ROBUST PLACE RECOGNITION WITHIN MULTI-SENSOR VIEW SEQUENCES USING BERNOULLI MIXTURE MODELS

Filipe Ferreira [*,1] **Vitor Santos** [*] **Jorge Dias** [**]

[*] *Department Of Mechanical Engineering, University of Aveiro, Portugal*
[**] *Department Of Electronics Engineering and Computer Science, University of Coimbra, Portugal*

Abstract: This article reports on the use of Hidden Markov Models to improve the results of Localization within a sequence of Sensor Views. Local image features (SIFT) and multiple types of features from a 2D laser range scan are all converted into binary form and integrated into a single, binary, Feature Incidence Matrix (FIM). To reduce the large dimensionality of the binary data, it is modeled in terms of a Bernoulli Mixture providing good results that were reported in an earlier presentation. We have improved the good performance of the approach by incorporating the Bernoulli mixture model inside a Bayesian Network Model, an HMM, that accumulates evidence as the robot travels along the environment.

Keywords: Bernoulli Mixture model, Binary data, Expectation Maximisation, Dimensionality reduction, Robot Localization.

1. INTRODUCTION

Improving the robustness of localisation is a critical problem in the context of appearance and view-based localization since the appearance of an environment changes over time. Previous work by the authors(Ferreira *et al.*, 2006) lies at the heart of the place recognition approach presented here. The method can handle a large number of features originating from multiple sensors. After leading the robot, once, through a path in the environment, and allowing it to collect a sequence of [sensor] Views, our method allows the robot to localize itself when it travels through the same stretch of environment(within the original sequence of Views), a second time. The problem is reduced to the alignment of two sequence of views as shown in Fig. 1.

We first present examples of some approaches that have used range finders and cameras to perform view-based localization. Methods that depend on range sensors have used landmark and free-space boundary depictions to represent places. To increasing sensory reliability, many range-sensor based methods extract lines and other primitive features from the range scans. The extraction of lines from the laser scan continues to be a popular approach in the robust segmentation of laser scan data, see (Nguyen *et al.*, 2005) and (Sack and Burgard, 2004) for recent reviews of popular line-extraction algorithms. Other approaches eschew segmentation into simple primitive features and favour the description of the 2D Laser scan in some reduced variable space, such as in (Sooyong Lee, 2000) where each feature extracted from the laser range scan is given a symbol and each scan is described in the form of a string for example mMmMmMmMmDCm, where the string alphabet in this case is (M)axima, (D)iscontinuity, (m)inima, (c)onnection).

[1] Partially supported by the EU-BACS FP6-IST-127041 project

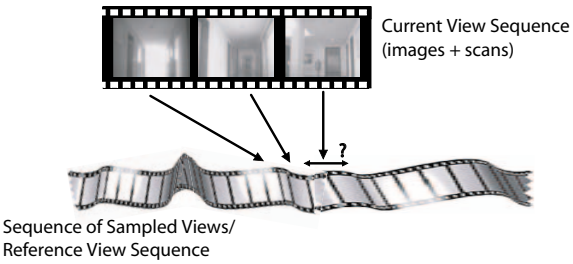


Fig. 1. A schematic description of the Localization problem

In our work we have used multiple types of range scan features, namely 1)wall-like (line) features, 2)scan contour (HU) properties and 3)scan region properties. Line segments were extracted using the incremental method (Nguyen *et al.*, 2005). Scan countour features were identified at scan discontinuities and the values of the Hu Moments (Gonzalez and Woods, 2002), (Hu, 1962), for regions around these discontinuities were used as descriptors. A further set of features included some properties of the laser scan including the area within the scan and its distribution around the range scanner.

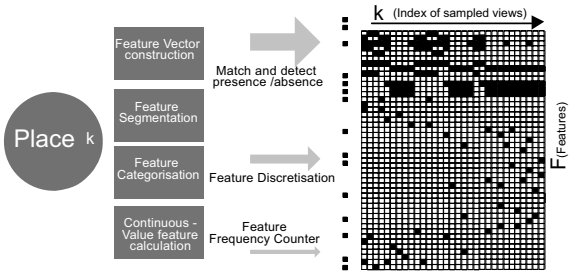
The use of cameras on mobile robots has become widespread over the last few years. Feature extraction for vision-based robots varies from local-image descriptors to global image properties derived over the entire image. Both approaches seek to avoid having to store the entire image itself. In seminal work, Murase and Nayar (Murase and Nayar, 1997), attempted to represent objects in terms of a 'parametric Eigenspace representation'. Among other feature extraction methods, Baker (Baker, 1998), attempts to create a generalised descriptor for local image features and the introduction to his thesis provides a perspective on the development of gradient based methods.

