Proc. 2001 Australian Conference on Robotics and Automation
Sydney, 14-15 November 2001

Application of Vision in Simultaneous Localization & Mapping

Trevor Fitzgibbons and Eduardo Nebot
Australian Centre for Field Robotics
Rose Street Building, J04
The University of Sydney, 2006

Abstract

This paper presents the strategies and algorithms for implementing video data into the Simultaneous Localization and Mapping (SLAM) problem, with the emphasis for outdoor applications. The video information is obtained from standard cameras. Natural feature detection algorithm are implemented to obtain relative bearing information. This information is then used in a Simultaneous Localization and map building framework to bound the dead-reckoning errors. Experimental results in an outdoor environment are presented.

1 Introduction

Localization through *a priori* map has been a solved problem for sometime, as has mapping from observations at known positions [Elfes, 1989], [Stentz *et al.*, 1999], [Durrant-Whyte, 1996]. More difficult is the combination of localization and mapping, which is known as the Simultaneous Localization & Mapping problem, commonly referred to as SLAM. This infers no *a priori* information is known, and all localization is done as the map is built [Leonard and Durrant-Whyte, 1991].

The extended Kalman Filter (EKF) can be used to solve a SLAM problem [Leonard and Durrant-Whyte, 1991], [Williams *et al.*, 2000], [Guivant *et al.*, 2000], as long as models can be provided for the vehicle's motion and sensors. Increased complexity comes from taking this application into outdoor environments [Guivant and Nebot, 2001], due to the difficulty of extracting and mapping natural landmarks.

The use of vision has been applied to localization and mapping. Extracting structure and motion from video [Dellaert *et al.*, 2000a] is a currently pursued field, which parallels the efforts of SLAM. The distinction between the two is that SLAM aims to carry its operation in a sequential manner, where 'structure and motion' is performed in batch mode.

The use of visual information for localization has been approached by [Dellaert *et al.*, 1999b], [Fox *et al.*, 1999], who used a Monte Carlo filter to localize their position, and both [Davidson and Murray, 1998], [Lacroix *et al.*, 2001], using stereo-vision to aid in applying SLAM.

One way to use video information is by extracting bearing to natural features selected as targets. As such initialization can only be performed with at least two observations of the same landmark. This raises problems on data association and landmark validation. Furthermore, since all pixels are occupied by some feature in the environment, association between a known landmark contained in the map and its appearance in the image can be difficult. These problems are addressed with techniques that enable association using information derived from classical imaging methods and statistical validation.

The paper is structured as follows: Section 2 will provide information on the modelling of cameras and images. Section 3 introduces the SLAM problem and how the extended Kalman filter is applied. Section 4 discusses the selection process employed in obtaining well-conditioned features. Section 5 examines the problems with initializing for bearing-only SLAM. Section 6 looks at data association between landmarks and video images. Section 7 has the presentation of experimental results using the algorithms presented in this paper as used in an outdoor environment. Finally Section 8 presents a conclusion and future paths of this research.

2 Fundamentals of Cameras

The properties of the camera must be first understood for modelling it as a sensor and developing data association techniques. The advantages for using a camera are that provides 3-D information on the environment and delivers a large amount of information in each return.

The data delivered is a 2-dimensional image, formed from ray casting from the object to the camera's focal point and onto a CCD array. Each pixel value is a measure of the light intensity that is returned from the environment. This is made up of the amount of illumination that is incident to the scene and the amount of light reflected from the object it's self. These two components are known as the illumination (α) & reflectance components (r), with the light intensity (p) being the product of the two [Gonzalez and Wintz, 1987]. As such the image model can be described as a function of the illumination and reflectance components;

$$p(u,v) = \alpha(u,v) r(u,v). \qquad (1)$$

The pixel values are then used to identify and associate landmarks as they are viewed.

The camera model works upon the principle that the image is a result of the projection of a point P onto the 'image' plane, typically the capturing CCD array. The perspective origin, O, acts as the origin of the reference frame, XYZ. The image plane lies parallel to the XY-plane at a distance known as the focal length, *f*, along the Z-axis. The point at which the Z-axis intersects with the image plane is known as the principle point, [*Cu, Cv*]. It should be noted that the focal length will differ in the X & Y direction depending on the resolution along these axes.

The position of the object on the image plane, p, is the projection of the pencil from P to the perspective origin, *O*.

