Toward Selecting and Recognizing Natural Landmarks *

Erliang Yeh David J. Kriegman
Center for Systems Science, Department of Electrical Engineering
Yale University, New Haven, CT 06520-8267

Abstract

Landmarks are often used as a basis for mobile robot navigation. In this paper, we consider the problem of automatically selecting from a set of 3D features the subset which is most likely to be recognized from noisy monocular image data and is least likely to be confused with any other group of features. Assuming perspective projection, real valued recognition functions are constructed for a set of features. The value returned from such functions are invariant to changes of viewpoint and can be evaluated directly from image measurements without prior knowledge of the position and orientation of the camera. With image noise, the recognition function no longer evaluates to a constant value. Because of the possibility of false matches, a Bayes detector is used to determine the optimal range of values of the recognition function that will be accepted as image features of the model. The model with the lowest Bayes cost is selected as the most distinguishable landmark. We show implementation results for real 3D objects.

1 Introduction

There have been two approaches to mobile robot navigation in the literature: reconstructionist versus reactive. In the more traditional reconstuctionist approach, sensor information is used to construct a 3D model or map of the robot's environment [4, 5]. Much effort is required to maintain a consistent and accurate representation of the geometry of the world. On the other hand, the reactive paradigm, initially championed by Brooks [2] and adopted by many others, bases robot behavior more directly on immediate sensor data and less on a stored representation. In particular explicit, large scale reconstruction is avoided because as argued by proponents, the world is not static, it is difficult to maintain a consistent representation, and perhaps more importantly, it is unnecessary for most navigation tasks.

In the work of Taylor and Kriegman, algorithms were developed for systematically exploring a bounded 2D configuration space in search of a recognizable object [11]. A prototypical task for an indoor mobile robot operating in an office setting might be fetching output from a printer. Clearly the robot must be able to recognize the printer when it is in sight. In addition to recognizing its goal, the robot takes advantage of objects that it can recognize along the way. As a byproduct, the algorithm constructs a "topological representation" akin to a level of Kuiper's spatial semantic hierarchy [7]. The representation essentially encodes which recognizable objects are visible in the vicinity of a particular object, and this leads to a natural graph structure. A robot can execute a plan, defined by a path through this graph, using a combination of boundary following and "visual servoing" to approach the recognizable object. Exploration is then cast as the process of learning this graph and terminating when the recognizable object has been found. As a byproduct, the learned graph can be used for future navigation tasks. Note that this is not a quantitative reconstruction of the geometric structure of the environment but instead encapsulates the qualitative relationship of recognizable objects. Besides our own work, landmarks have been critical to many other approaches to navigation [6, 8, 9]. Note also that once the landmarks are identified, they can be used to localize the robot [1, 10]

The exploration/navigation algorithm described above has been implemented on our mobile robot [11], and the objective of this work as to show *how* object recognition could be used to solve navigation and exploration problems rather than using reconstruction. The actual problem of object recognition was trivialized by tacking targets on objects (essentially bar codes) which are easily recognized in cluttered scenes. A more compelling alternative to artificial targets would be to store 3D models of those objects that the robot is likely to encounter and employ established techniques to recognize these objects. While prior models are useful for describing the destination, such an approach would be ineffective during the course of navigation when the robot encounters unmodelled objects. Instead, the robot should be able to learn about the new objects that it encounters and retain models of those objects that are useful for the task.

In this paper, we consider the problem of recognizing and learning about perceptually salient objects or landmarks from image data. Thus, a robot would not have to be preprogrammed with CAD-like models of important objects and instead would learn from what it encounters. What the robot uses as a landmark will be driven by the statistical distribution of objects and features that it encounters in the world rather than some prior set of preprogrammed models.

The goal of identifying and later recognizing perceptually distinctive objects (also termed landmarks)

*This work was supported by the Office of Naval Research under grant N00014-93-1-0305 and the National Science Foundation under Grant NYI-IRI-9257990.

can be cast as the following problem: *Given a set of features, select a subset of these features which in a monocular image is most likely to be recognized and least likely to be confused with any other group of features.* Here, we assume that landmarks are selected from a set of indistinguishable viewpoint independent 3D features (e.g. points or lines); that is, they cannot be differentiated by local geometry, color, or texture nor can they be distinguished by adjacency information (e.g. connectivity by edges). If such information were available, it would naturally simplify the resulting combinatorics and improve accuracy.

