Learning Generative Models of Scene Features

Robert Sim and Gregory Dudek
{simra,dudek}@cim.mcgill.ca
Centre for Intelligent Machines
McGill University 3480 University St., Montreal, Canada H3A 2A7

Abstract

We present a method for learning a set of generative models which are suitable for representing variations of selected image-domain features of the scene as a function of changes in the camera viewpoint. Such models are important for robotic tasks, such as probabilistic position estimation (i.e. localization), as well as visualization. Our approach entails the selection of image-domain features, as well as the synthesis of models of their visual behavior. The model we propose is capable of generating maximum likelihood views of automatically selected features, as well as a measure of the likelihood of a particular view from a particular camera position. Training the models involves regularizing observations of the features from known camera locations. The uncertainty of the model is evaluated using cross validation. The features themselves are initially selected automatically as salient points by a measure of visual attention, and are tracked across multiple views. While the motivation for this work is for robot localization, the results have implications for image interpolation, virtual scene reconstruction and object recognition. This paper presents a formulation of the problem and illustrative experimental results.

1 Introduction

This paper describes a technique for learning generative models of image-domain features of an environment, and then using them for camera position estimation in a Bayesian framework. The models capture not only projective geometry, but also appearance variation due to perspective and illumination phenomena. We also measure our confidence in each model so as to deliver likelihood estimates of future observations. Our goal is to employ these models for a variety of visualization and robotics tasks. In this paper we consider the task of robot localization.

For many robotic tasks an important problem is that of evaluating the likelihood of an observation $\mathbf{z}$ of the environment given some piece of relevant information $\mathbf{q}$, such as the location, or pose, of the camera, or a particular object model hypothesis. The likelihood function $p(\mathbf{z}|\mathbf{q})$ is useful for the task of *Bayesian Inference*, which allows for the computation of the maximum likelihood location or model $\mathbf{q}^*$:

$$p(\mathbf{q}|\mathbf{z}) = \frac{p(\mathbf{z}|\mathbf{q})p(\mathbf{q})}{p(\mathbf{z})} \quad (1)$$

$$\mathbf{q}^* = \arg\max_{\mathbf{q}} p(\mathbf{q}|\mathbf{z}) \quad (2)$$

As a very simple example, Figures 1 a) and c) depict images from a laboratory environment from known poses $q_0 = 0$ and $q_1 = 1$. Given the image in Figure 1 b), taken from an unknown pose q which lies somewhere between q_0 and q_1, the task of localization is to find a q^* which maximizes the likelihood of the image according to Equation 2.

Rather than computing the likelihood of the entire image, which is a computationally complex problem, this paper addresses the problem of learning generative models of local image features that can be used to compute the likelihood of observations of these features from a particular pose. This is accomplished for any given feature f by computing the maximum likelihood observation $\mathbf{z}^*$ given the pose of the camera $\mathbf{q}$, and employing an associated model uncertainty to compute the likelihood function $p(\mathbf{z}|\mathbf{q})$ based on $||\mathbf{z}-\mathbf{z}^*||$. Due to the generative nature of the model, we effectively produce virtual observations of scene features from novel views. As such, our approach also has useful implications for image interpolation and scene reconstruction.

Our approach operates by automatically selecting potentially useful features $\{f_i\}$ from a set of training images of the scene taken from a variety of camera poses (i.e. samples of the *configuration space* of the sensor). The features are selected from each image at each position on the basis of the output of a visual attention operator and are tracked over the training images. This results in a set of observations of many features from different positions. For a given feature f, the reconstruction task then becomes one of learning the imaging function $F_f(\cdot)$, parameterized by camera pose, that

