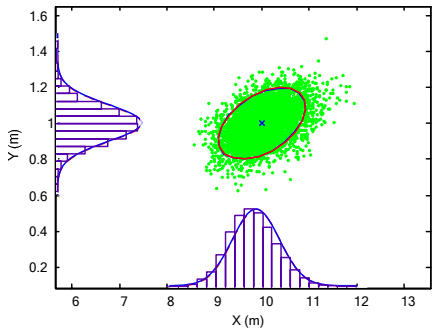
1.) Initialisation of features in the map still requires a linearization.

2.) During the prediction stage, it still requires linearization in the form $\mathbf{HP}^{-}\mathbf{H}^{T}$ in order to propagate covariance from world coordinates to the image coordinates. $\mathbf{H}$ is the Jacobean of the relevant observation function and $\mathbf{P}^{-}$ is the predicted covariance of the corresponding map elements.

So far we have looked at two possible alternative observation models and their effects on the SLAM implementation from the filter consistency perspective. Through simulated data it was shown that for narrow baseline stereo vision, an EKF filter could become inconsistent owing to the linearization approach adopted in the implementation. Fig. 12 shows the paths generated by different implementations using simulated data. Path generated by the



(a) B=0.09m



(b) B=0.5m, linearized and approximated Gaussians are overlapping

Fig. 10 Errors in projective mapping

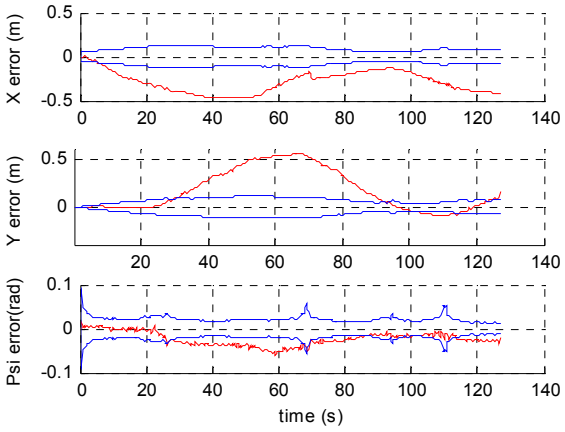


Fig. 8 Error plot: Projected *x,y,z* as observations (B = 0.09m)

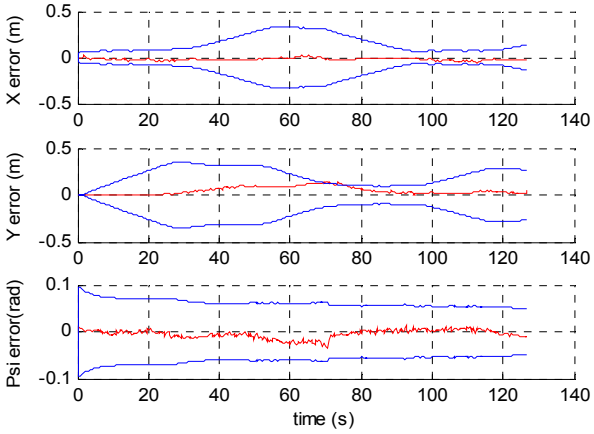


Fig. 9 Error plot: Projected *x,y,z* as observations (B=0.5m)

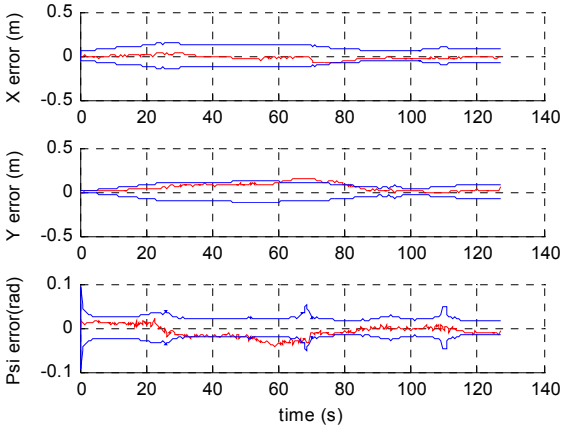


Fig. 11 Error plot: *u,v,d* as Gaussian observation

filter using the $[u,v,d]^T$ as Gaussian observations agrees more with the true path whereas the projected form of the $[x,y,z]^T$ observations show clear divergence from the true path. As expected, when the $[x,y,z]^T$ observations are truly Gaussian, the path estimates agrees with the truth comprehensively.

Similar trends could be seen with experimental data, which were captured using a Pioneer mounted stereo camera as shown in Fig. 13. Apart from the linearization issues discussed previously, several other factors including spurious data and slowly drifting fake features contribute to the degradation in performance with real data. We are currently working on a more elegant solution for the high nonlinearities presence in the observation equations as in stereo vision system, in order to achieve consistent filter performance in substantially larger loops.

