Learning Generative Models of Scene Features

Paper 256

Abstract

We present a method for learning a set of generative models which are suitable for synthesizing views of scene features from arbitrary camera positions. The model is capable of generating maximum likelihood views of 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 and navigation, 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 image-domain features of an environment, predicting their subsequent appearance, and using them for robot position estimation. Our approach is an essentially probabilistic one, so that the confidence of these features can also be estimated. In particular, we are interested in doing this over large environments where simple global appearance-based method are inadequate.

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 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 $\mathbf{q}^*$:

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

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

For example, Figure 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, 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, this paper addresses the problem of learning generative models of local image features that can be used to compute the likelihood of an observation of an image feature 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 $p(\mathbf{z}|\mathbf{q})$. 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.

Research in biological vision indicates that the visual system is selective as to what parts of a scene are processed. It is understood that successful application of attention can significantly reduce the computational cost of performing generic visual tasks [2, 16]; in fact it may be a necessity in many contexts. By the same token, it is useful for computational vision systems to exploit the output of an attention operator in order to reduce the computational burdens posed by a particular problem. In the context of scene reconstruction, it is clear that such an approach has benefits not only in reducing computational complexity, but also allows the system to be sensitive to local, as opposed to global, variations in scene properties.

Our approach operates by automatically selecting potentially useful features from a set of training images of the scene taken from a variety of camera poses (i.e. sample 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 gives rise to the imaged observation $\mathbf{z}^*$ of

(a) (b) (c)

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

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 each difficult and costly to recover[15, 6]. Traditional approaches to the problem of inferring $F_f(\cdot)$ have either focussed on recovering the geometric properties of the feature under strict surface or illumination constraints (shape-from-X), or developed implicit appearance-based representations (Principal Components Analysis) which 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, as we have already mentioned, the application of an attention operator allows one to focus on the local behaviours of features, which may be easier to model than global properties.

In the next section we consider prior work on the problem.

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[5, 9]. In more recent work, Dellaert *et al.* have demonstrated the feasibility of employing a vision sensor in the Markov framework[3]. However, the model of the environment is reduced to a simple overhead planar mosaic, and the sensor model is reduced to a single pixel. In other work, Pourraz and Crowley and Nayar *et al.* have each examined an appearance-based model of the environment and perform localization by interpolation in the manifold of principal components[14, 11]. These approaches 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 work by Lowe, and Dudek and Jugessur in the problem domain of object and place recognition demonstrates that object descriptions are captured well by local pseudo-invariants[10, 4]. 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 the work of both Lowe and Dudek and Jugessur 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 prior work has demonstrated the utility of feature-based localization[1]. In that work, observed features are projected into the subspace spanned by previous observations of the 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. This paper addresses these issues by reconsidering the problem in the context of learning generative feature models which are useful for probabilistic localization.

In the subsequent sections we describe the learning framework and present experimental results.

3 Selecting and Tracking Features

The essential primitives of our approach are a set of image subwindows that are matches of previously seen regions modeled during a training phase. These image-domain features define the camera pose by virtue of both their appearance and their position.

Before features can be modelled, they must first be selected and tracked. We take the approach developed in our prior work on the task of visual landmark acquistion, which is summarized in Figure 2. During an initial exploration of the scene, a set of training images are collected from various camera positions (large rectangles). In this work we employ a simple measure of visual attention which operates on the basis of edge density to extract a set of candidate features from each image (small squares). In other work, we have evaluated a variety of attention operators for this task[2].

Our technique for feature tracking operates by matching the extracted features to previously observed *prototypes*. Matching is performed on the basis of local image correlation. If a feature fails to match to any prototype within a reasonable error threshold, it becomes a prototype for a new feature. Whereas there is usually some instability in the attention operator, a local search is performed in the image neighbourhood of the feature in order to find the best fit to the matching prototype.

A *tracked feature* constitutes the set of observations used to train a generative model of the feature. Figure 3 depicts a tracked feature, where each observation

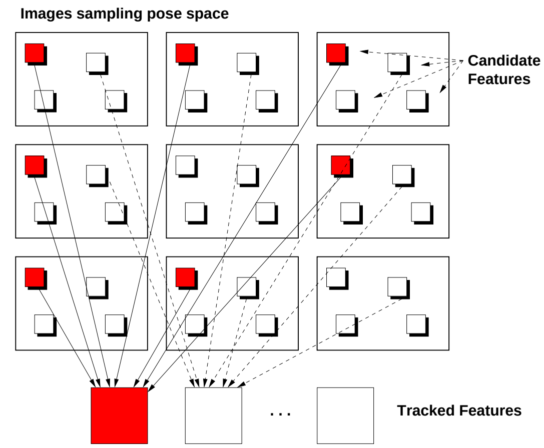


Figure 2: Feature Selection and Tracking.

of the feature is placed at the camera position where it was observed from, where the camera has explored a 2-D pose space with uniform sampling. Empty grid locations indicate that the feature was not detected by the attention operator from that position. Note also that while the feature varies slightly in appearance from pose to pose, the translation parameters of the feature in the image also vary as a function of pose.