We take the following approach: From a set of 3D features, a subset of the features becomes a hypothetical landmark model. For this set of features, a recognition function can be constructed which evaluates to zero for any noiseless image of these features. Applying this function to actual image data, a set of features is taken to be an instance of the model when the function evaluates to zero. Because of image noise, it will not evaluate precisely to zero, and a range of values (presumable about zero) must be accepted. Knowing the probability distribution of image measurements, an optimal range can be selected based on a Bayes detector. Furthermore, the probabilities of mistaking some other object as a landmark (false positives) or missing a landmark (false negatives) can be computed. For a set of hypothetical landmarks, the one which minimizes the Bayes cost is selected as the most salient landmark and used for robot navigation.

In this paper, we simplify the problem in the following way: we assume that the mobile robot only travels along a horizontal ground plane, and the only features considered are vertical lines. Taken together, the problem can be reduced to selecting landmarks from point features in the plane. We assume Gaussian image noise, though other models could be employed. We also assume that the 3D features are visible from any viewpoint within a certain distance (i.e., no occlusion). Taken together, these assumptions allow for tractable formulation. Future work will include methods for relaxing some of these assumptions to a richer set of features, more realistic noise models, and using representations like aspect graphs to handle occlusion.

2 Recognition Functions

Recently, Weinshall introduced the notion of "model-based invariants" for object recognition [13]. From a set of m 3D features called the model $\mathcal{M}$, a real valued recognition function $\mathcal{I}(\mathbf{a})$ can be constructed where $\mathbf{a}$ is a vector of the image measurements. The recognition function $\mathcal{I}(\mathbf{a})$ evaluates to zero for an image of the model $\mathcal{M}$ from any viewpoint. Thus, given an image with n features, an algorithm for recognizing $\mathcal{M}$ is to choose all $\binom{n}{m}$ subsets of m features and evaluate $\mathcal{I}(\mathbf{a})$. The subset of m features which minimizes $|\mathcal{I}(\mathbf{a})|$ is considered to be the recognized object.

In this paper, we assume that the robot moves on a horizontal ground plane and that the camera is modeled by perspective projection. Note that for a camera whose optical axis is parallel to the ground plane, the image of vertical 3D lines will also be vertical [5]. Using vertical segments as features, and projecting both the features and the image plane onto the ground plane, the problem can be modelled in two dimensions. The features project to points in the plane, the camera position is given by one orientation and two translation parameters, and the image plane can be considered an image line.

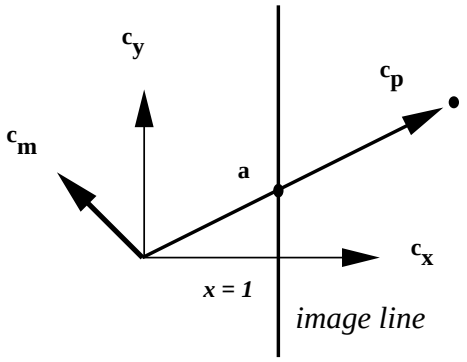


Figure 1: The perspective projection imaging model.

Given the coordinates of a point in the world frame ${}^w\mathbf{p}_i = ({}^wx_i, {}^wy_i)$, the coordinates of the point in the camera frame are given by ${}^c\mathbf{p}_i = {}^c_w\mathbf{R}\ {}^w\mathbf{p}_i + {}^c\mathbf{t}$, i.e.,[1]

$$\begin{bmatrix} {}^cx_i \\ {}^cy_i \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} {}^wx_i \\ {}^wy_i \end{bmatrix} + \begin{bmatrix} {}^ct_x \\ {}^ct_y \end{bmatrix}$$

where ${}^c_w\mathbf{R}$ is a 2D rotation matrix of the world frame relative to the camera frame, and ${}^c\mathbf{t}$ is the translation vector in camera frame.

