Bearing-Only SLAM using Colour-based Feature Tracking

Trevor Fitzgibbons

Australian Centre for Field Robotics

Rose Street Building J04

The University of Sydney, 2006.

tfitzgibbons@acfr.usyd.edu.au

Eduardo Nebot

Australian Centre for Field Robotics

Rose Street Building J04

The University of Sydney, 2006.

nebot@acfr.usyd.edu.au

Abstract

This paper presents identifies and addresses the difficulties that arise from implementing visual information into the Simultaneous Localization and Mapping (SLAM) problem, with an emphasis for outdoor applications. Through identifying these problems, techniques for integrating the visual & navigation are proposed with results from their preliminary applications. Video data is gathered through a standard colour camera. With the relative bearing obtained from the extracted features, the Simultaneous Localization & Mapping framework then bounds the dead-reckoning errors.

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 & Durrant-Whyte, 1991].

The extended Kalman Filter (EKF) can be used to solve a SLAM problem [Leonard & Durrant-Whyte, 1991],[Williams & Newman, 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 & 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., 2000] 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.,1999],[Fox et al., 1999], who used a Monte Carlo filter to localize their position, and both [Davidson & 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. Furthermore, the extraction of an estimate from any two observations is a non-linear problem, due to the lack of constraint on the relative range of the feature. This introduces problems as to how to introduce such an estimate into a SLAM filter or how to represent its probability distribution for error evaluation. This is a similar problem to the identification of an object by it's colour, or more precisely, it's colour distribution. These will be addressed with techniques that enable such bearing-only observations to be used in the SLAM framework as well as defining an object's particular colour signature.

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 visual 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 itself. These two components are known as the illumination (α) & reflectance components (r), with the light intensity (p) being the product of the two [Tomasi & Kanade, 1991]. 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, $[C_u, C_v]$. The position of the object on the image plane, p, is the projection of the pencil from P to the perspective origin, O.
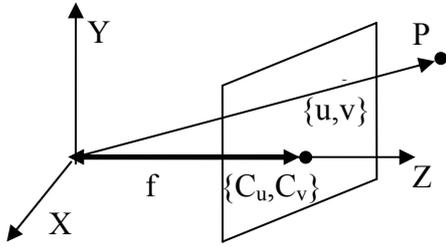

Figure 1 - Camera Reference Frame

The camera model can then be expressed as

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \begin{bmatrix} X/Z \\ Y/Z \\ 1 \end{bmatrix} \quad where \quad K = \begin{bmatrix} f_u & 0 & C_u \\ 0 & f_v & C_v \\ 0 & 0 & 1 \end{bmatrix} \qquad (2)$$

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

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.
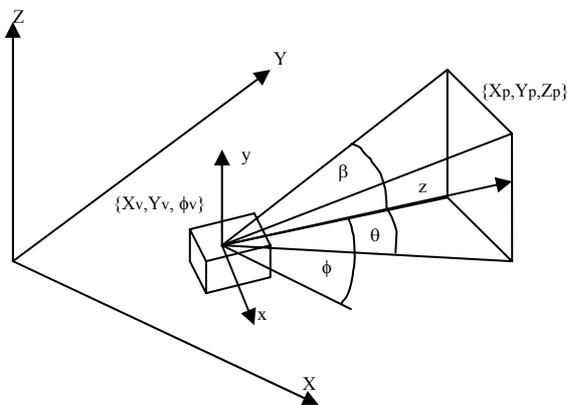

Figure 2 - Camera Model related to the Reference Frame

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;

$$\theta_i = \tan'\left(\frac{u_i - C_u}{f_u}\right) = \tan'\left(\frac{X_i}{Z_i}\right) \qquad (3)$$

The same bearing relative to the vehicle are described as

$$\theta_i = \phi_v - \arctan\left(\frac{y_i - y_v}{x_i - x_v}\right) \qquad (4)$$

With 'u_i' the observation we are using, the observation model becomes;

$$u_i = f_u \tan\left(\phi_v - \arctan\left(\frac{y_i - y_v}{x_i - x_v}\right)\right) + C_u \qquad (5)$$