0-7695-1272-0/01 $10.00 © 2001 IEEE

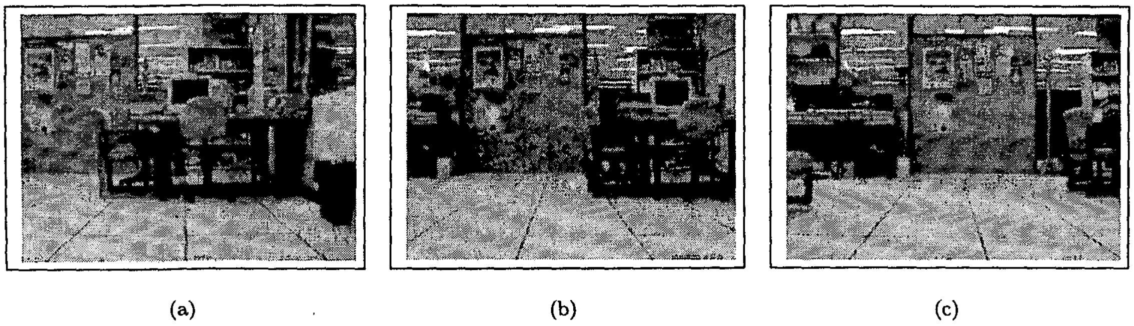

Figure 1: Laboratory Scene: a) known pose q=0, b) q unknown, c) known pose q=1

gives rise to the imaged observation $\mathbf{z}^*$ of f:

$$\mathbf{z}^* = F_f(\mathbf{q}) \tag{3}$$

Clearly, the imaging function is also dependent on scene geometry, lighting conditions and camera parameters, which are difficult and costly to recover[12]. Traditional approaches to the problem of inferring $F_f(\cdot)$ have either focussed on recovering properties of the feature under strict surface or illumination constraints (c.f. [1]), or developed implicit appearance-based representations (e.g. principal components analysis) which often fail to account for the effects of geometry, and hence lead to blurred interpolations between views. Our work addresses the problems inherent in appearance-based representations by capturing feature geometry implicitly. That is, both the appearance and geometric attributes of the feature are captured in a single regularization framework. We accomplish this by representing geometry in the space of affine transformations of the image in the neighborhood of the feature. The best-fit transformation parameters are clearly dependent on the camera position, and can be applied as a precursor to developing an appearance-based representation, which is better suited to representing variation due to radiosity and illumination conditions. Furthermore, the application of an attention operator allows one to focus on the local behaviors of features, which may be easier to model than global properties. In addition, an attention-based approach provides some robustness to complications such as scene dynamics and sensor occlusion.

In the next section we consider prior work on the problem of vision-based robot localization.

2 Prior Work

Our work is motivated by a need to address the task of probabilistic robot mapping, localization and navigation using a vision sensor. Prior work on this task has been successful using sonar and other range-sensing modalities. Recent work by Pourraz and Crowley, as well as Nayar *et al.* have examined an appearance-based model of the environment and perform localization by interpolation in the manifold of principal components[8, 5]. In other work, Dellaert *et al.* have demonstrated the feasibility of employing a vision sensor in the Markov framework[2]. However, the model of the environment is reduced to a simple overhead planar mosaic, and the sensor model is reduced to a single intensity measurement at each camera location. While these approaches demonstrate the utility of appearance-based modeling, they suffer due to the dependency of the result on global sensor information. Furthermore, it is not clear that a strict PCA-based representation can scale for larger environments.

Recent works by Lowe, by Jugessur and Dudek and by Schmid in the problem domains of object and of place recognition demonstrate that object descriptions are captured well by local pseudo-invariants[4, 3, 10]. An attention-style mechanism is employed to extract a set of local object features, and the features are matched against previously learned features for each object class. The benefits of local representations include robustness to partial occlusion and sensor noise. An important aspect of both works is the task of recognizing invariants under changes in viewing conditions. In particular, the attention operators developed are respectively robust to changes in scale and planar rotation. For the localization problem, it is not only important to be able to recognize pseudo-invariants, but to be able to parameterize the effects changes in pose have on the feature. While our current work considers only translation invariance, these prior works indicate the feasibility of including other parameterizations.

Our own prior work has demonstrated the utility and potential accuracy of feature-based localization[11]. In that work, observed features are projected into the subspace spanned by previous observations of the feature, and the resulting pose estimates are combined in a robust fashion. The drawback to this approach is that, like PCA-based methods, the construction of the feature subspace does not scale well for larger environments, and it entails a challenging parameter estimation subtask to permit the elimination of outliers. More importantly, our prior work did not employ a Bayesian framework, and as such did not model image features but imposed a one-to-one mapping between observations and pose. This paper addresses these issues by reconsidering the problem in the context of a generative model of feature behavior, and presenting experimental results for a larger pose space than we have considered in the past.