Assuming a camera with unit focal length, the image line (projection onto the ground plane of the image plane) is located at $x = 1$ in the camera frame. Let a_i be the image measurements of ${}^c\mathbf{p}_i$, then

$$a_i = \frac{{}^cy_i}{{}^cx_i} = \frac{{}^wy_i\ \cos\theta - {}^wx_i\ \sin\theta + {}^ct_y}{{}^wx_i\ \cos\theta + {}^wy_i\ \sin\theta + {}^ct_x}. \qquad (1)$$

As shown in Figure 1, from an image measurement a_i of a 2D point ${}^c\mathbf{p}_i$, we know that ${}^c\mathbf{p}_i$ lies on a ray defined by the optical center and the point $(a_i, 1)$. Considering the vector ${}^c\mathbf{m}_i = [a_i, -1]^t$ to be a vector that is orthogonal to this ray, we can derive the following constraint in the camera coordinate system:

$${}^c\mathbf{m}_i \cdot {}^c\mathbf{p}_i = 0$$

Expanding the above equation and expressing the coordinates of ${}^c\mathbf{p}_i$ in the world frame, we can construct an equation in three variables $(\theta, {}^ct_x, {}^ct_y)$,

$$(a_i\,{}^wx_i - {}^wy_i)\cos\theta + (a_i\,{}^wy_i + {}^wx_i)\sin\theta + a_i{}^ct_x - {}^ct_y = 0. \qquad (2)$$

Since each measurement provides one constraint on the values of $(\theta, {}^ct_x, {}^ct_y)$, the three variables $(\theta, {}^ct_x, {}^ct_y)$

[1] To represent the coordinates of a vector, we follow the notation established by Craig [3]; the leading superscript indicates the frame in which the coordinates are expressed. Premultiplying the coordinates of a vector written in frame w by a rotation matrix ${}^c_w\mathbf{R}$ yields the coordinates in frame c.

in Equation (2) can be determined using three points and their images. For four points, we can construct a model-based recognition function $\mathcal{I}(\mathbf{a})$, where $\mathbf{a} = (a_1, a_2, a_3, a_4)$ is a vector of the image measurements in camera frame. Without loss of generality, we can let ${}^w\mathbf{p}_1 = (0,0)$ and ${}^w\mathbf{p}_2 = (1,0)$ by translating, rotating and scaling the four points. The recognition function is then of the form:

$$\begin{aligned}\mathcal{I}(\mathbf{a}) &= {a_1}^2(k_1 + k_9a_2 + k_{10}a_3 + k_{11}a_4 + k_{18}a_2a_3 + \\ & k_{19}a_2a_4 + k_{20}a_3a_4) + {a_2}^2(k_2 + k_8a_1 + k_{12}a_3 + k_{13}a_4 \\ & + k_{21}a_1a_3 + k_{22}a_1a_4 + k_{23}a_3a_4) + a_1(k_4a_3 + k_5a_4) \\ & + a_2(k_6a_3 + k_7a_4) + a_1a_2(k_3 + k_{14}a_3 + k_{15}a_4) \\ & + a_3a_4(k_{16}a_1 + k_{17}a_2 + k_{24}a_1a_2). \qquad (3)\end{aligned}$$

where k_i are constants determined by the world coordinates of the 2D points ${}^w\mathbf{p}_3$ and ${}^w\mathbf{p}_4$ (Due to space limitations, the equations for k_i are omitted; they can be found in [14]). Note that I($\mathbf{a}$) is a quartic polynomial in the image measurements a_i. By construction, the function $\mathcal{I}(\mathbf{a})$ is independent of the coordinate system used to specify p_i. Any two point sets that differ by a similarity transformation lead to the same recognition function, and so they are indistinguishable under $\mathcal{I}(\mathbf{a})$. Two sets of points that differ by a reflection will also be indistinguishable.

3 Selecting distinguishable landmarks

The goal of identifying and recognizing distinguishable landmarks from a set of features becomes the following simplified 2D problem: Given a set of 2D points, generate model-based recognition functions for all four-point models and select the one with the lowest Bayes cost as the most recognizable landmark. We now consider the selection problem.