V. Conclusion

We have investigated the use of small baseline stereo in SLAM. Only modest attention has been given to this area of research which still poses substantial hindrance to a successful SLAM implementation. This work attempts to shed light on understanding the sensor behavior and sensor modeling. We have shown that with the use of small baseline stereo cameras the non linearity manifest within very short ranges leading to inconsistencies in SLAM. A simple linearization as in EKF can not handle such nonlinearities requiring a more elegant solution. We are now focusing our attention towards incorporating non Gaussian observations in SLAM context as a solution.

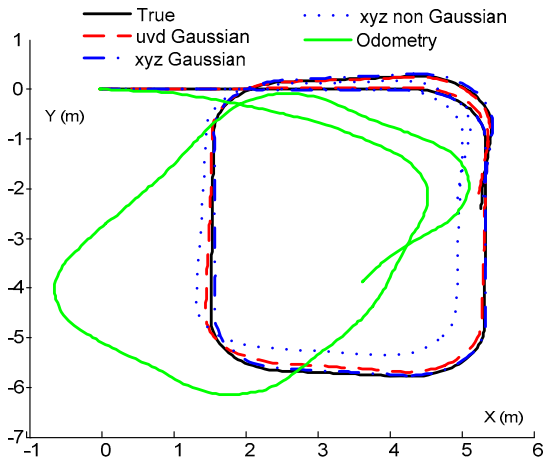


Fig. 12 Comparison of paths generated using simulated data.

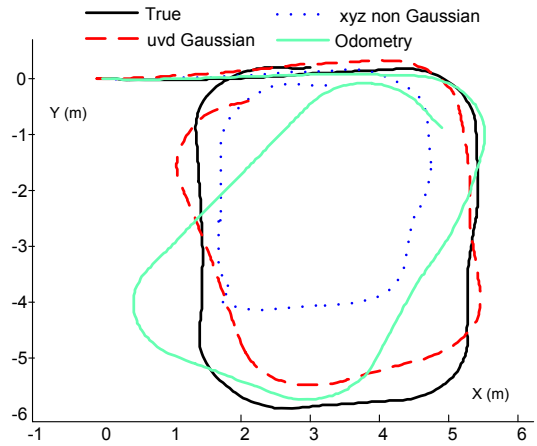


Fig. 13. Comparison of paths generated using experimental data

Acknowledgment

This work is supported by the ARC Centre of Excellence program, funded by the Australian Research Council (ARC) and the New South Wales State Government.

References

[1] L. Matthies and S. S., "Error modeling in stereo navigation," *IEEE Journal of Robotics and Automation*, vol. 3, pp. 239 - 248 1987.

[2] Y. Xiong and L. Matthies, "Error analysis of a real-time stereo system," presented at IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997.

[3] K. Konolige, "Small Vision Systems: Hardware and Implementation," presented at Eighth International Symposium on Robotics Research, 1997

[4] D. C. Herath, S. Kodagoda, and G. Dissanayake, "Simultaneous Localisation and Mapping: A Stereo Vision Based Approach," presented at IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006), Beijing, China, 2006.

[5] T. Kanade and M. Okutomi, "A stereo matching algorithm with an adaptive window: theory and experiment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 920 - 932 1994.

[6] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," presented at International Joint Conference on Artificial Intelligence (IJCAI '81), Vancouver, BC, Canada,, 1981.

[7] C. Tomasi and T. Kanade, "Detection and tracking of point features," Carnegie Mellon University CMU-CS-91-132, April 1991 1991.

[8] M. W. M. G. Dissanayake, P. Newman, S. Clark, H. F. Durrant-Whyte, and M. Csorba, "A Solution to the Simultaneous Localization and Map Building (SLAM) Problem," *IEEE Transactions On Robotics And Automation*, vol. 17, pp. 229-241, 2001.

[9] R. Smith, M. Self, and P. Cheeseman, "A Stochastic Map For Uncertain Spatial Relationships," presented at Fourth International Symposium on Robotics Research, 1987.

[10] S. B. Williams, "Efficient Solutions to Autonomous Mapping and Navigation Problems," in *Department of Mechanical and Mechatronic Engineering*, vol. Doctor of Philosophy. Sydney: The University of Sydney, 2001.

[11] S. Takezawa, D. C. Herath, and G. Dissanayake, "SLAM in Indoor Environments with StereoVision," presented at IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004), Sendai International Center, Sendai, Japan, 2004.

[12] I. K. Jung, "Simultaneous localization and mapping in 3D environments with stereovision," in *LAAS*, vol. PhD. Toulouse: Institut National Polytechnique, 2004, pp. 118.