In the subsequent sections we describe the feature model, present our learning framework, discuss the application of the model to the tasks of scene reconstruction and robot localization, and present experimental results.

3 The Generative Feature Model

We are interested in learning a model of a scene feature, given a set of observations of the feature from known camera positions. The model will be capable of producing maximum-likelihood virtual observations (predictions) of the feature from previously unvisited poses. It will also be capable of estimating the likelihood of a new observation, given the (possibly hypothetical) pose from which it was observed.

We will represent an observation of a feature f by the vector $\mathbf{z} = [\mathbf{t}^T\ \mathbf{i}^T]^T$, where $\mathbf{t}$ represents the parameters that specify an affine transformation of the image sub-window of f to achieve an optimal fit to a representative template of f, and $\mathbf{i}$ represents the local image of f after $\mathbf{t}^{-1}$ has been applied. In this paper, we consider only the translation of the feature in the image plane as the space of possible transformations- a more complete approach would be to also consider rotation and scaling, but we will defer this issue to future work. The observation $\mathbf{z}$ is a vector-valued function of the pose of the camera $\mathbf{q}$. We seek to learn an approximation $F_f(\cdot)$ of this function, as expressed in Equation 3.

The approach we take to learning $F_f(\cdot)$ is by modeling each element of $\mathbf{z}$ as a linear combination of radial basis functions (RBFs), each of which is centered at a camera locus of the observations.

Formally, each scalar element z_i of $\mathbf{z}$ is expressed as

$$z_i(\mathbf{q}) = \sum_j w_j G(\mathbf{q}, \mathbf{q}_j) \tag{4}$$

where $G(\cdot,\cdot)$ is an exponential function centered at the locus $\mathbf{q}_j$ of observation j,

$$G(\mathbf{q}, \mathbf{q}_j) = \exp(-\frac{||\mathbf{q} - \mathbf{q}_j||^2}{2\sigma^2}) \tag{5}$$

and the w_j's are weights that are learned from the training observations. The variance σ in Equation 5 is selected on the basis of the number of observations n and the maximal distance d in between observations in pose space:

$$\sigma = \frac{d}{\sqrt{2n}} \tag{6}$$

The computation of the weights w_j is well understood in the context of regularization and interpolation theory and is described elsewhere[14, 7]. The computational cost for n observations is that of an $O(n^3)$ singular values decomposition of an n by n matrix, followed by an $O(n)$ back-substitution for each z_i. The selection of σ according to Equation 6 induces an exact interpolation of the observations. In practice, however, it is desirable that the feature model be capable of extrapolating beyond the set of observed poses- increasing σ can accomplish that at the minor expense of a smoother interpolating function.

Figure 2 depicts three generated instances of the same feature from different poses. The predicted feature image $\mathbf{i}$ is plotted at the predicted image location $\mathbf{t}$. Note the variation in both appearance and position of the feature in the image.

3.1 Visibility

In addition to modeling the appearance and relative geometry of appearance-domain features, it is also valuable to model their *visibility*. That is, whether or not a particular feature is visible from a particular location in pose-space is informative for the task of localization and important for the problem of reconstructing the scene. We employ the same regularization framework to learn a visibility likelihood function $p(visible(f)|\mathbf{q})$, training the function with the binary observability of each feature from each visited pose in the training set[1]. This information is also useful for informing the question of where to collect new training examples.

3.2 Model Uncertainty

Given the predicted maximum likelihood observation $\mathbf{z}^*$ of a feature f, one might compute the likelihood of a new observation $\mathbf{z}$ on the basis of $||\mathbf{z} - \mathbf{z}^*||$. It is not clear, however, how a metric in the space of observations can be consistently defined (recall that an observation is a combination of pixel intensities and

[1] The regularizer can produce likelihood values less than zero or greater than one- we clamp these outputs when they occur