If there were no measurement noise or detector bias, every instance of a model would be correctly recognized; the only falsely identified or missed landmarks would arise either from objects that are equivalent to the model up to some transformation or would occur from an accidental viewpoint. With image noise, the situation is different; the recognition function will no longer evaluate to precisely zero, and so a range $\mathcal{R}$ of values is employed. If $\mathcal{I}(\mathbf{a}) \in \mathcal{R}$, then $\mathbf{a}$ is considered to arise from an instance of model $\mathcal{M}$. Two similar 3D objects are likely to be indistinguishable from many viewpoints since their images will be similar; consequently for images of both objects, $\mathcal{I}(\mathbf{a})$ may fall within $\mathcal{R}$. To determine the optimal $\mathcal{R}$ in a Bayesian sense, we first need to determine the probability distribution resulting from application of the recognition function to noisy image data over the set of viewpoints for which the features are visible.

3.1 Distribution for one viewpoint

First, let us consider the distribution of $\mathcal{I}(\mathbf{a})$ from a single viewpoint when $\mathbf{a}$ is corrupted by noise. For a model $\mathcal{M}$, the ideal image measurements $\mathbf{a}$ from a particular viewpoint $\mathbf{v} = (t_x, t_y, \theta)$ can be expressed as $\mathbf{a}(\mathcal{M}, \mathbf{v})$ as given in Equation (1). We assume that image measurements are corrupted by additive Gaussian noise, and that the noise associated with each measurement is independent. With noise, we have

$$\tilde{\mathbf{a}} = \mathbf{a}(\mathcal{M}, \mathbf{v}) + \mathbf{n}$$

where $\mathbf{n}$ is a vector of m independent, zero mean, Gaussian random variables, each with variance σ.

The result of applying the recognition function to $\tilde{\mathbf{a}}$ is another random variable $\mathcal{I}(\tilde{\mathbf{a}})$. The probability density $p(\mathcal{I} \mid \mathcal{M}, \mathbf{v})$ could be computed using $\mathcal{I}(\mathbf{a})$, $\mathbf{a}(\mathcal{M}, \mathbf{v})$, and the known statistics of $\mathbf{n}$. Because $\mathcal{I}(\mathbf{a})$ is nonlinear, $\mathcal{I}(\tilde{\mathbf{a}})$ will not be zero mean and will not have a normal distribution. However, it appears to be problematic to compute $p(\mathcal{I} \mid \mathcal{M}, \mathbf{v})$ analytically, and even if it can be found, it is cumbersome. Therefore, we will approximate the conditional density by a Gaussian and retain the first two moments.

Since the image is corrupted with independent Gaussian noise, we can compute the moments of the recognition function for a specific viewpoint using the fact that functions of independent random variables are also independent. See [14] for the details of the computation. The mean of $(\mathcal{I} \mid \mathcal{M}, \mathbf{v})$ is

$$\begin{aligned}\eta(\tilde{\mathbf{a}}) &= E\{\mathcal{I}(\tilde{\mathbf{a}})\} = \mathcal{I}(\mathbf{a}) + [k_1 + k_2 + k_8a_1 + k_9a_2 \\ & + (k_{10} + k_{12})a_3 + (k_{11} + k_{13})a_4 + k_{18}a_2a_3 + k_{19}a_2a_4 \\ & + (k_{20} + k_{23})a_3a_4 + k_{21}a_1a_3 + k_{22}a_1a_4]\sigma^2\end{aligned}$$

where k_i are coefficients of $\mathcal{I}(\mathbf{a})$ given in Eq. (3).

The variance of $(\mathcal{I} \mid \mathcal{M}, \mathbf{v})$ is

$$\sigma^2(\tilde{\mathbf{a}}) = E\left\{[\mathcal{I}(\tilde{\mathbf{a}}) - \eta(\tilde{\mathbf{a}})]^2\right\} = q_1 + q_2\sigma^2 + q_3\sigma^4 + q_4\sigma^6 + q_5\sigma^8$$

where the q_i's are functions of $\mathbf{a}$.

3.2 Distribution for all viewpoints

The computation of section 3.1 provides the density of $\mathcal{I}(\tilde{\mathbf{a}})$ from only one viewpoint. Now, we can consider the distribution of $\mathcal{I}(\tilde{\mathbf{a}})$ taken over the range of viewpoints $\mathcal{V}$ from which the features are visible.