3 Simultaneous Localization & Mapping

The SLAM algorithm [Guivant & Nebot, 2001] addresses the problem of a vehicle with known kinematics, 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 planning 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.
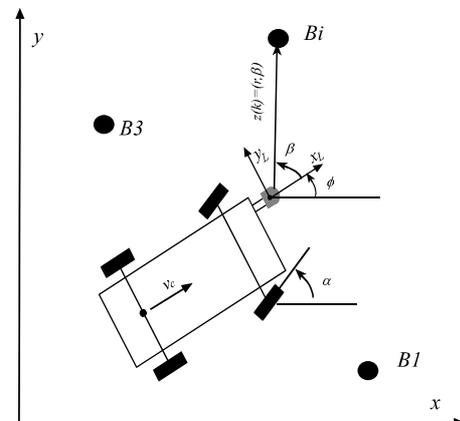

Figure 3 – Vehicle Coordinate System

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

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

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 (7), 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:

$$Z = \begin{bmatrix} z^1 \\ \vdots \\ z^m \end{bmatrix} \qquad (7)$$

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:

$$X = \begin{bmatrix} X_L \\ X_I \end{bmatrix}$$

$$X_L = (x_L, y_L, \phi_L)^T \in R^3 \qquad (8)$$

$$X_I = (x_1, y_1, .., x_N, y_N)^T \in R^{2N}$$

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:

$$X_L(k+1) = f(X_L(k)) + \gamma$$
$$X_I(k+1) = X_I(k) \qquad (9)$$

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

$$\frac{\partial F}{\partial X} = \begin{bmatrix} \dfrac{\partial f}{\partial \tilde{x}_L} & \varnothing \\ \varnothing^T & I \end{bmatrix} = \begin{bmatrix} J_1 & \varnothing \\ \varnothing^T & I \end{bmatrix} \qquad (10)$$

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

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*,

$$X(k+1,k) = F(X(k,k), u(k))$$
$$P(k+1,k) = J \cdot P(k,k) \cdot J^T + Q(k) \qquad (11)$$

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

$$S(k+1) = H \cdot P(k+1,k) \cdot H^T(k+1) + R(k+1)$$
$$W(k+1) = P(k+1,k) \cdot H^T(k+1) \cdot S^{-1}(k+1)$$
$$\vartheta(k+1) = Z(k+1) - h(X(k+1,k)) \qquad (12)$$
$$X(k+1,k+1) = X(k+1,k) + W(k+1) \cdot \vartheta(k+1)$$
$$P(k+1,k+1) = P(k+1,k) - W(k+1) \cdot S(k+1) \cdot W(k+1)^T$$

Where

$$J = J(k) = \frac{\partial F}{\partial X}\bigg|_{(X,u)=(X(k),u(k))}, \qquad H = H(k) = \frac{\partial h}{\partial X}\bigg|_{X=X(k)} \qquad (13)$$

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

4 Colour Feature – Selection & Tracking

Several advantages arise from the use of colour as a property for the use on feature selection and tracking. Firstly, in the case of the Red-Green-Blue scheme, a large range of identifiable colours are possible, which allows strong recognition. Furthermore, if we consider white-light illumination on the scene, then we can approach the measurement of the colour, independent of the level of illumination. Such colour measures commonly used are normalized RGB, CMY, YIQ & HSI [Tomasi & Kanade, 1991]. For the purposes here, normalized RGB has been used in these experiments. Using the principles used in equation (1), the normalized RGB measure, is obtained by;

$$\hat{R} = \frac{R}{R+G+B} \quad \hat{G} = \frac{G}{R+G+B}$$
$$\hat{B} = \frac{B}{R+G+B} \qquad (14)$$

; and eliminates the need to distinguish the illumination.

The concept of identifying features by unique intensity distribution has been investigated in recent years [Cornelissen & Groen, 2002], due to the fact that the normalized intensity distribution is invariant to motion, with their limitation lying within the number of pixels viewed, which is affected by scaled, angle of view and occlusion. The normalized intensity distribution is taken from the histogram of intensities.

To extract our features solely by colour, our approach has been to define a base model distribution for the features under investigation. The model is obtained from a series of images of the features, with the overall colour

histograms normalized to produce a raw distribution. From this, a Sum-of-Gaussians (SOG) distribution is fitted to the data. Figure 4 shows the colour model used in the identification of test poles, scaled by 256.
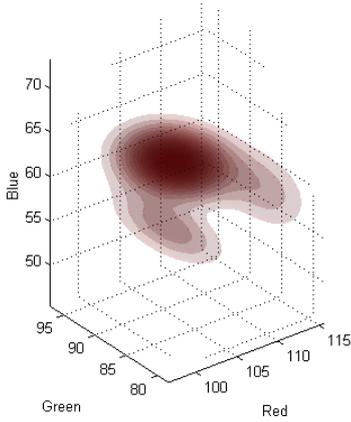

Figure 4 – Sum-of-Gaussian Distribution in RGB

To extract features from an image, each pixel is subjected to statistical gating to the SOG model. This creates a binary image of those that are within the bounds of the model, and those that are not.