The stability and repeatability of points extracted at local Maxima (or Minima) in gradient images that have been repeatedly smoothed using operators, has been known for some time (Koenderink, 1984) (A.P.Witkin, 1983), and research in the field finally culminated in the Scale-Space theory proposed by Lindeberg(Lindeberg, 1994). In work that combined the lessons of Scale-Space with the reliable characterisation of features, Lowe (Lowe, 1999) describes the use of gradient histograms taken at various points close to some point of interest. Since their introduction, SIFT features have been widely applied, among various applications, to object recognition (Pope and Lowe, 2000) (Lowe, 2001), in the panoramic assembly of images (Brown and Lowe, 2003) and in image retrieval (Ke *et al.*, 2004).

We typically extract between fifty and two hundred SIFT features per image and have adopted a simple procedure involving the creation of a number of intermediate KDTrees that are created from SIFT features extracted from images obtained as the robot progresses through the environment. The creation of these intermediate KDTrees would normally require the sorting of SIFT descriptors, and the pairwise comparison (without recourse to a KDTree) of descriptors required to check for duplicate descriptors. Adding a small amount of noise to the SIFT descriptors prior to the creation of the KDTree makes the creation of the KDTrees much faster, enabling the use of SIFT features for continuous image sequences, as described in Algorithm 1 in (Ferreira *et al.*, 2006).

Various approaches have been proposed to combine sensors and, given the variety of features-based methods using vision or range scans, the combinations are many (see the bibliography maintained by Keith Price at http://iris.usc.edu/Vision-Notes/bibliography/match-pl502.html). Place recognition, image retrieval and robot localisation methods (even single sensor platforms) typically make use of large numbers of features whose correlations among each other is unknown. Two principal approaches to feature integration are possible; filter-based and wrapper based. In filter-based methods, physical sensor models are imposed on new data as it comes in. wrapper-based approaches, on the other hand, attempt to facilitate the NP-hard mathematical procedures that are used to approximate the integration of all features simultaneously (Kohavi and John, 1997). Among methods that use the latter approach, some, such as (Newman *et al.*, 2006), employ a distance metric based on the number of repeated features in the entire set of images and a 'Rank-reduction' method to identify the important similarities between images. Others such as (Marsland *et al.*, 2001) attempt to 'learn' the features or landmarks that are good and use these for localisation.

For place recognition, speech recognition and other procedures that use a large number of features and which seek to explicitly reduce the dimensionality of the 'feature space', Principal Component Analysis (PCA) and more application-specific methods derived from PCA constitute an important class of data-reduction methods. Mixture models are another common solution to modeling data that is believed to follow non-parametric distributions and reducing its dimensionality, (Sajama and Orlitsky, 2005) (McLachlan and Peel, 2000). There is previous work that goes some way to demonstrate the usefulness of binary features (Kaban and Girolami, 2000) (Wang and Kaban, 2005) by modelleing binary data as mixtures of appropriate distribitions. Mixtures of Bernoulli distributions have been used to model data containing binary features, (Juan and Vidal, 2004), (García-Hernández *et al.*, 2004) and (Gonzalez *et al.*, 2001). Converting features into binary form offers significant advantages, the main ones being that binary data can represent both qualitative and categorical data and that this scheme allows us to integrate very disparate variables. Given that we wish to integrate thousands of features, in real-time we look at approximate techniques to reduce the dimensionality of the features.



(a) Creation of a view of Binary Features.(b) The Feature Incidence Matrix.

Fig. 2. The creation of binary features is performed in different ways for different sensors.

We have extracted features, using the methods described above, and converted them into binary form by one of the following 1)matching extracted features against a feature database to detect their presence (or absence), 2)categorising features and 3)discretizing continuous-value features as seen in 2a. We end up with a matrix of binary values Fig. 2b, where each row denotes a particular feature that was extracted from at least one image or laser range scan. Each column represents a place at which an image or scan was obtained. The presence of a 'one' in any column signifies that the feature was observed in an image or laser scan taken at that place.

