Mobile Robot Localization from Learned Landmarks

Robert Sim Gregory Dudek

Centre for Intelligent Machines
McGill University
3480 University St.
Montreal, QC H3A 2A7

Abstract

This paper presents an approach to vision-based mobile robot localization. In an attempt to capitalize on the benefits of both image and landmark-based methods, we describe a method that combines their strengths. Images are encoded as a set of visual features called landmarks. *Potential landmarks are detected using an attention mechanism implemented as a measure of uniqueness. They are then selected and represented by an appearance-based encoding. Localization is performed using a landmark tracking and interpolation method which obtains an estimate accurate to a fraction of the environment sampling density. Experimental results are shown to confirm the feasibility and accuracy of the method.*

1 Introduction

In this paper we address the problem of encoding the visual characteristics of an environment to permit accurate positioning. We assume that we know what general area the robot is in, but we have no *a priori* estimate of its precise position. Traditional approaches to this problem involve the detection of manually inserted landmarks, followed by a position estimation step based on triangulation. In our work, the primary objective is to avoid the requirement for artificial landmarks, or domain-specific features. As such, our problem is related to object recognition, where we wish to learn visual characteristics of interest.

The problem of landmark-based position recognition was first formalized by Sugihara as a computational geometry problem [1]. Since then, the problem was further explored by Avis and Imai [2], Sutherland and Thompson [3], and Boley, Steinmetz and Sutherland [4], among others. These authors have all pursued the problem under two assumptions: first that the world is planar (although their methods can be extended to three-space), and second that the problem of detecting (and sometimes distinguishing) landmarks in the environment has been solved. Several authors have also considered positioning using non-visual landmarks [5, 6, 7]. In practice, position estimates from visual sensors are typically combined with those from odometry using methods such as Kalman filtering [8, 9].

The problem of detecting landmarks has been approached in a variety of ways. Many vision-based robot localization methods rely on landmarks which are either artificially added to the environment [10], or based on strong assumptions with respect to the environment [11, 12]. For example, Krotkov [11] relies on the assumption that the environment is structured in such a way that vertical lines can be easily extracted as landmarks. This assumption is problematic in two ways. First, it places a restriction on the kinds of environments that can be explored, and second, it places a restriction on the pose of the camera. Basri and Rivlin [13] have also exploited the geometric behavior of landmarks in selected model images to provide navigation information. Exploiting an assumption of global invertibility of the imaging function, Nayar has shown that subspace methods can provide accurate positional feedback in sufficiently constrained environments [14]. A key assumption in that work is that each possible viewing position gives rise to a unique image. In similar work, Thrun derives a probabilistic approach to obtain a pose estimate using a neural net [15]. In the works of both Nayar and Thrun, however, all significant variations in the set of possible images, including those due to lighting variations, must be explicitly sampled and encoded. In other work, it has been shown that localization can

[*] This papers appears in Proc. IEEE/RSJ Conf. on Intelligent Robots and Systems (IROS), Victoria, BC, Oct. 1998.

be achieved despite unanticipated illumination variations [16]. That method can also deal with non-invertibility of the imaging transform, a problem that is typical in unconstrained environments.

Our approach uses *image landmarks* to perform position estimation, but learns these landmarks from a preliminary traversal of the environment (i.e. an off-line mapping phase). Preliminary landmark selection is based on a local distinctiveness criterion: this is later validated by verifying the appearance of the candidate landmarks. In this aspect our approach is also related to feature-based image representation used, for example, for image registration by Zoghlami and Faugeras [17]. In that work, a corner detector was used to define landmarks for the construction of an image mosaic. We are interested in images selected from a much wider range of imaging geometries.

Our approach to landmark selection is inspired by models of human visual attention where visual saccades are drawn to regions of high edge density [18]. We select extrema of the density of the edge distribution in each image as *landmark candidates* and extract a subwindow about each one. We then perform principal components analysis on these subwindows to produce low-dimensional descriptions of the appearance of each of the observed landmarks. In an off-line learning phase, the subspace encodings are employed to build *tracked landmarks*, which correspond to sets of landmark candidates that are tracked over configuration-space.