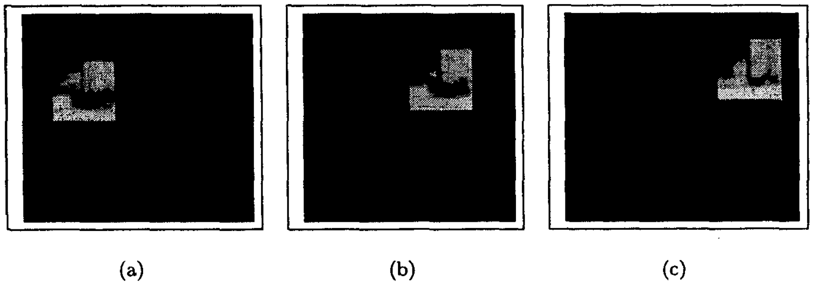

Figure 2: A Feature as generated from three different camera positions. Each feature image has been cropped and enlarged for the purposes of depicting appearance variations. The same cropping parameters were employed in each case, in order to depict the variation in feature position.

transformation parameters). Nor is it clear that the observation space is smooth and/or continuous. Furthermore, how does the likelihood behave as a function of the metric? We model the likelihood function as a Gaussian with a covariance C determined by leave-one-out cross-validation[6].

Given the very high dimensionality of the observation space, however, the covariance C is almost guaranteed to be rank-deficient, which poses problems for numerical stability in the presence of noisy observations. To overcome this problem, we represent the metric $||\mathbf{z}-\mathbf{z}^*||$ as the Euclidean metric of the vector space defined by

$$\mathbf{z}_e = f(\mathbf{z},\mathbf{z}^*) = \left[||\mathbf{i}-\mathbf{i}^*||_2^T, (\mathbf{t}-\mathbf{t}^*)^T\right]^T \qquad (7)$$

where $\mathbf{i}^*$ and $\mathbf{t}^*$ are the intensity and transformation components of the maximum-likelihood prediction, respectively. The likelihood function is then expressed as

$$p(\mathbf{z}|\mathbf{q}) = c\exp(-0.5\mathbf{z}_e^T C^{-1}\mathbf{z}_e) \qquad (8)$$

where $c = ((2\pi)^M \det C)^{-1/2}$, $\mathbf{z}_e$ is the transformed $\mathbf{z}-\mathbf{z}^*$ and M is the dimensionality of the transformed observation space.

The covariance C is not only useful as a model parameter, but is also a useful measure of model fit. Trained features whose model covariance contains large trace values can be eliminated from the set of features on the basis that the feature is not modeled well and will not be useful for feature reconstruction or camera localization.

4 The Learning Framework

In this section we present our approach to collecting and extracting observations of scene features. This process is necessary in order to a) instantiate models in the first place, and b) consider a wide variety of potential features.

4.1 Overview

Our learning approach operates as follows:

1. The robot explores the environment, collecting images from a sampling of positions. It is assumed that a mechanism is available for accurate pose estimation during the exploratory stage (such as a second observing robot[9], or the utilization of an expectation-maximization (EM) approach to map building[13]).
2. A subset of images are selected, and features are extracted from them using a model of saliency.
3. For each extracted feature, a generative feature model is initialized.
4. The generative model is applied in conjunction with the saliency measure to locate a match to each feature in each of the collected images (as described below). As new observations (matches) are found, the generative model is updated. In the interests of mediating computational efficiency and robustness, when the number of observations associated with a particular model exceeds a threshold, the model is split into two separate sets of observations and treated as two different feature models.
5. When the matching is complete, a confidence measure is computed for each feature model, and the models are stored for future use.

Note that while we have presented our approach as a batch computation over the training images, it is sequential in nature and the matching and model updating can be performed in conjunction with the collection of new training images.

In the following sections, we will discuss the details of how features are detected and tracked.

4.2 Feature Detection

As we have described, potential features are initially extracted from a subset of the training images using a model of visual saliency. In this work we employ edge density as our attention operator. The edge map from a given image is convolved with a Gaussian kernel and local maxima in the convolution are selected as salient features. Figure 3 depicts the selected features from an image as superimposed squares over the original, and the convolved edge map.

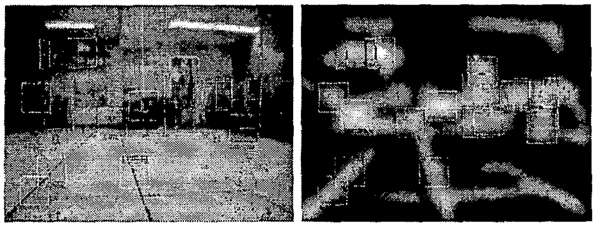

Figure 3: Detected Features in an Image. The original image, and the convolved edge map or density function. The extracted features are marked by squares.