In the next section, we shall briefly review the application of our method to integrate laser range finder and vision features for place-recognition, details of which appear in an earlier publication, (Ferreira *et al.*, 2006). In section 3 we shall present a framework by which the same procedure can be applied to multiple views in the Reference Sequence to improve the robustness of the localization process. Section 4 concludes the article by reviewing the results and providing suggestions for future work.

2. INTEGRATING LASER AND VISION FEATURES FOR PLACE RECOGNITION

Our robot platform is equipped with cameras capable of taking VGA- images and a SICK laser range finder which provides a set of 361 range measurements through a 180 degree interval, Fig 3.

Binary features from the Laser range scan are created by classifying the number of extracted lines and their distance from the range scanner, by matching the contour features and by classifying the free, open space within the range scan. In a similar way, in the case of the camera features, each SIFT feature in the KDTree is taken to be a separate binary feature.

The use of all the SIFT and LRF features results in a very large number of features, the information from all of which we want to integrate, in order to estimate the position of the robot. Each feature will be correlated, to varying extents, with other features. The correla-

Fig. 3. The Robuter mobile robot platform with two cameras and a Laser Range Finder.

tions between features will themselves be different at different parts of the environment, being significant in some regions and, in other regions, being not so significant. In order to capture some of these correlations and make better place-recognition estimates for the view that is currently available, we use a Bernoulli Mixture Model to classify the original, large number of features so that place recognition can be performed, in a smaller dimensional space, where the correlations between sets of features is taken into account.

To perform place recognition, the robot is first led through the environment during which the sensors sample the environment, generating a sequence of views, called the Reference Sequence. The record of binary features extracted from each of these views is represented within a Feature Incidence Matrix (FIM), $\mathcal{V}$. Each row i, of the FIM corresponds to a feature Y_i and each column j, to an index view, V_j, from the Reference Sequence (each entry in the FIM might be represented as $Y_{i,j}$ where the first subscript indicates the feature and the second subscript, the view). $Y_{i,j}$ takes value 1 if feature Y_i appears (is visible) in view V_j, 0 otherwise, see Fig. 2.

$$\mathcal{V} = \begin{bmatrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,K} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{N,1} & Y_{N,2} & \dots & Y_{N,K} \end{bmatrix} \quad (1)$$

Given that features arise in groups and persist/disappear as a result of the structure of the environment, an assumption of independence between the features does not hold. Inferences made by using this assumption would be biased toward certain views in the Reference Sequence as, in practice, some of the features are highly correlated while others are less. To address this problem, the Feature Incidence Matrix (FIM), $\mathcal{V}$ is modeled as a Bernoulli Mixture Model where any single View V^{obs} appears as a vector of binary features

$\{0,1\}^D$ which is obtained from a particular mixture of Bernoulli distributions, as in (2), where Θ denotes the parameters of the distribution of the views that compose our Mixture Model. These parameters include the M component vectors, the Θ_is, and the proportions in which these are mixed, the α_is. Each α_i represents the prior probabilities of the component i in the mixture model, subject to the constraint $\sum_i \alpha_i = 1$. The likelihood of matching the V^{obs} with each View k in the Reference Sequence can be determined using (3). The *Maximum Likelihood Estimation* approach is used to obtain the [best] matching view.

$$P(V^{obs}|\Theta) = \sum_{i=1}^{M} \alpha_i P_i(V^{obs}|\Theta_i) \tag{2}$$

$$P(V^{obs} = V_k) = \frac{\sum_{j=1}^{M} P(V_k) z_{ki} \alpha_j P(V^{obs}|\Theta_j)}{\sum_{k=1}^{K} \sum_{j=1}^{M} P(V_k) z_{kj} \alpha_j P(V^{obs}|\Theta_j)} \tag{3}$$

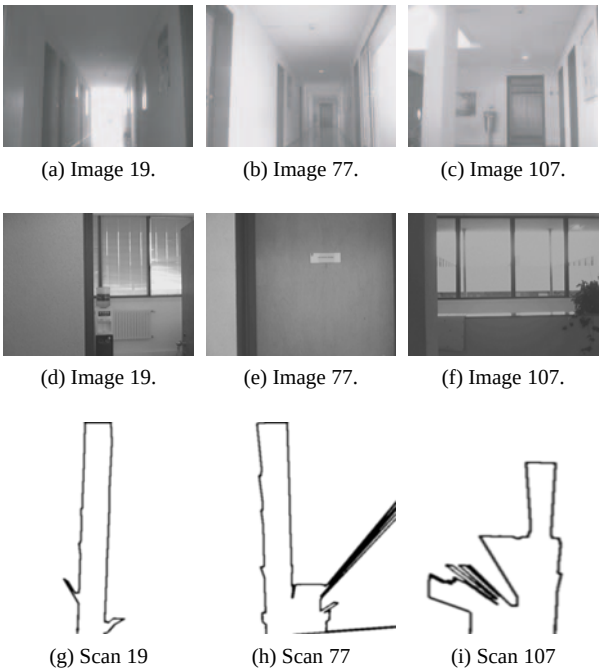
The parameters α_is, Θ_is and the Z(the hidden variables) of the Bernoulli Mixture Model and obtained by running the well known Expectation Maximisation (EM) Algorithm. More details can be found in (Ferreira *et al.*, 2006).