The online localization method exploits variation in the appearance-based encoding and other measures of the observed landmarks as a function of camera position. When the camera is in an unknown position, candidate landmarks are extracted from the image, matched to tracked landmarks in the database, and an estimate is obtained for each matched landmark based on a linear interpolation of landmark *feature vectors.* A final position estimate is obtained through a selective merging of individual estimates.

Section 2 briefly discusses the motivation for our representation of what constitutes a landmark. Section 3 presents details of our approach. Section 4 presents experimental results and section 5 discusses their implications.

2 Visual Cues for Positioning

Luminance edges appear to encode much of the relevant geometric content in images, yet edge operators suffer from instability due to sensor noise or variations in lighting conditions. Ideally one might wish to connect edge elements from an image to obtain extended geometric edges. In practice, however, existing methods for this task are either costly, domain specific, or exhibit other limitations [19, 20, 21]. Given these limitations, if we smooth the output of an edge operator over a small neighborhood then we can consitently determine the neighborhood of the edge. To this end, we propose the use of edge element *density* over the neighborhood of a pixel in order to detect regions of interest without the cost of geometric interpretation. The extrema in edge density over the image appear to be stable under variations in camera position, and hence will make good candidates for image domain landmarks. Therefore we will define a *candidate landmark* as a local extremum of a measure of image feature content.

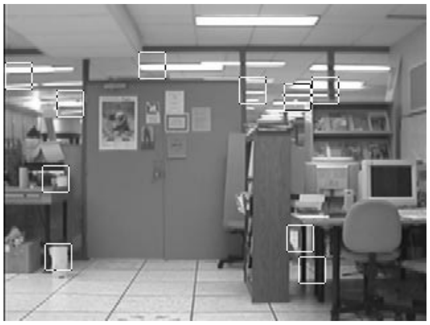

Figure 1: Detected Candidate Landmarks in an Image.

Figure 1 presents an example of the output of the landmark detector. The candidate landmarks, depicted as boxes, have been detected as local extrema in edgel density, as measured over a circular window of radius 10 pixels. Only those candidates which are maximal over their neighbourhood, and which exceed a user-defined threshold density are shown.

3 Method

Thus far, we have discussed a method for detecting *possible* landmarks in the environment. It should be noted at the outset that our method intentionally makes no restrictions on how landmark position in an image is related to position in the world. Landmarks might arise out of three-dimensional arrangements of arbitrary complexity. In addition, no restrictions are placed on camera pose itself. If the robot is moving over hilly terrain, landmarks will move in an irregular fashion, yet their position and appearance will still hold significant information for positioning. The key to our approach is the assumption that *locally*, the appearance and position of a good landmark can be predicted by a simple parametric function. Given that we cannot treat landmarks as projections of three-dimensional points, we are unable to invoke the standard motion estimation and triangulation methods [1, 2, 4, 11]. Recall also that we are interested in localization even in the absense of an *a priori* pose estimate, obviating the possibility of using Kalman filtering or optical flow techniques [8, 9].

Localization is a two-step process consisting of an *off-line* preprocessing stage and an *on-line* estimation stage. The off-line stage consists of building a representation of the environment in the form of a database, which is later used for positioning. The on-line stage uses the database to match currently observed landmarks to previously stored landmarks. Each of these matches are then used to compute individual position estimates, which are combined in a robust fashion to obtain a position estimate. The following subsections explain the details of each stage.

3.1 Building the Landmark Database

In order to describe the environment, images must be obtained from representative viewpoints. In practice, we select viewpoints that cover the pose space in a uniform grid. In ongoing work, we are considering methods to automatically select a minimal set of such viewing positions [22]. In the work described in this paper, viewpoints are selected such that the camera is facing in a consistent orientation, although this constraint can be relaxed using a technique described by Dudek and Zhang [16]. Once these images have been acquired, they are used to automatically compute a suitable set of landmarks for subsequent positioning.