4.3 Feature Matching

Once an initial set of features have been extracted, a generative feature model is initialized for each. The next phase is matching the detected features over the entire training image set. Each training image is searched in sequence for each feature. Given that the camera pose of any given training image is known, we use the generative model of the feature to predict the intensity image $\mathbf{i}_f$ of the feature for the training image being searched. We define the best match to $\mathbf{i}_f$ in the image to be the image sub-window $\mathbf{i}^*$ centered at position (x^*, y^*) that has maximal correlation ρ with the predicted image $\mathbf{i}_f$:

$$\rho = \cos\theta = \frac{\mathbf{i}_{(x,y)} \cdot \mathbf{i}_f}{||\mathbf{i}_{(x,y)}|| \, ||\mathbf{i}_f||} \quad (9)$$

When matching is complete, we have a set of matched features, each of which is comprised of a set of observations from different camera poses. Figure 4 depicts one such set, where each observation is laid out on an overhead view of the pose space; grid locations where there is no observation correspond to locations in the pose space where the feature was not found in the corresponding training image. Note that the generative nature of the matching mechanism allows the appearance of the feature to evolve significantly over the pose space.

5 Applications

5.1 Scene Reconstruction

Given a set of trained features and a particular pose $\mathbf{q}$, one can generate a maximum likelihood reconstruction of the scene features. The generated maximum likelihood observations are each weighted by their likelihood (which varies inversely with the determinant of the feature covariance C) and superimposed onto a blank image. For example, Figure 3 a) shows a training image from a laboratory scene for which training images have been collected at 25cm intervals over a 6.0m by 3.0m pose space; Figure 5 depicts the reconstruction of the same scene from a nearby pose. Note that the reconstruction cannot predict pixels for which there is no feature model, and as such, the lower edge of the image is left unshaded.

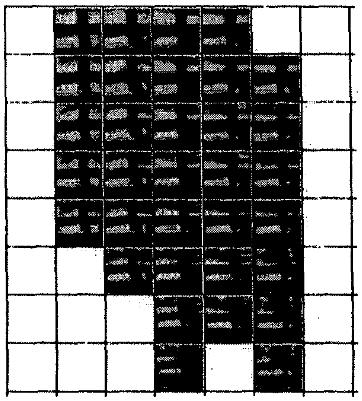

Figure 4: A set of observations of an extracted scene feature. The grid represents an overhead view of the pose space of the camera, and feature observations are placed at the pose corresponding to where they were observed. Note that the observations capture variation in feature appearance.

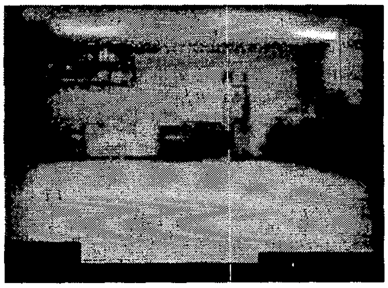

Figure 5: A reconstruction of the laboratory scene depicted in Figure 3, as predicted from a nearby camera pose.

5.2 Localization

Given a set of feature models, the task of robot localization can be performed using Bayesian Inference. When the camera is at an unknown position, an observation is obtained and optimal matches to the learned features are detected in the image, $Z = \{\mathbf{z}_f\}$. Each feature observation $\mathbf{z}_f$ then contributes a probability density function (pdf) $p(\mathbf{z}_f|\mathbf{q})$, which is defined as the product of the pdf due to the maximum likelihood prediction of the model (Equation 8) and the feature visibility likelihood $p(visible(f)|\mathbf{q})$. In the absence of informative priors, the pose $\mathbf{q}^*$ that maximizes the joint likelihood of the observations is considered to be the maximum likelihood position of the robot, as illustrated by Equation 2. Numerically, the joint likelihood can be difficult to compute, as it requires summing over all permutations of successful and unsuccessful feature matches. Instead, we approximate the likelihood as the sum of the individual pdf's:

$$p(Z|\mathbf{q}) \approx \sum_{\mathbf{z}_f \in Z} p(\mathbf{z}_f|\mathbf{q}) \qquad (10)$$

Note that a complete description of the probability density function should take into account the likelihood of the match between each detected feature and all possible generated observations. However, such an approach is not only intractable, but taking an approximation as a joint probability also leads to catastrophic cancellation in the presence of outlier matches. In the following section we present our experimental results for the task of camera localization.