3. ROBUST PLACE RECOGNITION USING HIDDEN MARKOV MODELS

The views in a Reference Sequence are taken in real life conditions and could include people moving in the environment and very similar or unchanging stretches of environment. To gain robustness for our place recognition, we have attempted to integrate the information that is available from the matching of multiple Views within the Reference Sequence. This Reference Sequence is modeled as a simple, left-to-right Markov Chain as shown in Fig. 6a.

Since we know of only one route that connects each pair of consecutive views, the action/behaviour that is recorded along with each view in the Reference Sequence will take us to the next view and, any other action will take us somewhere else (where, we do not know!). Also, depending on the frequency with which the scans and images are taken during localization, relative to the frequency of sampling in the Reference Sequence, the robot might some times end up in-between the Views of the Reference Sequence. As a result, the Markov Chain depicted in Fig. 6a will be modified to Fig. 6b, where a $Lost\_Place$ is inserted between every pair of places in the original Reference Sequence. The incomplete, dotted lines represent the state transitions that have not been drawn in order to avoid cluttering the figure.

This modified Markov Chain is used as a model for the transition between the 'hidden states' of the Hidden Markov Model, shown in Fig. 5. As a result of this,



(a) Image 19. (b) Image 77. (c) Image 107.

(d) Image 19. (e) Image 77. (f) Image 107.

(g) Scan 19 (h) Scan 77 (i) Scan 107

Fig. 4. Representative images and laser range scans from a sequence taken by Camera 1 (top row), Camera 2 (middle row) and the Laser range finder (bottom row).

the parameters of the HMM are expressed as in (4), where $N = 2 \times K$ corresponds to the number of states, $M = K + 1$ the total number of possible observations, π represents the initial probability on the states, the α_{ij}s correspond to the transition probabilities between a pair of states i and j and $b_i(n)$ represents the probability of viewing symbol m at state n. An additional, hypothetical, observation is added to the existing observations, i.e. the views of the Reference Sequence. This observation is the most likely observation that can be obtained at any one of the lost places.

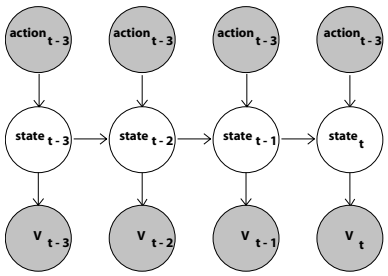
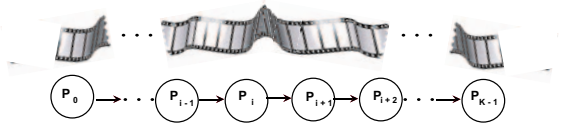


Fig. 5. A classic representation of a HMM for localization in a Reference Sequence showing an action that will propel the robot from one view or place to another.

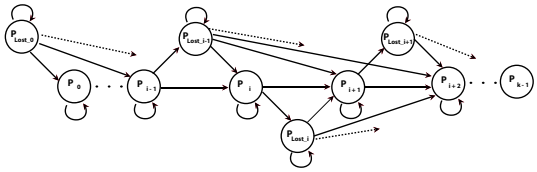
$$\lambda = \langle N, M, \{\pi_i\}, \{a_{ij}\}, \{b_n(m)\} \rangle \tag{4}$$

$$b_n(LostView\_m) = \frac{1}{K+1} \tag{5}$$

The Viterbi algorithm, a type of Dynamic Programming algorithm, is commonly used in the context



(a) Markov Chain for the original Reference Sequence



(b) Markov Chain for the Complete Reference Sequence with Lost Places

Fig. 6. The original Reference Sequence, at top, is modified to create the complete Reference Sequence, at bottom, by introducing 'lost places' in between original views.

of HMMs to determine the most probable sequence of hidden states (Places) that gave rise to a particular sequence of observations (Views)(Forney, 1973), (Rabiner, 1989). Using one hidden state at a time, the Viterbi algorithm calculates all the outcomes that could be possible for that state - and then keeps only the most likely sequence of states. After traversing the length of the HMM, the 'surviving' sequence of places is the sequence that is most likely to have generated the complete sequence of observations.