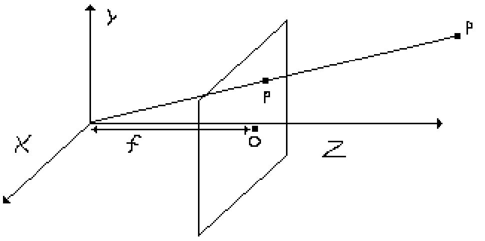

Figure 1: Camera Reference Frame.

The camera model can then be expressed as

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \begin{bmatrix} X/Z \\ Y/Z \\ 1 \end{bmatrix}$$
$$K = \begin{bmatrix} f_u & 0 & C_u \\ 0 & f_v & C_v \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

$$u = f_u \frac{X}{Z} + C_u$$
$$v = f_v \frac{Y}{Z} + C_v. \tag{3}$$

Where {X,Y,Z} are relative to camera reference frame, *{fu,fv,Cu,Cv}* are the intrinsic parameters of the camera, and {*u,v*} are the resulting coordinates of the image.

To develop a sensor model for localisation purposes , a point {X,Y,Z} needs to be converted into the global coordinates.

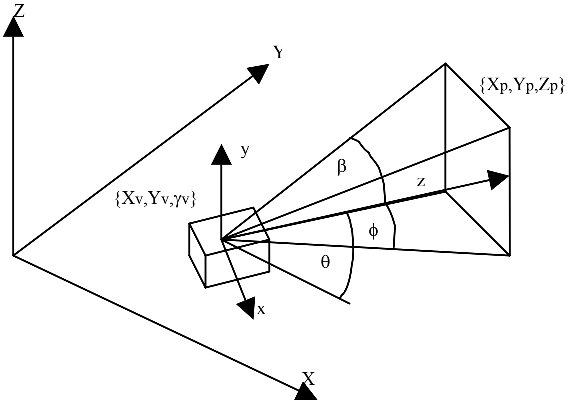

Figure 2: Camera Model fixed to Vehicle, related to the Global Reference Frame.

For this we assume that the camera is fixed forward on the vehicle, so that the camera's Z-axis points in the direction of ϕ. The vehicle will move in the XY plane (2D), where the landmarks will be 3D. The coordinates can be converted to the bearing it makes to the landmarks, such that;

$$\theta_i = \tan'\left(\frac{u_i - C_u}{f_u}\right) = \tan'\left(\frac{X_i}{Z_i}\right)$$
$$\beta_i = \tan'\left(\frac{v_i - C_v}{f_v}\right) = \tan'\left(\frac{Y_i}{Z_i}\right). \tag{4}$$

The same bearings relative to the vehicle are described as

$$\theta_i = \phi_L - \tan'\left(\frac{y_i - y_L}{x_i - x_L}\right)$$
$$\beta_i = \tan'\left(\frac{z_i}{\sqrt{(x_i - x_L)^2 + (y_i - y_L)^2}\cos\theta_i}\right). \tag{5}$$

3 Simultaneous Localization and Mapping

The SLAM algorithm [Guivant and Nebot, 2001] addresses the problem of a vehicle with known kinematic, starting at an unknown position and moving through an unknown environment populated with some type of features. The algorithm uses dead reckoning and relative observation to detect features, to estimate the position of the vehicle and to build and maintain a navigation map. With appropriate planing the vehicle will be able to build a relative map of the environment and localize itself. If the initial position is known with respect to a global reference frame or if absolute position information is received during the navigation task then the map can be registered to the global frame. If not the vehicle can still navigate in the local map performing a given task, explore and incorporate new areas to the map.

A typical kinematic model of a land vehicle can be obtained from Figure 3. The steering control α and the speed υ_c are used with the kinematic model to predict the position of the vehicle. The external sensor information is processed to extract features of the environment, in this case called $B_{i(i=1..n)}$, and to obtain relative range and bearing, $z(k) = (r, \beta)$, with respect to the vehicle pose.

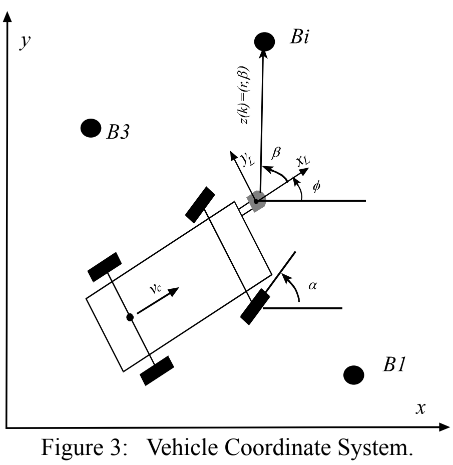

Figure 3: Vehicle Coordinate System.