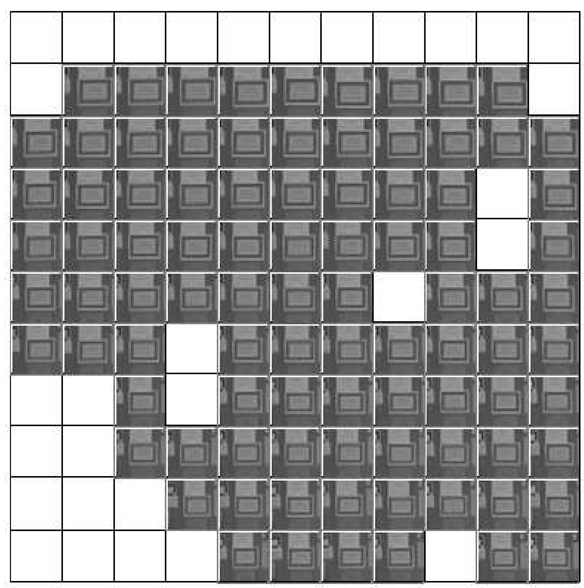


Figure 3: A typical tracked feature.

As we elaborated at the outset, it is desirable to be able to generate a maximum likelihood observation of the feature from an arbitrary pose, as well as a measure of the likelihood of an arbitrary observation. In the following section we describe the modelling mechanism.

4 Modelling Features

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 the affine transformation of the image subwindow of f to achieve the best fit to its matching prototype, and $\mathbf{i}$ represents the local image of f under the inversion of the transformation $\mathbf{t}$. In this paper, we consider only the translation of the feature in the image plane as the space of possible transformations. 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 modelling each element of $\mathbf{z}$ as a linear combination of radial basis functions (RBF), each of which is centered at a camera locus of the observations. Poggio and Edelman have suggested similarities between RBF representations and those maintained in human perception[13].

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

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

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

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

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:

$$\sigma = \frac{d}{\sqrt{2N}} \qquad (6)$$

The computation of the weights w_j is well understood in the context of regularization and interpolation theory and is described elsewhere[13, 7]. The computational cost is that of a singlular values decomposition of an N by N matrix ($O(N^3)$) followed by a backsubstitution for each z_i.

Figure 4 b) depicts a maximum likelihood prediction of a feature from the laboratory environment depicted in Figure 1. The generated feature is due to a previously unexplored pose falling between the observations taken in Figure 4 a) and c).

4.1 Visibility

In addition to modelling 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 point in configuration 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 $p(visible(f)|\mathbf{q})$, setting the function to 1 where an observation is obtained and 0 where it is not observed. This information is also useful for informing the question of where to collect new training examples. This visibility data is linearly interpolated between sample positions.

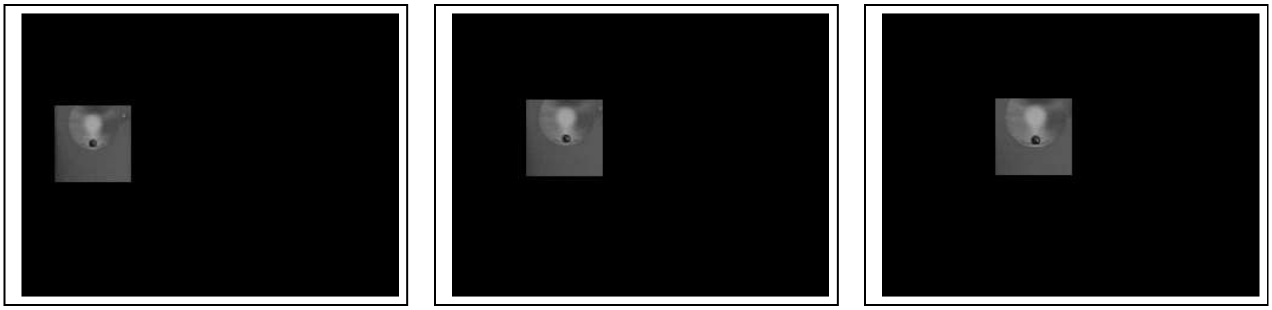
5 Model Uncertainty

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 transformation parameters). 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[12, 8]. Given the high dimensionality of the observation space, we first transform $\mathbf{z} - \mathbf{z}^*$ to ensure numerical stability. The transformation is achieved by combining the differences in the pixel intensities to a single magnitude and leaving the transformation components as-is. The likelihood function is then expressed as

$$p(\mathbf{z}|\mathbf{q}) = c \exp(-0.5\mathbf{z}_e^T C^{-1} \mathbf{z}_e) \qquad (7)$$

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 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 modelled well and will not be useful for feature reconstruction or camera localization.