6 Experimental Results

In this experiment, we test the learning framework and feature models on the task of robot localization. The laboratory scene depicted in Figure 3 was explored by taking 291 training images at uniform intervals of about 25cm over a 3.0m by 6.0m pose space. An observing robot was deployed to estimate the ground-truth position of the exploring robot to an accuracy of about 4cm, as described in [9]. For the purposes of this experiment, the robot attempted to take training images at the same global orientation. However, noise in the robot's odometry, as well as the observing robot's estimator, led to some noise in this orientation from pose to pose.

A set of initial features were extracted from a small subset of the training images, and more than 117 feature models were trained. Those models with high uncertainty, or with too few observations were removed, resulting in 80 reliable feature models.

To validate the learned models, a set of 93 images were collected from random poses, constrained to lie anywhere within the 3.0m by 6.0m training space. These test images were used to compute maximum-likelihood estimates of the camera's position, and the maximum likelihood estimates were compared against the "ground truth" estimates provided by the observing robot. The maximum-likelihood (ML) estimates themselves were computed by exhaustive search in the training space, selecting the hypothesized $\mathbf{q}$ that maximized Equation 10. Note that in a production environment, a more efficient estimator, such as Monte Carlo sampling, could be deployed.

In practical settings, one is not always interested in the ML pose estimate, but rather the entire pdf of possible poses. Figure 6 depicts the un-normalized pdf resulting from evaluating Equation 10 for a single test image over a uniform grid of poses. The figure clearly indicates a region where the pose is highly likely, as well as a second, wider less-likely region. The second region is likely due to a mis-classified feature (a failure in the matching stage), or possibly some self-similarity in a trained feature.

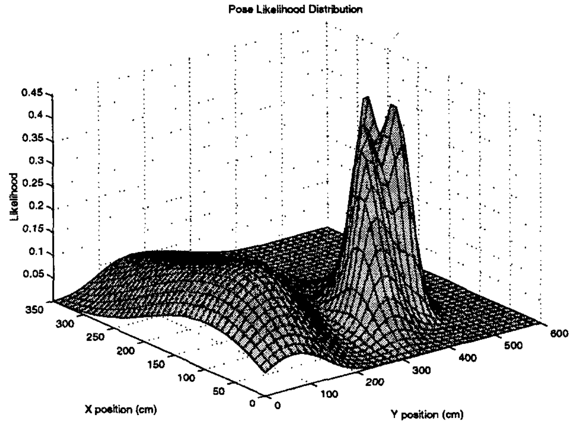

Figure 6: Likelihood function of pose of robot over 3.0m by 6.0m pose space. Note that the distribution is not unimodal, possibly due to a misrecognized feature, or model self-similarity at different poses.

Given that each ML estimate has an associated likelihood, it is possible to reject pose estimates that do not meet a particular likelihood threshold. In this way, four of the 93 estimates were rejected. Interestingly, the majority of these estimates were associated with images that were obtained when the robot was very close to the wall it was facing, where it was difficult to reliably track features at the large training interval (25cm).

Figure 7 plots the location of the unrejected ML estimates for the test images ('x') against the "ground truth" camera position ('o'). The mean absolute error is 17cm, (7.7cm in the x direction vs 15cm in the y direction). The larger error in the y direction corresponds to the fact that the camera was pointed parallel to the positive y axis, and changes in observations due to forward motion are not as pronounced as changes due to side-to-side motion. The smallest absolute error was 0.49cm, which is insignificant compared to the "ground truth" error, and the largest error was 76cm. Note that most of the larger errors occur for large values of y. This is due to the fact that the camera was closest to the wall it was facing at these positions y, and as has been mentioned, tracking scene features over 25cm pose intervals became nearly impossible.