Applying a region-growing segmentation algorithm to this binary image with the boundary function allowing growth only with pixels previously accepted as fitting with the model.

After heuristic removal of segments that are too small to provide reasonable amount of information, the remaining segments are taken as features.

The figure 5 shows application of this approach, using the normalized RGB model of the pole, shown in figure 4.

Figure 5 – Segmented Image using SOG Colour Model

The white segments are features that have been extracted, with the red line the centroid in a 2D plane. The acceptance of pixels other than the pole's but close to it are directly from noise in the camera, but more so due to bleeding of colours into each other from the camera's CCD display.

5   SLAM Initialization

5.1   Bearing-Only SLAM Initialization

The basis of this investigation was, that as visual information on the features in our environment is return as

relative bearing to the SLAM framework, the number of degrees of information will always be one less than what is required to estimate the landmark's position. In this case, at least two observations will always be required. But any triangulation to achieve the landmark's position is non-linear, with results that may not be suitably be fitted by a Gaussian distribution. But with enough observations of the landmark from varying angles, we can constrain the probability distribution until it approaches a Gaussian fit.

Since the distribution can not be easily modelled, we have approached this through the use of particle filters [], to track the landmarks before they are added to the map.

The state vector consists of landmark's position;

$$\underset{\sim}{x}_i = \begin{bmatrix} x_p \\ y_p \end{bmatrix} \qquad (15)$$

; with the initial distribution of particles dictated by the vehicle and bearing uncertainty model as Gaussian and a range term with an uniform distribution from 0 to ∞ (or a large enough number to represent ∞).

Each consecutive bearing observation of the feature is then used to update the particle distribution, as outlined in [Gordon et al., 1993]. At each particle update, before the application of resampling, the particle distribution is tested as to whether is has a Gaussian distribution. The approach for this is the apply a Goodness of Fit test [Montgomery, 1998] to the hypothesis that the particle distribution is Gaussian, with the Chi-squared measure given by;

$$\chi_0^2 = \sum_{i=1}^{N} \frac{(d_i - e_i)^2}{e_i} \qquad (16)$$

; where for all sample bins, d_i is the number of particles found in the ith bin and e_i is the expected number of particles in the ith bin. e_i is calculated from the number of particle and the distribution under question, which in this case is a Gaussian distribution.

If the hypothesis is found to be true, by;

$$\chi_0^2 \leq \chi_{\alpha, N-1}^2 \qquad (17)$$

; then we can extract the landmark estimate and uncertainty from the mean and covariance of the particle distribution, and incorporate it into our map.

5.2   Implementing Colour Distribution Tracking

The next logical step to integrating the visual aspects with the navigation side of the algorithm is to update each feature's colour distribution with each observation. Prior to initialization, the state vector for the particles can be extended to include the normalized RGB components, so that;

$$\underset{\sim}{x}_i = \begin{bmatrix} x_p \\ y_p \\ \hat{r} \\ \hat{g} \\ \hat{b} \end{bmatrix} \qquad (18)$$

; and the observation consisting of the bearing along with

the normalized RGB components from each pixel within the extracted segment. This has not been yet implemented.

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

$$\chi^2 \geq v^T S^{-1} v \qquad (19)$$

where v is the innovation, $v = \begin{bmatrix} (u_{obs} - u_{est}) & (v_{obs} - v_{est}) \end{bmatrix}^T$

Knowing that the observation has 1-degree-of-freedom, the Chi-squared level for 95% confidence is 5.02.
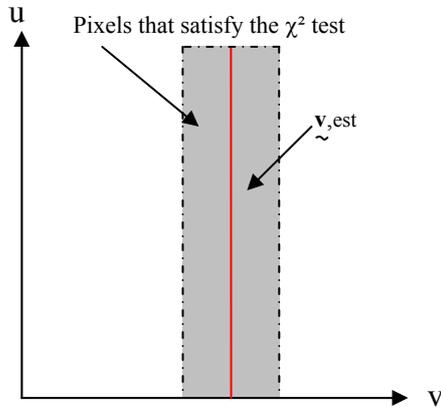

Figure 6 - Innovation Covariance projected onto Image Plane

The bounding function from the test can be expanded to represent an rectangular region in the image defined by (20).

$$5.02 \geq \frac{\Delta v^2}{S}$$

$$where; \qquad (20)$$

$$\Delta v = v - \hat{v}$$

; where, $\hat{v}$ is the expected centriod and S is the innovation covariance, which are all scalar.