Considering that the vehicle is controlled through a demanded velocity υ_c and steering angle α the process model that predicts the trajectory of the centre of the back axle is given by

$$\begin{bmatrix} \dot{x}_c \\ \dot{y}_c \\ \dot{\phi}_c \end{bmatrix} = \begin{bmatrix} v_c \cdot \cos(\phi) \\ v_c \cdot \sin(\phi) \\ \dfrac{v_c}{L} \cdot \tan(\alpha) \end{bmatrix} + \gamma , \tag{6}$$

where L is the distance between wheel axles and γ is white noise. The observation equation relating the vehicle states to the observations is given by equation (5), where z is the observation vector, (x_i, y_i) is the coordinates of the landmarks, x_L , y_L and ϕ_L are the vehicle states defined at the external sensor location and γ_h the sensor noise.

In the case where multiple observation are obtained the observation vector will have the form:

$$Z = \begin{bmatrix} z^1 \\ \vdots \\ z^m \end{bmatrix} . \tag{7}$$

Under the SLAM framework the vehicle starts at an unknown position with given uncertainty and obtains measurements of the environment relative to its location. This information is used to incrementally build and maintain a navigation map and to localize with respect to this map. The system will detect new features at the beginning of the mission and when the vehicle explores new areas. Once these features become reliable and stable they are incorporated into the map becoming part of the state vector.

The state vector is now given by

$$\begin{gathered} X = \begin{bmatrix} X_L \\ X_I \end{bmatrix} \\ X_L = (x_L, y_L, \phi_L)^T \in R^3 \\ X_I = (x_1, y_1, \ldots, x_N, y_N)^T \in R^{2N} , \end{gathered} \tag{8}$$

where $(x,y,\phi)_L$ and $(x,y)_i$ are the states of the vehicle and features incorporated into the map respectively. Since this environment is consider to be static the dynamic model that includes the new states becomes:

$$\begin{gathered} X_L(k+1) = f(X_L(k)) + \gamma \\ X_I(k+1) = X_I(k) \end{gathered} . \tag{9}$$

It is important to remarks that the landmarks are assumed to be static. Then the Jacobian matrix for the extended system is

$$\frac{\partial F}{\partial X} = \begin{bmatrix} \dfrac{\partial f}{\partial \tilde{x}_L} & \varnothing \\ \varnothing & I \end{bmatrix} = \begin{bmatrix} J_1 & \varnothing \\ \varnothing & I \end{bmatrix} . \tag{10}$$

$$J_1 \in R^{3x3}, \varnothing \in R^{3xN}, I \in R^{2Nx2N}$$

These models can then be used with a standard EKF algorithm to build and maintain a navigation map of the environment and to track the position of the vehicle.

The Prediction stage is required to obtain the predicted value of the states X and its error covariance P at time k based on the information available up to time *k-1*,

$$\begin{gathered} X(k+1,k) = F(X(k,k), u(k)) \\ P(k+1,k) = J \cdot P(k,k) \cdot J^T + Q(k). \end{gathered} \tag{11}$$

The update stage is function of the observation model and the error covariances

$$\begin{gathered} S(k+1) = H \cdot P(k+1,k) \cdot H^T(k+1) + R(k+1) \\ W(k+1) = P(k+1,k) \cdot H^T(k+1) \cdot S^{-1}(k+1) \\ \nu(k+1) = Z(k+1) - h(X(k+1,k)) \\ X(k+1,k+1) = X(k+1,k) + W(k+1) \cdot \nu(k+1) \\ P(k+1,k+1) = P(k+1,k) - W(k+1) \cdot S(k+1) \cdot W(k+1)^T , \end{gathered} \tag{12}$$

where

$$J = J(k) = \left.\frac{\partial F}{\partial X}\right|_{(X,u)=(X(k),u(k))} \quad and \quad H = H(k) = \left.\frac{\partial h}{\partial X}\right|_{X=X(k)} \tag{13}$$

are the Jacobian matrices derived from vector functions *F(X,u)* and *h(k)* with respect to the state *X(k)* . *R* and *Q* are the error covariance matrices characterizing the noise in the observations and model respectively.