In order to collect repeated observations of the same landmark from different viewpoints, we *track* observed landmarks over the database by incrementally growing *tracked landmarks*. The tracked landmarks are initially defined by the sets of single landmark candidates observed in a selected *bootstrap* image from the configuration-space (typically the centroid of the covered configuration-space). These landmark candidates then become templates for matching. Matching is based on a minimization of the Euclidean distance between the principal components encodings of the template and of the candidate landmark. Principal components analysis (PCA), sometimes referred to as *eigenfaces*, operates by constructing a linear subspace which maximizes the distance between the classes to be discriminated. PCA has enjoyed considerable success in the domains of face and object recognition, and is favoured over correlation and other methods for its desirable computational and numeric properties, particularly the maximization of the signal-to-noise ratio of the training set. [23, 14, 24, 25].

Given an intial set of templates, the candidate landmarks in each image are considered for inclusion in one of the sets. Consideration for inclusion in a tracked landmark is based on the following methodology:

1. For each candidate landmark l_i in the image, and
   (a) for each tracked landmark t_j in the database,
      i. perform a local search on the image in the neighbourhood of l_i for a better match to t_j, according to minimal Euclidean distance in the subspace.[1] If a better match l' is found, it replaces l_i as a candidate for t_j.
   (b) Select the tracked landmark t_j for which the best match to l_i was found in 1a.
2. If l_i is the best match to t_j over all other landmarks in the image and l_i matches t_j within a reasonable threshold, add it to t_j, otherwise, create a new set with l_i as the template.

The goal of this method is to grow tracked landmarks over pose space so that a candidate landmark can be matched to the correct target over a large portion of space. Figure 2 shows a typical tracked landmark. Each thumbnail image corresponds to the landmark as detected in the image taken at the corresponding grid position in camera space. Clearly, any changes in landmark appearance over this region are subtle.

3.2 On-line Localization

On-line localization is performed by matching candidate landmarks to tracked sets, and exploiting a transformation of each landmark into a subspace defined by its corresponding tracked set. This section will discuss the matching and estimation procedure.

When a position estimate is required, an image is obtained and landmark candidates are extracted. The extracted landmarks must then be matched to the *tracked landmarks* in the database. Matching is accomplished using the same procedure outlined above in Section 3.1. That is, each landmark l undergoes a local adjustment to find a best match to each tracked

[1] This search is employed in order to counter the effects of any instabilities in the underlying landmark detector.

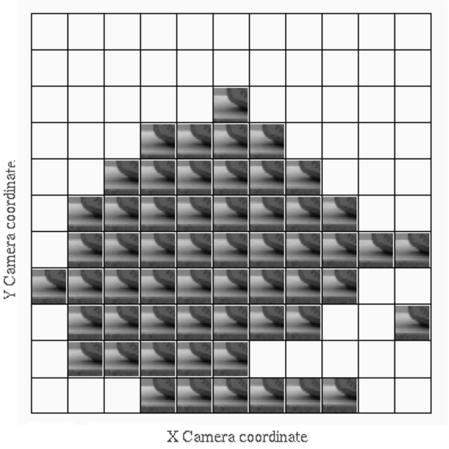

Figure 2: A typical tracked landmark. Each thumbnail corresponds to the landmark as detected in the image taken at the corresponding grid position in camera space. Grid positions where the image does not appear correspond to camera positions where the either the landmark simply wasn't detected, the landmark differed significantly from the template, or another nearby landmark dominated the local search for a better match.

set, and the set whose template is unambiguously closest to the encoding of l is selected as the match.