(a)

(b)

(c)

Figure 4: A Feature: a),c) Prior observations, b) generated feature from a novel view between the two observations

5.1 Partial 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 C) and superimposed onto the image. For example, Figure 5 a) shows an image of a simple scene which is trained over a relatively high sampling density whereas Figure 5 b) depicts the reconstruction of the same scene from the same pose, but where figure a) is not used as a training example.

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 the features detected in the image are matched to the prototypes of the learned features. Each feature observation $\mathbf{z}_f$ then contributes a probability density function (pdf) $p(\mathbf{z}_f|\mathbf{q})$, which takes into account both the maximum likelihood observation and the feature visibility likelihood. The pose $\mathbf{q}^*$ that maximizes the joint likelihood of the observations is considered to be the position of the robot. Numerically, the joint likelihood can be difficult to compute, due to the catastrophic cancellation caused by outliers. Instead, we compute the likelihood as the sum of the individual pdf's:

$$p(\mathbf{Z}|\mathbf{q}) = \sum_{\mathbf{z}_f \in Z} p(\mathbf{z}_f|\mathbf{q}) \tag{8}$$

In theory, a complete description of the probability density function would take into account the likelihood of the match between each detected feature and each prototype. Under this interpretation, taking the sum of the pdf's is reasonable since there is a non-neglible likelihood that any particular match is incorrect. In the following section we present our experimental results for scene reconstruction and camera localization.

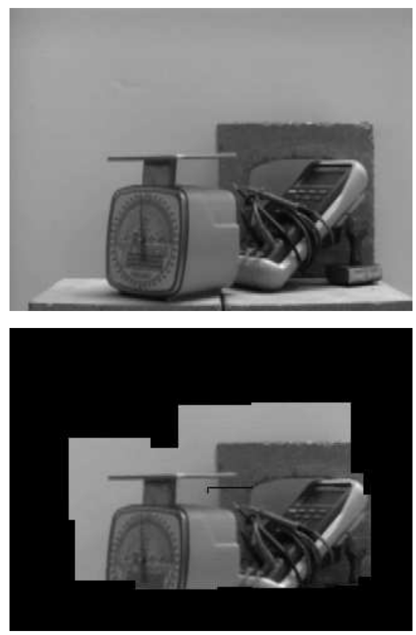


Figure 5: A simple scene Top: Actual Bottom: Synthesized

6 Experimental Results

6.1 Reconstruction Performance

In our first experiment, we evaluate the feature reconstruction performance. We employ the simple scene depicted in Figure 5, where the camera is affixed to the end-effector of a gantry-mounted robot arm. The feature models are trained by collecting training images from known positions at 1cm intervals over a 10cm by 10cm grid. The orientation of the camera is fixed to face forward in a direction parallel to the positive y axis. Images from extreme corners of the grid are depicted in Figure 6. A second set of 20 test images were obtained from random known positions within the boundary of the grid. The camera position for each test image was then used to generate the maximium likelihood observation of the image from the training examples. The mean difference between rendered pixels and the actual observation was 1.3% (3.2 intensity values out of a range of 256) with a standard deviation of 2.2%.

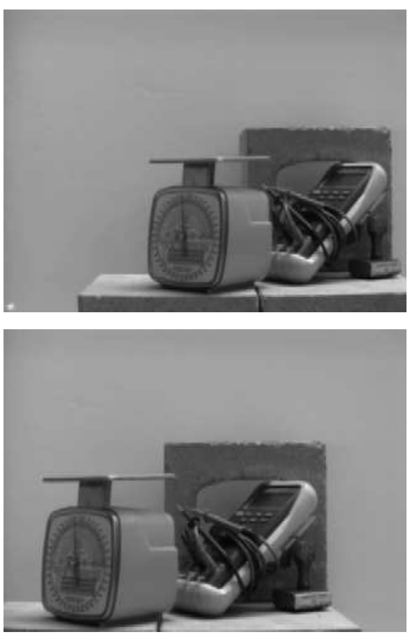


Figure 6: Two images from the simple scene

6.2 Localization Performance

We employ the same scene and training images for the purpose of evaluating the localization performance of the feature models. The twenty test images are provided as input to the system, each of which generates a joint probability density function as a function of pose. Figure 7 depicts the pdf of an image collected from a position close to the center of the training space (coordinates (14.32,6.75)). While the distribution has a clear maximum at (14.45,6.78), the it is clearly not unimodal in nature.