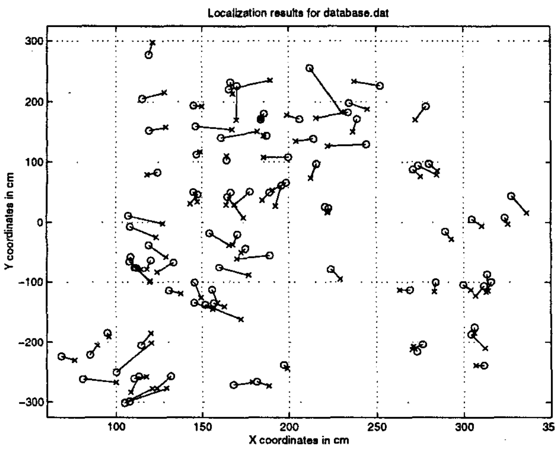

Figure 7: The set of maximum-likelihood pose estimates ('x') plotted against their "ground truth" estimates ('o').

7 Discussion and Conclusions

We have presented a method for learning generative models of visual features. The method operates by matching image features over a set of training images, and learning a generating function parameterized by the pose of the camera which can produce maximum likelihood feature observations. The system also models the uncertainty of the generated features, allowing for Bayesian inference of camera pose.

The experimental results demonstrate the utility of the learned feature models for pose estimation, as well as other tasks, such as scene reconstruction. Our experiments have demonstrated the stability and smoothness of the resulting pdf over camera pose, and we were able to detect most outliers by thresholding the likelihood of the ML estimates. In addition, the scope of our experiments surpass that of our prior works, where only very small regions of the pose space were explored, and the feature models were not suited to larger environments. However, important issues are raised in this work with respect to the density of training samples. In order to capture aspects of the scene that change significantly, one must sample at higher densities. One possible solution is to select the robot's viewing direction before sensing in order to take in more stable parts of the environment (for example, point the camera at the farthest wall). Our future work is addressing some of the issues raised here, as well as expanding the approach to much larger environments.

References

[1] P. Belhumeur and D. Kriegman. What is the set of images of an object under all possible lighting conditions. In *Computer Vision and Pattern Recogition*, page 270, San Francisco, CA, 1996. IEEE Press.

[2] F. Dellaert, W. Burgard, D. Fox, and S. Thrun. Using the condensation algorithm for robust, vision-based mobile robot localization. In *Computer Vision and Pattern Recognition.* IEEE Press, June 1999.

[3] G. Dudek and D. Jugessur. Robust place recognition using local appearance based methods. In *IEEE Conf. on Robotics and Automation*, San Francisco, April 2000. IEEE Press.

[4] D.G. Lowe. Object recognition from local scale-invariant features. In *International Conf. on Computer Vision*, pages 1150–1157, Corfu, Greece, September 1999. IEEE Press.

[5] S.K. Nayar, H. Murase, and S.A. Nene. Learning, positioning, and tracking visual appearance. In *IEEE Conf on Robotics and Automation*, pages 3237–3246, San Diego, CA, May 1994.

[6] R.R. Picard and R.D. Cook. Cross-validation of regression models. *J. of the American Statistical Association*, 79(387):575–583, 1984.

[7] T. Poggio and S. Edelman. A network that learns to recognize 3d objects. *Nature*, 343:263–266, January 1990.

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

[9] I. Rekleitis, G. Dudek, and E. Milios. Multi-robot collaboration for robust exploration. In *Proc. IEEE Conf. on Robotics and Automation*, San Francisco, CA, April 2000.

[10] C. Schmid. A structured probabilistic model for recognition. In *Computer Vision and Pattern Recognition*, pages 485–490, Ft. Collins, CO, June 1999.

[11] R. Sim and G. Dudek. Learning and evaluating visual features for pose estimation. In *IEEE International Conference on Computer Vision*, Kerkyra, Greece, sept 1999. IEEE Press.

[12] G. P. Stein and A. Shashua. Model-based brightness constraints: On direct estimation of structure and motion. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(9), September 2000.

[13] S. Thrun, W. Burgard, and D. Fox. A real-time algorithm for mobile robot mapping with applications to multi-robot and 3d mapping. In *IEEE Conference on Robotics and Automation*, San Francisco, CA, May 2000. IEEE Press.

[14] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems.* V.H. Winston & Sons, John Wiley & Sons, Washington D.C., 1977. Translation editor Fritz John.