Since we assumed that measurement noise $\mathbf{n}$ is independent and white, the probability density for model $\mathcal{M}$ is:

$$p(\mathcal{I} \mid \mathcal{M}) = \int_{\mathbf{v} \in \mathcal{V}} p(\mathcal{I} \mid \mathcal{M}, \mathbf{v}) p(\mathbf{v}) d\mathbf{v} \qquad (4)$$

where $p(\mathbf{v})$ is the likelihood of the camera being located at viewpoint $\mathbf{v}$.

Since we assumed that measurement noise $\mathbf{n}$ is independent and white, the mean and variance of $(\mathcal{I} \mid \mathcal{M})$ can be computed from

$$\bar{\eta}(\tilde{\mathbf{a}}) = \int_{-\infty}^{\infty} \int_{\mathcal{V}} \mathcal{I}\, p(\mathcal{I} \mid \mathcal{M}, \mathbf{v}) p(\mathbf{v}) d\mathbf{v}\, d\mathcal{I} = \int_{\mathcal{V}} \eta(\tilde{\mathbf{a}})\, p(\mathbf{v}) d\mathbf{v}$$

$$\bar{\sigma}^2(\tilde{\mathbf{a}}) = E\{\mathcal{I}^2\} - \bar{\eta}^2 = \int_{\mathcal{V}} (\sigma^2 + \eta^2) p(\mathbf{v}) d\mathbf{v} - \bar{\eta}^2.$$

Thus, the average and variance of $p(\mathcal{I} \mid \mathcal{M})$ is determined by integrating the moments (e.g. $E\{p(\mathcal{I} \mid \mathcal{M})\}$) with respect to θ, t_x, and t_y for all viewpoints within the visible area $\mathcal{V}$. Supposing that the observer is

equally likely to be at any viewpoint, then $p(\mathbf{v}) = 1/\int_{\mathbf{v}\in\mathcal{V}} d\mathbf{v}$.

The visible area itself is the set of viewpoints for which all of the features in a model are visible. This set depends on camera resolution, the field of view of the camera, and possible occlusion by other surfaces in the scene. Here, we will not be concerned with possible occlusion by opaque objects, but will handle the other two issues. Because of the limited field of view of a camera, the model will only be visible from an interval of orientation. Expressing the location of the camera center in polar coordinates (r, α) as $t_x = r\cos\alpha$ and $t_y = r\sin\alpha$, the range of camera orientations for which all of the features are visible is a function of r, i.e., $\theta \in (\theta_{min}(r), \theta_{max}(r))$. To account for the finite resolution of the camera, we assume that all feature points are visible for all viewpoints in the region between two circles of radius d and R. Thus, the *visible area* is taken to be an annulus.

Unfortunately, the mean $\bar{\eta}(\tilde{\mathbf{a}})$ and the variance $\bar{\sigma}^2(\tilde{\mathbf{a}})$ cannot be integrated analytically, and so they are approximated numerically by computing the finite sum with suitably fine sampling. Using the spatial probability distribution $p(\mathcal{I} \mid \mathcal{M})$, we can now compute the recognition interval of the model $\mathcal{M}$ using the Bayes criterion.

3.3 Computing recognition intervals

Consider what happens when $\mathcal{I}(\mathbf{a})$ is applied to noisy images of some other set of feature points $\mathcal{G}$. From $\mathcal{I}(\mathbf{a})$, $\mathcal{G}$, and the known statistics of $\mathbf{n}$, the density function $p(\mathcal{I} \mid \mathcal{G})$ can be determined for observing $\mathcal{G}$ from all viewpoints. Thus we have the distribution of applying $\mathcal{I}(\mathbf{a})$ to the correct model $\mathcal{M}$ and to an incorrect set of points $\mathcal{G}$. Typically the distribution of $p(\mathcal{I} \mid \mathcal{M})$ is nearly zero mean and has a fairly small variance, whereas $p(\mathcal{I} \mid \mathcal{G})$ is likely to have a mean that is far from zero and a rather broad distribution. That is, for a mismatch, the value of $\mathcal{I}(\mathbf{a})$ is likely to be far from zero and to vary quite a bit with viewpoint.