4 Feature Selection

The type of feature that is to be used in the map, should be detectable and easily distinguishable within the image. Working with an outdoor environment, the landmarks available usually consist of trees and buildings, which are reasonable to use as they are stationery and not closely grouped.

The corner detection technique used in this work is described in [Tomasi and Kanade, 1991]. This involves calculating the 2x2 gradient matrix, G;

$$G = \begin{bmatrix} \Sigma g_x^2 & \Sigma g_x g_y \\ \Sigma g_x g_y & \Sigma g_y^2 \end{bmatrix} , \tag{14}$$

and then taking its eigenvalues. Large eigenvalues will indicate that the point is above the image noise, and if both are within a relative scale, then the point is well-conditioned.

For outdoor navigation, the selection of good features is hampered by non-uniform illumination. As such any comparison of pixel values or their gradients will not only be comparing the structure of the scene but also the illumination, shown in (1) & (15)

$$\begin{gathered} g_x = r\frac{\partial \alpha}{\partial x} + \alpha\frac{\partial r}{\partial x} \\ g_y = r\frac{\partial \alpha}{\partial y} + \alpha\frac{\partial r}{\partial y} \end{gathered} . \tag{15}$$

The eigenvalues derived from the G matrix cannot be compared to gain sight into the structure of the scene unless the illumination of the scene is known. This has shown that an area with little structural distinctiveness can overweight over with high structure but shaded.

To address this problem in practical applications, it is assumed that the illumination is constant in the small region surrounding any given pixel. If we let the illumination be a constant value I, for this region, such that

$$\alpha\left(u,v\right)=I \tag{16}$$

then the gradient functions are proportional to the gradient of the scene structure (17)

$$\begin{aligned} g_x &= I\frac{\partial r}{\partial x} \\ g_y &= I\frac{\partial r}{\partial y}. \end{aligned} \tag{17}$$

Direct ratio comparisons can then be done to get a relationship of the scene structure for that region. This was applied as a criteria in eliminating features that were selected due to large values of *α*, *dα/dx* & *dα/dy*.

The ratio comparison used requires the ratio of the average of a pixel's neighbours over the pixel in question. This provides an intuitive reasoning that if the pixel in question was a corner then its value would be notably higher than its neighbours.

$$\begin{aligned} C_{u,v} &= 1-\frac{average\left(q_{neighbours}\right)}{q_{u,v}} \\ &= 1-\frac{\Sigma q_{i,j}}{8q_{u,v}}. \end{aligned} \tag{18}$$

This ratio however does not carry any information on the strength of the corner, and thus should not be used to compare to other features in the image. The picture below shows the applied corner test with the removal of pixels with a coefficient, C, of 0.65 or less.

Figure 4: Corners selected in Outdoor Environment, C >= 0.65.

The features shown above are the 25 features with the highest minimum eigenvalues that satisfies the ratio criteria. This however does not scale the remaining features.

5 SLAM Initialization

Although the selection process will return a well-conditioned feature, it must be verified as being a stable landmark, since a single observation provides no information whether the feature is stationary. Furthermore, since the camera provides only bearing angles to the feature in question, initial estimations of its position will require at least two observations.

In the case where two observations are made of the same landmarks, then the position can be calculated as

$$\left[x_p,y_p,z_p\right]=f\left(x_1,y_1,\gamma_1,u_1,v_1,x_2,y_2,\gamma_2,u_2\right)$$

$$\begin{aligned} &let: \\ &\theta_i=\gamma_i-\tan'\left(\frac{u_i-C_u}{f_u}\right) \\ &\alpha_i=\gamma_i-\tan'\left(\frac{y_p-y_i}{x_p-x_i}\right) \\ &x_p=\frac{-y_1+y_2+x_1\tan\theta_1-x_2\tan\theta_2}{\tan\theta_1-\tan\theta_2} \\ &y_p=\left(x_p-x_1\right)\tan\theta_1+y_1 \\ &z_p=\left(\frac{v_1-C_v}{f_v}\right)\left(\sqrt{\left(x_p-x_1\right)^2+\left(y_p-y_1\right)^2}\cos\alpha_1\right), \end{aligned} \tag{19}$$

where the first observation is $Z_1=\left[u_1 \quad v_1\right]$ at pose, $X_1=\left[x_1 \quad y_1 \quad \gamma_1\right]$, and the second, $Z_2=\left[u_2 \quad v_2\right]$ at $X_2=\left[x_2 \quad y_2 \quad \gamma_2\right]$.

