Comparing image-based localization methods

Robert Sim and Gregory Dudek
Centre for Intelligent Machines, McGill University
{simra,dudek}@cim.mcgilLca

Abstract

This paper compares alternative approaches to pose estimation using visual cues from the environment. We examine approaches that derive pose estimates from global image properties, such as *principal components analysis* (PCA) versus from local image properties, commonly referred to as *landmarks.* We also consider the failure-modes of the different methods. Our work is validated with experimental results.

1 Introduction

A considerable amount of research has been conducted on the problem of using vision to localize a robot in a known environment. Two basic approaches have emerged; the first correlates global properties of the image, such as a principal components subspace projection; whereas the second correlates a set of local properties, or landmarks. Each approach has its own strengths and weaknesses, and is based on its own assumptions about the environment. It is often argued that local feature-based representations are more robust to scene dynamics and illumination variation than globally-derived representations. It is also widely assumed that feature-based approaches require less online computation time than global approaches.

The goal of this paper is to critically examine a selection of image-based pose estimation methods and compare them for their performance under a variety of conditions. We consider accuracy and running time as the two major indicators of performance.

2 Previous Work

Pose estimation from monocular vision data is non-intrusive and conceptually appealing as a research direction. Position estimation entails a combination of estimating local displacement with recognizing familiar locations. It has been approached as both a coarse place-recognition problem (e.g. based on color histograms for familiar locations [Ulrich and Nourbakhsh, 2000]) and also as a problem of recognizing a set of local features associated with a specific location (for example [Se *et al.,* 2002]). In several approaches, authors have suggested interpolating between familiar locations to produce a position estimate more accurate than the density of training data, using either global image features ([Nayar *et al.,* 1994; Pourraz and Crowley, 1999]) or sets of local image features ([Sim and Dudek, 2001]). In this paper we compare the performance of some of these alternatives..

3 Localization Methods

Framework We assume the same framework for all of the methods presented below. First, a set of training images has been collected from known poses in the environment. A model, or models, are constructed by building an interpolant over the set of training poses. We employ bilinear interpolation of the observation space over the Delaunay triangulation of the pose space. The cost of constructing the triangulation and any other preprocessing of the training images (such as tracking features) is referred to as the *offline cost.*

When a pose estimate is required, a probability distribution is computed using a multi-resolution grid-based discretization of the pose space. At each grid position a predicted observation is generated and compared with the actual observation. The probability is computed using Bayes' Rule, whereby, assuming a uniform prior, the probability of a pose q is proportional to the probability of the observation z given the pose:

$$p\{q|z\} \propto p\{z|q\}$$

Once a cell with maximal probability is determined, a higher resolution discretization is computed in the neighborhood of that cell, and the process recurses until a desired level of resolution is achieved. By setting the maximum discretization to a fixed value, we fix the number of times the likelihood function is evaluated, thus allowing us to benchmark the performance of the method. This running time is referred to as the *online cost.*

3.1 Global Methods

Edge Density Out first globally-based method computes a map of the local edge density of the image. The density is approximated by convolving the image edge map with a wide (33 pixels) Gaussian kernel. The edge map is relatively resistant to illumination changes, and measuring edge density propagates edge information locally. Depending on the width of the kernel, a linear combination of neighboring training edge densities can loosely approximate the motion of edges as the camera moves.

Principal Components PCA operates by computing a low-dimensional subspace of the space spanned by the training images. For this work, we employ the first twenty principal components. When pose likelihoods are computed online, only the low-dimensional projections of the training inputs are interpolated. The input observation must be projected into the subspace; an operation that is performed once; and subsequent comparisons are between low-dimensional vectors.

3.2 Local Methods

The alternative to computing global image features is to extract a set of local image features, or landmarks, and model their behavior as a function of position. Local features must be *selected* and *tracked,* and subsequently they are *modeled.* Finally, when online they must be *matched* to features in the input observation. Once matching is complete, the individual estimates from different landmarks must be combined in a robust manner.

We use the landmark-based method proposed by Sim and Dudek [Sim and Dudek, 2001]. That model selects landmarks using saliency, and, once they are tracked across the training space, models the landmark behaviors as a function of pose. The models compute a linear interpolant of the landmark attributes over the triangulation of the pose space.