Since the current tests are done in 2D with landmarks that can provide a 2D centroid, this approach is suitable for our purposes. Previous research used this approach for point features, using ellipsoid boundaries [Fitzgibbons & Nebot, 2001].

As long as the feature's centriod coordinates lie within this rectangle, it satisfies the chi-squared test and is thus a valid candidate for a match. Using this, the rectangle becomes the bounds for potential seed pixels for the feature extraction process. As such, after the image has been segregated, those pixels that potentially represent the feature and lies within the above defined boundaries are

used to seed the region-growing segmentation.

If it is found that an extracted feature's centroid lies within the set boundaries, the feature is taken as a possible observation of the landmarks, and is added to a list of individually validated observations. From this list, the set with the highest number of jointly-compatible matches obtained by employing the Joint Compatibility Branch & Bound technique [Neira & Tardos, 2000]. This filters out mismatches using the EKF's statistical information.

7 Results

Current tests in the application of particle filter SLAM initialization have been performed using bearing & range lasers and compared to true paths obtained from differential GPS and range & bearing SLAM. The plot below shows one such run.
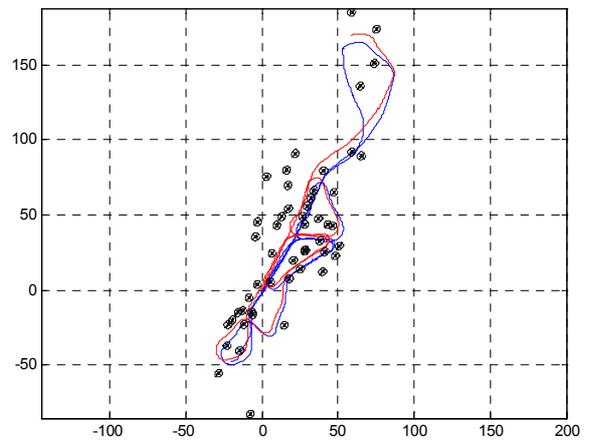

Figure 7 – Comparison of Bearing-Only SLAM to Range & Bearing SLAM (grid in meters).

The plot compares the bearing-only SLAM (red) with the range & bearing SLAM (blue) which was verified by GPS, with the tracked landmarks in black. The Bearing-only SLAM follows the same path as the blue path when it is in view of updatable landmarks. This is due to the maintenance of the information of the observations over time. The offsets in the paths, noticeably at the beginning and moving to the top part of the loop, are from the delay in initialization of landmarks, which puts the paths out by a translation error of 5m throughout the run.

Currently the colour tracking has only been applied to full sequences of with features extracted, as shown in the previous section. The application of Innovation Covariance bounds have already been applied to tracking landmarks as point features [Fitzgibbons & Nebot, 2001], and are currently being applied to the SLAM framework.

8 Conclusions

The results from the current research and previous applications that made the basis of the material covered here, point to this SLAM application as not only practical but also robust.

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. Conference 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 & 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 & Newman, 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 & 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., 2000] 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.,1999] 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 & 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

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

[Neira & Tardos, 2000] J. Neira & J.D. Tardos, "Data Association in Stochastic Mapping: The fallacy of the Nearest Neighbour", W4: Mobile Robot Navigation and Mapping workshop at IEEE International Conference on Robotics and Automation, 2000.

[Cornelissen & Groen, 2002] L. A. Cornelissen & F. C. A. Groen, "Automatic Colour Landmark Detection and Retrieval for Robot Navigation", 7th International Conference on Intelligent Autonomous Systems, 2002.

[Montgomery, 1998] D. C. Montgomery, "Engineering Statistics", John Wiley & Sons, Inc., 1998

[Fitzgibbons & Nebot, 2001] T. Fitzgibbons & E. Nebot, "Application of Vision in Simultaneous Localization & Mapping", Australian Conference on Robotics & Automation 2001.

[Gordon et al., 1993] N. J. Gordon, D. J. Salmond & A. F. Smith, "Novel Approach to Non-Linear/Non-Gaussian Bayesian State Estimation", IEE Proceedings, Vol. 140, No. 2, April 1993.