To initialize these landmarks and to verify the choice in landmark, a 3-point verification method is employed.

1. Upon each camera observation, viable features are extracted and stored. The process of searching for matches to known landmarks is run, and any feature that lay within the search windows are eliminated from the list. This is done to prevent selection of landmarks that could be mistaken in later image searches, and to prevent initializing a second instance of an existing landmark in the map.
2. The remaining observations with their poses are added to a list of previous observations. From this list, a landmark position estimate is made for each possible pairing with its associated covariance. Possible pairing of observations are those that:
   a) the bearings converge to a point within both the pose's field of view (i.e. in front of the vehicle)
   b) the poses are far enough to pair well-conditioned observations for the calculation
   c) are from a different feature set (i.e. different time instance).
   d) Their appearance is similar, as found with the application of the correlation test.

3. Then for each pair, a third observation is required to verify that it is a valid landmark. When comparing to other observations, the estimated observation is calculated from the estimated landmark position and the pose of the camera for the compared observations, along with its covariance. Taking the innovation from the actual and estimated observations, a chi-squared test can be performed, where χ^2 of 7.38 is used for a confidence of 95%.

Validated results are added to the map, and covariance used in the re-calculation of the state covariance matrix.

6 Data Association between Landmarks and Images

The data association between a mapped landmark and a detected observation is usually verified using the statistical information in the EKF. This can be performed by either:

- 2-sigma gate test
- Chi-Squared test

The approach taken was to use the Chi-squared test as the bounding function for the coordinates of a search window. A equation by which the Chi-squared value is calculated by, is

$$\chi^2 \geq v^T S^{-1} v, \tag{20}$$

where v is the innovation, $v = \left[\left(u_{obs} - u_{est} \right) \quad \left(v_{obs} - v_{est} \right) \right]^T$

Knowing that the observation has 2-degrees-of-freedom, the Chi-squared level for 95% confidence is 7.38.

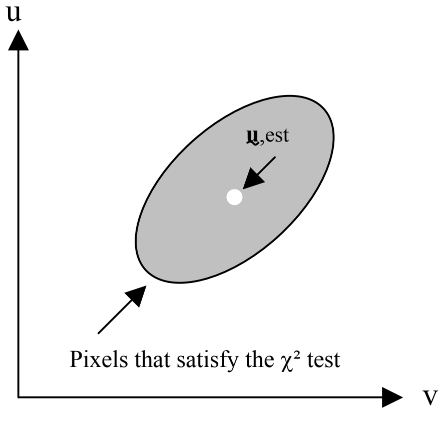

Figure 5: Innovation Covariance projected onto Image Plane.

The bounding function from the test can be expanded to represent an ellipsoid region in the image defined by

$$7.28 \geq \frac{1}{\det(S)} \left(S_{22} \Delta u^2 - \left(S_{12} + S_{21} \right) \Delta u \Delta v + S_{11} \Delta v^2 \right).$$

$$\Delta u = u_i - u_{est} \tag{21}$$

$$\Delta v = v_i - v_{est}$$

As long as the pixel coordinates lie within this ellipse, it satisfies the chi-squared test and is thus a valid candidate for a match. Using this, the ellipse becomes the bounds for a search window, with each pixel that lies within tested for a match. To determine if a given pixel matches to the landmark, a correlation test is performed on the pixel, comparing it and its neighbours to the last 3x3 image of the landmark.

If the pixel correlation is high enough to satisfy a match, the feature is taken as a new observation of the landmark and the SLAM filter is updated with this. Correlation tests are not affected by a difference of brightness, but can not handle significant changes in scale or skew that come about from different points of view. To compensate for this, the tolerance for the correlation is kept approximately 0.75-0.85 and upon each successful observation, the 3x3 image is re-sampled from the match.

7 Experimental Results

The experimental tests were done on a utility vehicle, modified to carry a sensor suite. The images were captured on an ELMO TSP481 CCD camera, through a Matrox Meteor II video capture board. The dead reckoning data was obtained from a wheel encoder, situated on the rear left wheel, and the steering through an LVDT attached the steering rack. The true path was obtained through a G24 Ashtech GPS set in differential mode with an accompanying base station nearby. Logging was done by a 400MHz Celeron computer.