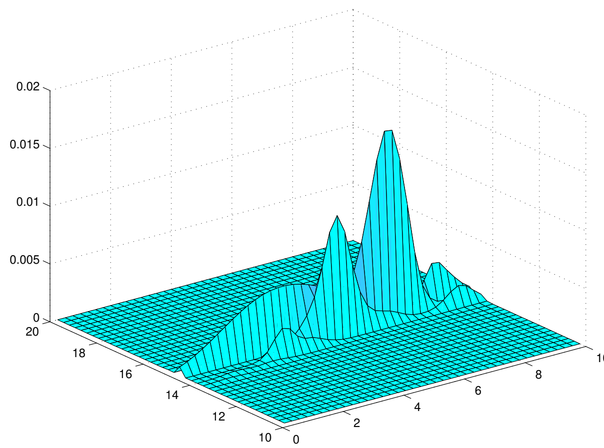


Figure 7: Probability density function of pose of robot over 10cm by 10cm pose space.

Figure 8 plots the location of the maximum of each computed pdf for each test image ('x') versus the true camera position ('o'). The mean absolute error is 0.52cm, and the standard deviation is 0.86cm. Note however, the presence of two significant outliers. These outliers are likely due to a failure in the feature matching. Note also that the mean error in the x coordinate is only 0.09cm versus 0.51cm in the y coordinate. The difference is due to the fact that the model representation (the combination of image domain translation and appearance) is better suited to capturing side to side motion (image translation) versus variations due to forward motion (image scaling).

7 Discussion and Conclusions

The experimental results demonstrate the utility of the approach for camera pose estimation and suggest its utility for image interpolation. Our experiments in a larger environment, such as the one shown in Figure 1 indicate slightly less accurate performance relative to the density of the training samples. These results are due primarily to the question of how to

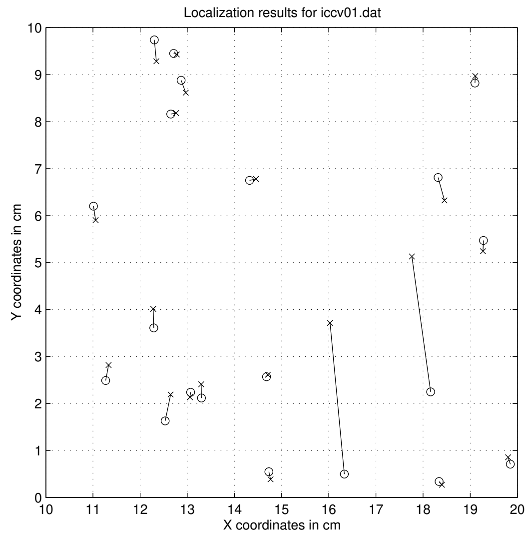


Figure 8: The set of pose estimates obtained for the simple scene shown in Figure 6.

reliably track features that undergo a more significant change in appearance from position to position. We believe that this can be ameliorated by using a more powerful representation of feaure behavior in the matching stage. Outlier feature matches during training can corrupt the behaviour of the feature reconstruction, and outlier matches during pose estimation can corrupt the joint pdf. This situation is slightly complicated by the fact that most state of the art tracking methods assume a one dimensional sequence of images, which may not be relevant to the task of tracking features over two or more dimensional configuration spaces. Despite these issues, we are optimistic that robust tracking can lead to significant improvements in performance.

We have presented a method for learning generative models of visual features. The method operates by tracking 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. We have indicated reliable experimental results for a simple scene, and discuss the issues encountered toward implementation in a practical environment. Our future work will pursue this direction, as well as methods for determining optimal sampling trajectories.

References

[1] The Author. " ". In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Corfu, Greece, September 1999. IEEE, IEEE Press.

[2] The Author. " ". Submission to the 2001 IEEE International Conference on Robotics and Automation, 2000.

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

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

[5] D. Fox, W. Burgard, and S. Thrun. Active markov localization for mobile robots. *Robotics and Autonomous Systems*, 1998. To appear.

[6] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The lumigraph. In *Computer Graphics Proceedings, Annual Conference Series, Proc SIGGRAPH '96*, pages 43–54, New Orleans, August 1996. ACM SIGGRAPH.

[7] S. Haykin. *Neural Networks*. MacMillan College Publishing Company, 1994.

[8] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 1995 International Joint Conference on Artificial Intelligence*, Montréal, August 1995.

[9] John J. Leonard and Hugh F. Durrant-Whyte. Mobile robot localization by tracking geometric beacons. *IEEE Transactions on Robotics and Automation*, 7(3):376–382, June 1991.

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

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

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

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

[14] 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.

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

[16] John K. Tsotsos. The complexity of perceptual search tasks. In *International Joint Conference on Artificial Intelligence*, pages 1571–1577, 1989.