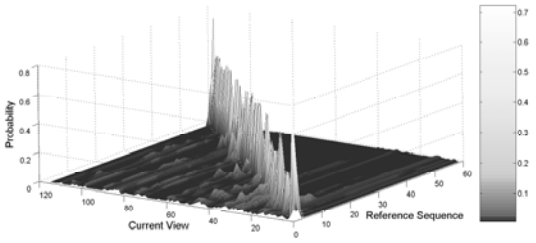
The robot is moved along a stretch of corridor and 2 camera images and a laser range scan obtained at regular intervals. The images and laser scans obtained at three places in the environment are shown in Fig. 4. The robot is then guided along the same stretch of corridor and, once more, acquires images and scans using to perform place recognition against the Reference Sequence.

The application of the Bernoulli mixture model to the 2 cameras and a laser range finder was evaluated over the entire path and the posterior probability distribution over all the views in the Reference Sequence is shown in Fig. 7, with and without recourse to HMMs.

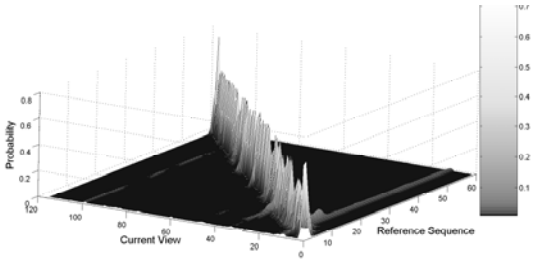
The results of the application of a plain Bernoulli Mixture model (the posterior probability distribution over all the views in the Reference Sequence) to integrate features from all three sensors, but without using an HMM, is shown in Fig. 7a. No motion model was utilised and the prior probability in (3) was assumed to be uniform(there would be no consistent way of maintaining such a probability without explicit use of an estimation filter).

The use of the HMM allows us to modify the prior distribution for Place recognition of subsequent Views. By constraining the positions that the robot can take at any time the place recognition results become much more reliable, as can be seen in Fig. 7b.

The experiment was repeated a number of times for the same Reference Sequence. The failed attempts at



(a) Cam 1 + Cam 2 + LRF, Single View localisation



(b) Cam 1 + Cam 2 + LRF, Localization using an HMM with 5 consecutive Views

Fig. 7. Posterior Probability distribution when comparing two sequences.

Table 1.

Mission Number	Reference Sequence	No-HMM failure.	HMM failure.
1	6	5	2
2	6	2	2
3	6	5	1
4	6	6	1
5	6	3	3
6	6	3	0

Place Recognition are compared, in Table 1, for the cases in which the HMM was used and that in which no HMM was used. As can be seen there are situations in which the HMM was not able to improve on the number of Place-recognition failures mostly because the environments had changed too much since the creation fo the Reference Sequence, because of lighting conditions or because of the presence of people.

4. CONCLUSIONS

Robustness in the place recognition has been increased by accumulating evidence from sequential views using a Hidden Markov Model. Place recognition is performed independently for each view using the Bernoulli Mixture model developed earlier. The use of the Hidden Markov Model allows the introduction of a prior probability in the Bernoulli Mixture Model in a consistent way which greatly improves the Place Recognition results and seems to be a promising approach for appearance-based localization methods to deal with dynamic environments.

Improvements that must be made include the development of schemes to handles features from sensors with different error models. We need to make modifications to our application of the Bernoulli Mixture Model so

that variation of more accurate features appearing in smaller numbers is taken into account. We are also looking at ways to modify the parameters of the HMM in order to improve the probability of correctly detecting the places represented in the Reference Sequence.

References

A.P.Witkin (1983). Scale-space filtering. In: *Proc. 8th Joint conference on Artificial Intelligence*. 8th Joint conference on Artificial Intelligence. Karlsruhe, W. Germany. pp. 1019–1023.

Baker, Simon (1998). Design and Evaluation of Feature Detectors. PhD thesis. Columbia University.

Brown, Matthew and David G. Lowe (2003). Recognising Panoramas. In: *Tenth International Conference on Computer Vision (ICCV 2003)*.