Once landmark matching is accomplished, we exploit an assumption of linear variation in the landmark characteristics in order to interpolate a position estimate for each match[2]. For the remainder of this section, let us assume that we have observed a single landmark l in the world and it has been correctly matched to tracked landmark T. Let us define a *feature-vector* $\mathbf{f}$ of a landmark as the initial principal components encoding of the landmark $\mathbf{k}$, which was the same subspace encoding used for matching, concatenated with two vector quantities: the image position $\mathbf{p}$ of the landmark, and the camera position $\mathbf{c}$ from which the landmark was observed:

$$\mathbf{f} = \left| \begin{array}{ccc} \mathbf{k} & \mathbf{p} & \mathbf{c} \end{array} \right| \tag{1}$$

Given $\mathbf{f}_i$ for each landmark l_i in the tracked landmark T, we construct a matrix $\mathbf{F}$ as the composite matrix of all $\mathbf{f}_i$, arranged in columnwise fashion, and then take the singular values decomposition of $\mathbf{F}$ to obtain $\mathbf{U}_F$, representing the set of decreasing eigenvectors of the feature vectors of T, arranged in columnwise fashion. Note that in this case, we have encoded camera position along with appearance. Now consider the feature vector $\mathbf{f}_l$ defined by l, the observed landmark for which we have no pose information. For the moment, let us assume that the $\mathbf{c}$ portion of $\mathbf{f}_l$ is initialised to the mean camera position of the landmarks contained in T^3. If we project $\mathbf{f}_l$ into the subspace defined by $\mathbf{U}_F$ to obtain

$$\mathbf{g} = \mathbf{U}_F^T \mathbf{f}_l \tag{2}$$

and then reconstruct $\mathbf{f}_l$ from $\mathbf{g}$ to obtain the feature vector

$$\hat{\mathbf{f}}_l = \mathbf{U}\mathbf{g} \tag{3}$$

then our observation is that the resulting reconstruction $\hat{\mathbf{f}}_l$ is augmented by a camera pose estimate that accurately interpolates between the nearest eigenvectors in $\mathbf{U_F}$. This assumes that the camera pose does not play a significant role in the subspace defined by $\mathbf{U}_F$- We aid this assumption by scaling down the value of $\mathbf{c}$ when we construct $\mathbf{f}$. In practice, the initial value of the camera pose will play a role in the resulting estimate, and so we substitute the new estimate back into $\mathbf{f}_l$ and iterate, reconstructing $\hat{\mathbf{f}}_l$ until the estimate reaches a steady state. Note that $\hat{\mathbf{f}}_l$ corresponds to the least-squares approximation of $\mathbf{f}$ in the subspace defined by the feature vectors of the tracked landmark T.

Given a set of position estimates from the set of observed landmarks in an image, a final position estimate is obtained by first detecting and removing outliers using a median filter, and then finding the mean of the remaining estimates. Section 4 will demonstrate the effectiveness of the method based on experimental results.

4 Experimental Data and Discussion

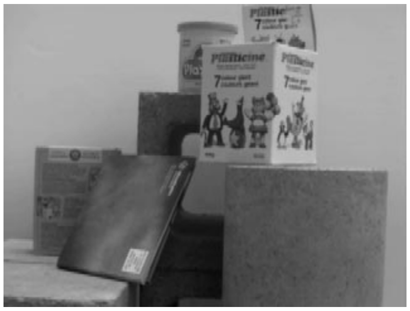

Figure 3: The test environment.

[2] We can measure *a priori* how well this assumption applies to a particular tracked landmark by cross-validating the localization method on each candidate in the tracked landmark.

[3] In practice, this initial value may be an *a priori* estimate.

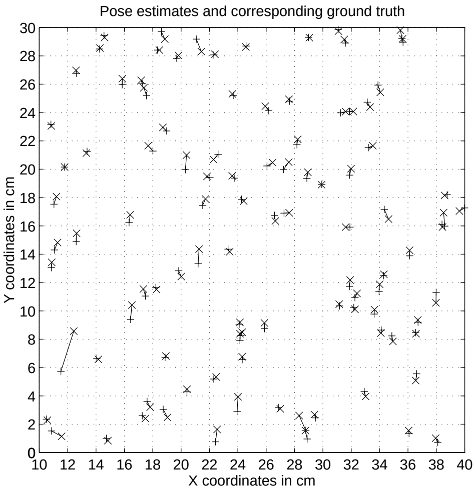

Figure 4: Position estimates and corresponding ground truth for one hundred random samples. Each 'x' marks an estimate as obtained from a single landmark set. The corresponding '+' marks the actual position from which the image was taken. Grid crossings mark the locations of the training images.