Two data sets were used in the testing of this algorithm: the first on a hockey field within university grounds and the second from a farm. These were chosen for their variety of landscape and potential landmark features.

Using the criterion described in the previous section for feature selection, a variety of results were obtained. On the hockey field, these features were predominantly building vertices and the corners of windows. Figure 6 shows one such frame from this run.

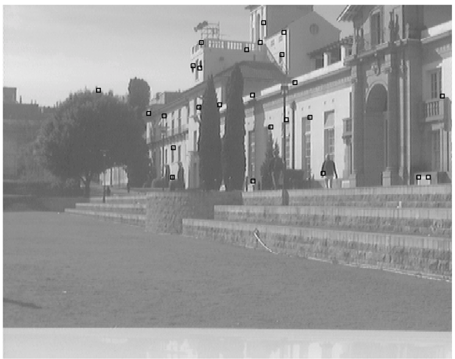

Figure 6 – Corners selected in Hockey Field, C >= 0.65.

On the farm, trees were more commonly selected, particularly their canopies and branches, as shown in figure 7.

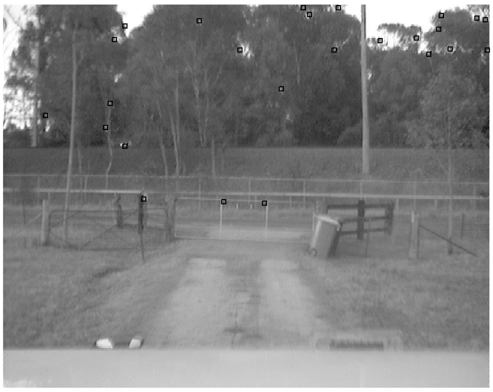

Figure 7: Corners selected in Outdoor Environment, C >= 0.65.

Without the ratio comparisons, the majority of selected features were taken along the edges of regions, which did not provide unique point features for the algorithm.

A separate test was performed to assess the individual components of landmark initialization and covariance search. The SLAM algorithm was run with GPS input to update the vehicle position and the number of updates per beacon analyzed (figure 8). Under these conditions using the data obtained from the farm, the algorithm found 167 landmarks throughout the course of the run. Of these landmarks only six were not observed after initialization. This suggests that these were spurious landmarks, being the result of mismatching of observations in the initialization.

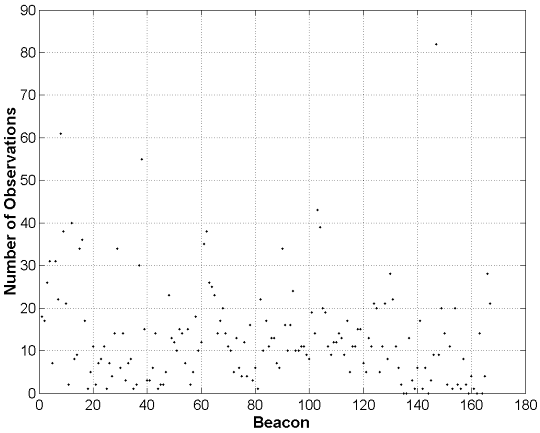

Figure 8: Number of Updates per Beacon in Farm data run.

Of the landmarks that were considered real, the mean number of updates per beacon was 13.287 with a standard deviation of 12. This was performed with all correlation tests (for initialization & data association) at 0.95 confidence. Mismatches in the data association could not be recorded in this test.

Figure 9 shows the resulting vehicle path and the map of the features used in the SLAM algorithm, with the dots representing landmarks. This was obtained from the hockey field data.

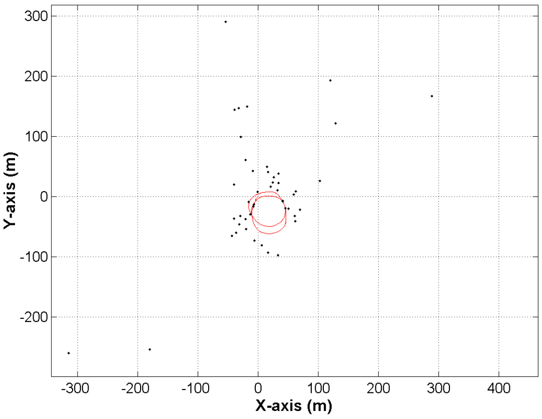

Figure 9: Vehicle Path and Feature Map.