We employ two versions of the landmark model; the first, referred to as DynamicLM models the appearance of the landmarks as a function of pose, and thus matching may require several instantiations of the landmark before it is successfully matched. The second, referred to as FixedLM, fixes the landmark appearance, speeding up the matching, but reducing the range of poses that the landmark can model.

4 Experimental Approach

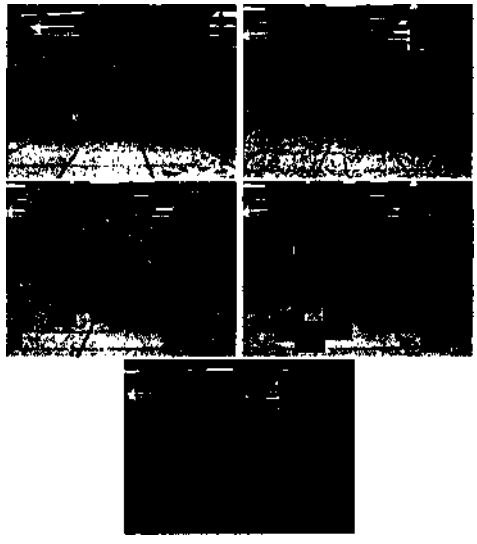

Figure 1: Images from a) the training set (top,left), b) *theNoisy* verification set (top,right), c) the *Occlusion* set with noisy occluders (middle,left), d) the *Occlusion* set with solid occluders (middle,right), and e) the *Illumination* set (bottom).

	Mean Err. (cm)	Max Err. (cm)	Tot. Train'g Time(s)	Online Preproc. (s/img)	Online Loc'zn (s/img)
Global Methods					
Edge Density	8.22	22.4	31.4	0.309	38.1
NN (EdgeDens)	11.1	22.3	n/a	n/a	n/a
PCA	6.06	12.4	2170	0.234	0.208
NN (PCA)	9.7	22.3	n/a	n/a	n/a
Local Methods					
DynamicLM	8.49	23.09	1581	145	19.8
FixedLM	7.61	19.7	1529	10.6	17.9

Table 1: Localization results for *Normal* set.

Training We used a robot to collect 121 images in a grid pattern at 20cm intervals, over a 2m by 2m area (Figure Ia)). Ground truth was obtained by mounting a laser pointer on the robot and measuring by hand the position of the laser point on the floor. As such, the ground truth is accurate to ~0.5cm.

The same set of training images was provided to each localization method for preprocessing and training, and the running times for this phase were recorded.

Verification A set of 29 verification images were collected with the robot from random locations at the same orientation as the training images. Ground truth was measured by hand. The images were collected under the same illumination conditions and observed the same static scene. These images constitute our *Normal* verification set. Gaussian white noise was added to these images to create a *Noisy* verification set (Figure Ia)), and a set of occluders were randomly painted into the images to generate two *Occlusion* sets (Figure Ib) and Ic)). The first set of occluders consists of white noise, and the second consists of solid tiles.

A second set of ten images was collected under low-light conditions from random locations. Ground truth was recorded by hand. These images are referred to as the *Illumination* verification set (Figure Id)).

Each localization method was applied to each verification set, and the mean error between the maximum-likelihood (ML) estimate and ground truth was compiled on a per-method, per-set basis. The computational cost of preprocessing the input images and computing the ML estimate was also recorded. For each image, the multi-resolution grids were evaluated at a total of 2300 distinct poses.

Measuring performance Consider the expected error using a method that selects the nearest training image with 100% accuracy. This *magic number* can be computed to be 7.6cm for our experiments. A mean error of less than 7.6cm could be considered to be successful at interpolating over the training set. In practice a nearest-neighbor implementation is unlikely to be perfect, and will generate larger errors. We have applied a nearest neighbor (NN) approach to the two global methods on the *Normal* set.

5 Experimental Results

Table 1 depicts the results for the *Normal* set. The mean localization error, maximum outlier, and offline and online run-

	Mean (cm)	Max (cm)	Estimates
Global Methods			
Edge Density	8.4	21.9	29/29
PCA	5.9	11.9	29/29
Local Methods			
DynamicLM	11.8	26.4	29/29
1 FixedLM	11.0	27.9	6/29

Table 2: Localization results for *Noisy* set.