In experimental trials, the landmarks selected and tracked by our procedure seem very effective for localization. In this paper, we present results from data acquired by using a camera mounted on a gantry robot for which ground truth positioning can be measured at an accuracy to one tenth of one millimetre. The camera is directed towards a simple constructed scene (Figure 3), which is positioned about 1m from the camera and training images are collected in a 30cm by 30cm grid at 2cm intervals.[4] In addition, one hundred test images are collected, taken from random positions.

Figure 4 is a plot of the mean position estimates (after median filtering) for all of the test images. Each '+' represents the actual position of a test image while the corresponding 'x' marks the position estimate. In this particular case, the average deviation from the correct position is measured to be 3.8mm., less than 20% of the grid sampling density.

In a second experiment, we sample the environment depicted in Figure 1 over a 1.2m by 3.0m configuration-space at 20 cm intervals, using a camera mounted on a mobile robot. In this particular experiment, the ground truth is estimated only by rough dead reckoning (accurate to about 5cm), and at times the robot is not perfectly aligned with the grid axes. In spite of these difficulties, the localization method demonstrates accuracy to 6.8cm in ten trials, suggesting that the method scales reasonably well in indoor environments.

5 Conclusion

In this paper we have described a new technique for position estimation using visual data. Rather than attempting to construct and use a *generic landmark*, we have developed a *generic landmark generation* framework. By using *learned* landmarks, we believe the technique can be used in a much broader range of environments than standard localization methods. Our current work involves experimentally validating this claim. The approach we have taken here is based on learning domain-specific landmarks using a subspace projection method based on principal components analysis. Position estimation involves selecting potential landmarks in an image using a model of visual attention which is based on maxima of the edgel density distribution.

During the online position estimation phase, landmarks are matched to known tracked landmarks based on a subspace encoding. Finally, local variations in the appearance of the landmarks themselves allow a position estimate to be computed. Our implementation computes a position estimate for each landmark that is matched to a tracked set by employing a "fill-in-the-blanks" least squares interpolation.

Experimental testing has demonstrated the validity of our approach. The technique produces an unambiguous position estimate using real data. The use of discrete landmarks generated by an encoding of the landmark sub-images makes our method potentially robust against isolated changes in the environment. In addition, it allows for the post-processing of selected landmarks to apply additional criteria.

Finally, the fact that the landmark representations are learned suggests that the technique can be applied to a wide range of different environmments, as illustrated by figures 1, and 3.

References

[1] K. Sugihara, "Some location problems for robot navigation using a single camera", *Computer Vision, Graphics, and Image Processing*, vol. 42, pp. 112–129, 1988.

[4] In this experiment, motion in the x coordinate corresponds to a sideways translation of the robot, while motion in the y coordinate corresponds to front-to-back motion.

[2] D. Avis and H. Imai, "Locating a robot with angle measurements", *Journal of Symbolic Computation*, , no. 10, pp. 311–326, 1990.

[3] Karen T. Sutherland and William B. Thompson, "Pursuing projections: Keeping a robot on path", in *Proc. IEE Conference on Robotics and Automation*, San Diego, CA, May 1994, pp. 3355–3361, IEEE Computer Society Press.

[4] D.L. Boley, E.S. Steinmetz, and K.T. Sutherland, "Robot localization from landmarks using recursive total least squares", in *Proc Conf. Robotics and Automation, 1996*, Minneapolis, April 1996, IEEE.

[5] M. Drumheller, "Mobile robot localization using sonar", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 2, pp. 325–331, February 1987.