Ferreira, Filipe, Vitor Santos and Jorge Dias (2006). Integration of multiple sensors using binary features and a bernoulli mixture model. In: *IEEE Conference on Multisensor Fusion and Integration*. Heidelberg, Germany.

Forney, G. D. (1973). The viterbi algorithm. In: *Proceedings of the IEEE*. Vol. 61. pp. 268– 278.

García-Hernández, José, Vicent Alabau, Alfons Juan and Enrique Vidal (2004). Bernoulli mixture-based classification. In: *Proc. of the LEARNING04, ISBN 84-688-8453-7* (A. R. Figueiras-Vidal et al., Ed.). Elche (Spain).

Gonzalez, J., A. Juan, P. Dupont, E. Vidal and F. Casacuberta (2001). A Bernoulli Mixture Model for Word Categorisation. In: *Symposium Nacional de Reconocimiento de Formas y Analises de Imagenes*.

Gonzalez, R. C. and R. E. Woods (2002). *Digital Image Processing*. Addison-Wesley Pub. Co.

Hu, Ming-Kuei (1962). Visual pattern recognition by moment invariants. *IEEE Transactions on Information Theory* **8**(2), 179– 187.

Juan, Alfons and Enrique Vidal (2004). Bernoulli Mixture Models for Binary Images. In: *International Conference on Pattern Recognition (ICPR'04)*. Vol. 3. pp. 367–370.

Kaban, Ata and Mark Girolami (2000). Initialized and Guided EM-Clustering of Sparse Binary Data with Application to Text Based Documents. In: *Proceedings of the 15th International Conference on Pattern Recognition (ICPR'00)*. Vol. 2.

Ke, Yan, Rahul Sukthankar and Larry Huston (2004). An efficient parts-based near-duplicate and sub-image retrieval system. In: *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*. ACM Press. New York, NY, USA. pp. 869–876.

Koenderink, Jan J (1984). The Structure Of Images. *Biological Cybernetics* **50**(5), 363–370.

Kohavi, Ron and George H. John (1997). Wrappers for feature subset selection. *Artificial Intelligence* **97**(1-2), 273–324.

Lindeberg, Tony (1994). *Scale-Space Theory in Computer Vision*. Kluwer Academic Press.

Lowe, David G. (1999). Object Recognition from Local Scale-Invariant Features. In: *Proc. of the International Conference on Computer Vision, Corfu*. pp. 1150–1157.

Lowe, David G. (2001). Local Feature View Clustering for 3D Object Recognition. In: *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*. Kauai, Hawaii. pp. 682–688.

Marsland, Stephen, Ulrich Nehmzow and Tom Duckett (2001). Learning to select distinctive landmarks for mobile robot navigation. *Robotics and Autonomous Systems* **37**, 241–260.

McLachlan, Geoffrey and David Peel (2000). *Finite Mixture Models*. John Wiley and Sons.

Murase, Hiroshi and Shree K. Nayar (1997). Detection of 3d objects in cluttered scenes using hierarchical eigenspace. *Pattern Recognition Letters* **18**, 375–384.

Newman, P., D. Cole and K. Ho (2006). Outdoor slam using visual appearance and laser ranging. In: *ICRA 06*.

Nguyen, Viet, Agostino Martinelli, Nicola Tomatis and Roland Siegwart (2005). A comparison of line extraction algorithms using 2d laser rangefinder for indoor mobile robotics. In: *International Conference on Intelligent Robots and Systems*.

Pope, Arthur and David G. Lowe (2000). Probabilistic Models of Appearance for 3-D Object Recognition. *IJCV* **40**(2), 149 – 167.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**, 257–286.

Sack, Daniel and Wolfram Burgard (2004). A comparison of methods for line extraction from range data. In: *IAV 2004*.

Sajama and Alon Orlitsky (2005). Supervised dimensionality reduction using mixture models. In: *Proceedings of the 22 nd International Conference on Machine Learning*. Bonn, Germany.

Sooyong Lee, Nancy M. Amato, James Fellers (2000). Localization based on Visibility Sectors using Range Sensors. In: *Procceedings of the IEEE Int. Conference Robot. Autom. (ICRA),*. pp. 3505–3511.

Wang, Xin and Ata Kaban (2005). Finding Uninformative Features in Binary Data. In: *Sixth International Conference on Intelligent Data Engineering and Automated Learning (IDEAL '05)*.