Further results of localization, feature identification and data association aspects will be presented at the conference.

8 Conclusions

This work provides a template for the use of cameras in SLAM as bearing only sensors. The concepts covered were: modelling the camera for bearing observations, selection of point features for outdoor environments, initialization of landmarks from bearing observations, and the application of statistical and image processing techniques for data association.

References

[Elfes, 1989] Elfes A., "Occupancy Grids: A Probabilistic framework for Robot Perception and Navigation", PhD. Thesis, Department of Electrical Engineering, Carnegie Mellon University, 1989.

[Stentz *et al.*, 1999] Stentz A., Ollis M., Scheding S., Herman H., Fromme C., Pedersen J., Hegardorn T., McCall R., Bares J., Moore R., "Position Measurement for Automated Mining Machinery", *Proc. of the Int. Conferrence on Field and Service Robotics*, August 1999, pp 299-304.

[Durrant-Whyte, 1996] Durrant-Whyte Hugh F., "An Autonomous Guided Vehicle for Cargo Handling Applications". *Int. Journal of Robotics Research*, 15(5): 407-441, 1996.

[Leonard and Durrant-Whyte, 1991] Leonard J., Durrant-Whyte H., "Simultaneous Map Building and Localization for an Autonoumous Mobile Robot", *Proc. of IEEE Int. Workshop on Intelligent Robots and Systems*, pp 1442-1447, Osaka, Japan, 1991.

[Williams *et al.*, 2000] Williams SB., Newman P., Dissanayake MWMG., Rosenblatt J., and Durrant-Whyte H., "A Decoupled, Distributed AUV Control Architecture", *31st International Symposium on Robotics 14-17 May 2000, Montreal PQ, Canada.*

[Guivant *et al.*, 2000] Guivant J., Nebot E., Baiker S., "Localization and Map Building Using Laser Range Sensors in Outdoor Applications", *Journal of Robotic Systems*, Volume 17, Issue 10, 2000, pp 565-583.

[Guivant and Nebot, 2001] Guivant J., Nebot E.M., "Optimization of the Simultaneous Localization and Map Building Algorithm for Real Time Implementation", *Proc. of IEEE Transaction of Robotic and Automation*, vol 17, no 3, June 2001 pp 242-257.

[Dellaert *et al.*, 2000a] Dellaert F., Seitz S., Thorpe C., Thrun S., "Structure from Motion without Correspondence", *Proc. of IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition (CVPR'00)*, June, 2000

[Dellaert *et al.*, 1999b] Dellaert F., Burgard W., Fox D., Thrun S., "Using the Condensation Algorithm for Robust, Vision-based Mobile Robot Localization", *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Fort Collins, CO, 1999.

[Fox *et al.*, 1999] Fox D., Burgard W., Dellaert F., Thrun S., "Monte Carlo Localization: Efficient Position Estimation for Mobile Robots", *Proc. of the Sixteenth National Conf. on Artificial Intelligence (AAAI'99).*, July, 1999.

[Davidson and Murray, 1998] Davidson A., Murray D., "Mobile Robot Localization Using Active Vision", *European Conf. on Computer Vision* (*ECCV)* 1998

[Lacroix *et al.*, 2001] Lacroix S., Jung I., Mallet A., "Digital Elevation Map Building from Low Altitude Stereo Imagery", *9th Symposium on Intelligent Robotic Systems* (*SIRS)*, Toubouse, July 2001

[Mohr and Triggs, 1996] Mohr. R & Triggs. B., "Projective Geometry for Image Analysis", *Tutorial at Intl Soc. for Photogrammetry and Remote Sensing, (ISPRS), Vienna,* 1996

[Tomasi and Kanade, 1991] C. Tomasi & T. Kanade., "Shape and Motion from Image Streams: A Factorization Method - Part 3 : Detection and Tracking of Point Features", Tech. report CMU-CS-91-132, Computer Science Department, Carnegie Mellon University, April, 1991.

[Gonzalez and Wintz, 1987] R. C. Gonzalez & P. Wintz, *Digital Image Processing: 2ed.*, Addison-Wesley Publishing Company, 1987.