[6] J. J. Leonard and H. F. Durrant-Whyte, "Mobile robot localization by tracking geometric beacons", *IEEE Transactions on Robotics and Automation*, vol. 7, no. 3, pp. 376–382, 1991.

[7] Lindsay Kleeman and Roman Kuc, "An optimal sonar array for target localization and classification", in *Proc. Intl Conf. Robotics and Automation*, San Diego, May 1994, pp. 3130–3135, IEEE Press.

[8] R. Chatila and J. Laumond, "Position referencing and consistent world modelling for mobile robots", in *IEEE International Conference on Robotics and Automation*, 1985, pp. 138–170.

[9] Akio Kosaka, Min Meng, and A. C. Kak, "Vision-guided mobile robot navigation using retroactive updating of position uncertainty", in *Proceedings of the International Conference of Robotics and Automation*, Atlanta, GA, May 1993, vol. 2, pp. 1–7, IEEE Computer Society Press.

[10] C. Lin and R. Tummala, "Mobile robot navigation using artificial landmarks", *Journal of Robotic Systems*, vol. 14, no. 2, pp. 93–106, 1997.

[11] Eric Krotkov, "Mobile robot localization using a single image", in *Proceedings 1989 IEEE International Conference on Robotics and Automation*, pp. 978–983. 1989.

[12] J. R. Beveridge, R. Weiss, and E. M. Riseman, "Combinatorial optimization applied to variable scale 2d model matching", in *Proceedings of the 10th International Conference on Pattern Recognition*, June 1990, pp. 18–23.

[13] R. Basri and E. Rivlin, "Localization and homing using combinations of model views", *Artificial Intelligence*, vol. 78, no. 1-2, pp. 327–354, October 1995.

[14] S.K. Nayar, H. Murase, and S.A. Nene, "Learning, positioning, and tracking visual appearance", in *Proc. IEEE Conf on Robotics and Automation*, San Diego, CA, May 1994, pp. 3237–3246.

[15] Sebastian Thrun, "Finding landmarks for mobile robot navigation", in *Proc. IEEE Robotics and Automation*, Leuven, Belgium, May 1998, pp. 958–963.

[16] G. Dudek and C. Zhang, "Vision-based robot localization without explicit object models", in *Proc. Int. Conf. on Robotics and Automation*, 1996.

[17] I. Zoghlami, O. Faugeras, and R. Deriche, "Using geometric corners to build a 2d mosaic from a set of images", in *Proc. Computer Vision and Pattern Recognition*, San Juan, PR, June 1997, pp. 420–425, IEEE Computer Society Press.

[18] David Noton and Lawrence Stark, "Eye movements and visual perception", *Scientific American*, vol. 224, no. 6, pp. 33–43, June 1971.

[19] Steven W. Zucker, Chantal David, Allan Dobbins, and Lee Iverson, "The organization of curve detection: Coarse tangent fields and fine spline coverings", in *Proceedings of the 2nd Interlnational Conf. on Computer Vision*, Tarpon Springs, Fla., Dec. 1988, pp. 568–577, IEEE.

[20] J.H. Elder and S. W. Zucker, "Computing contour closure", in *Proc. 4th European Conference on Computer Vision*, Cambridge, UK, 1996, vol. 2, pp. 399–412.

[21] A. A. Farag and E. J. Delp, "Edge linking by sequential search", *Pattern Recognition*, vol. 28, no. 5, pp. 611–633, May 1995.

[22] Eric Bourque, Gregory Dudek, and Philippe Ciaravola, "Robotic sightseeing - a method for automatically creating virtual environments", in *Proc IEEE Conference on Robotics and Automation*, Leuven, Belgium, May 1998.

[23] Matthew Turk and Alex Pentland, "Face processing: Models for recognition", *Mobile Robotics IV*, Nov. 1989.

[24] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition", in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, June 1994, pp. 84–90, IEEE Press.

[25] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, July 1997.