	Mean (cm)	"Max" (cm)	Estimates
Global Methods			
Edge Density	60.3	215	29/29
PCA	11.2	21.7	29/29
Local Methods			
DynamicLM	22.0	205	29/29
1 FixcdLM	27.3	180	28/29

Table 3: Localization results for noisy *Occlusion* set.

ning times are depicted. For the two global methods, nearest neighbor results are also tabulated for comparison. Only the PCA and FixedLM methods approach the magic number. The nearest neighbor approaches perform poorly against the magic number. As expected, PCA presents a large offline computational cost, but very low online cost.

Table 2 depicts the results for the *Noisy* set. In the case of the landmark-based approach, we have indicated error results for only those images where at least one landmark was successfully matched. The number of estimates is indicated in the last column, out of the total number of verification images. Both global methods saw an improvement in performance.

Table 3 depicts the results for the *Occlusion* set with noisy occluders. The local methods perform significantly worse, despite improved matching output in the FixedLM case. PCA continues to perform well, but the edge density approach is confounded by the introduction of regions of high edge density.

Table 4 depicts the results for the *Occlusion* set with solid occluders. The local methods demonstrate a degradation in performance, although not as great as for the white noise occluders. There is also a reversal in performance for PCA versus Edge Density.

Table 5 indicates the results under altered illumination conditions. While all of the methods perform poorly, the Edge Density approach suffers the least degradation.

	Mean (cm)	"Max (cm)	Estimates
Global Methods			
Edge Density	10.5	30.8	29/29
PCA	96.3	222	29/29
Local Methods			
DynamicLM	12.1	48.6	29/29
1 FixedLM	15.6	100	28/29

Table 4: Localization results for solid *Occlusion* set.

	Mean (cm)	Max (cm)	Estimates
Global Methods			
Edge Density	30.9	103	10/10
PCA	149	217	10/10
Local Methods			
DynamicLM	105	204	10/10
| FixedLM	138	271	9/10

Table 5: Localization results for *Illumination* set.

6 Discussion and Conclusions

Both PCA and the landmark-based methods present significant offline costs, calling into question their practicality for applications such as simultaneous localization and mapping. However, PCA was robust to some types of adverse imaging conditions, and was demonstrably faster in computing online pose estimates.

While the Edge Density method was less accurate than PCA, it took significantly less time to train, at the cost of slower online performance. However, the ongoing increase in CPU speeds makes global methods such as these increasingly practical.

Finally, the landmark-based methods presented a smooth degradation in performance as the imaging conditions became increasingly adverse. In general, these methods performed well, but were significantly more expensive online due to the cost of matching and frequent invocation of the landmark prediction model. Online performance of these approaches depends to a large extent on the nature of the prediction model itself.

We have presented a comparison of several localization methods in a known environment against a variety of imaging conditions. The results challenge some widely held assumptions about the various advantages and disadvantages of local and global methods.

References

[Nayar et al., 1994] S.K. Nayar, H. Murasc, and S.A. Nene. Learning, positioning, and tracking visual appearance. In *Proc. IEEE Confon Robotics and Automation,* pages 3237-3246, San Diego, CA, May 1994.

[Pourraz and Crowley, 1999] F. Pourraz and J. L. Crowley. Continuity properties of the appearance manifold for mobile robot position estimation. In *Proceedings of the 2nd IEEE Workshop on Perception for Mobile Agents,* Ft. Collins, CO, June 1999. IEEE Press.

[Se *et al,* 2002] S. Se, D. Lowe, and J. Little. Global localization using distinctive visual features. In *Proc. Int. Conf on Intelligent Robots and Systems (IROS),* pages 226 231, Lausanne, Switzerland, October 2002.

[Sim and Dudek, 2001] R. Sim and G. Dudek. Learning generative models of scene features. In *Proc. IEEE Conf Computer Vision and Pattern Recognition (CVPR),* Lihue, HI, December 2001. IEEE Press.

[Ulrich and Nourbakhsh, 2000] Iwan Ulrich and Illah Nourbakhsh. Appearance-based place recognition for topological localization. In *Proc. IEEE Intl. Conf on Robotics and Automation,* pages 1023-1029. IEEE Press, April 2000.