3.3.1 Bayes criterion

The recognition problem is to decide, based on the value returned by the recognition function from a single observation, whether or not a set of image features is identified as $\mathcal{M}$. We call hypothesis H_0 the event that image measurements are the image of $\mathcal{M}$ and the alternative hypothesis H_1 that the features do not arise from $\mathcal{M}$. There is a probabilistic description corresponding to each hypothesis. We know that either H_0 or H_1 is true. A criterion for making the decision must be selected. That is, given a recognition value $\mathcal{I}(\mathbf{a})$, which hypothesis is most probably true? The Bayes criterion can be used to determine the optimal range $\mathcal{R}$ of values of $\mathcal{I}(\mathbf{a})$ which will be accepted as images of $\mathcal{M}$.

There are two kinds of errors that can be made. One is to choose H_0 given H_1 is true (false negative), the other one is to select H_1 when H_0 is true (false positive). Depending upon the application, the consequences of each type of error may not be equally important, and so costs are assigned to each type of error. Let C_{ij} denote the cost associated with choosing hypothesis H_i when in fact hypothesis H_j is true. Without loss of generality, let $C_{00} = C_{11} = 0$ and $C_{10} > C_{00}$ and $C_{01} > C_{11}$. The Bayes criterion is to select $\mathcal{R}$ so that the average cost will be minimized. Thus the region $\mathcal{R}$ where H_1 is chosen is [12]:

$$\mathcal{R} = \{\mathcal{I} \in \mathbb{R} : p(\mathcal{I} \mid H_1) > p(\mathcal{I} \mid H_0)\,(\frac{p(H_0)}{p(H_1)})(\frac{C_{10}}{C_{01}})\}$$

Choosing $C_{10} = C_{01}$ results in a region which minimizes the total error rate and is known as a maximum *a posteriori* (MAP) classifier. We denote H_1 as the hypothesis that $\mathcal{M}$ is present and H_0 as the hypothesis that $\mathcal{M}$ is not present. Let $\mathcal{G}_j$ denote some model other than $\mathcal{M}$. If the only features in the scene arise from the hypothetical models $\mathcal{M}$ and $\mathcal{G}_j$, then

$$p(H_1) = p(\mathcal{M}) \text{ and } p(H_0) = \sum_{j=1}^{n-1} p(\mathcal{G}_j) = 1 - p(\mathcal{M})$$

Assuming that all features and consequently all models are equally probable, then we have $p(\mathcal{M}) = \frac{1}{n}$, $p(H_0) = \frac{n-1}{n}$, and $p(\mathcal{G}_j) = \frac{1}{n-1}$. Furthermore, the conditional probabilities for the two hypotheses are

$$p(\mathcal{I} \mid H_1) = p(\mathcal{I} \mid \mathcal{M}) \text{ and } p(\mathcal{I} \mid H_0) = \sum_{j=1}^{n-1} p(\mathcal{I} \mid \mathcal{G}_j) p(\mathcal{G}_j)$$

where Eq. (4) can be used to compute both $p(\mathcal{I} \mid \mathcal{M})$ and $p(\mathcal{I} \mid \mathcal{G}_j)$. Bayes' rule can then be used to determine the $\mathcal{R}$ that minimizes the average Bayes cost:

$$\mathcal{R} = \{\mathcal{I} \in \mathbb{R} : p(\mathcal{I} \mid \mathcal{M}) > (\frac{\mathcal{C}_{10}}{\mathcal{C}_{01}}) \sum_{j=1}^{n-1} p(\mathcal{I} \mid \mathcal{G}_j)\}$$

The optimal range $\mathcal{R}$ may be composed of a set of disjoint intervals. Rather than employing all of the intervals of $\mathcal{R}$, we use the single interval about the mean of $p(\mathcal{I} \mid \mathcal{M})$. In particular we denote the interval $(x_l, x_r) \subset \mathcal{R}$ such that $\bar{\eta} \in (x_l, x_r)$ as the recognition interval of model $\mathcal{M}$. Since the conditional probability density functions are differentiable, we can use Newton's method to solve for the limits of the recognition interval (x_l, x_r).

3.4 Selecting the landmark

We are now ready to select the most salient or easily recognized constellation of features as a model from a given set of features. The n features in the set can be grouped into $l = \binom{n}{4}$ hypothetical models $\mathcal{M}_i$ with $i \in [1, \ldots, l]$ containing $m = 4$ points, and the corresponding recognition function $\mathcal{I}_i$ can be constructed. For a model $\mathcal{M}_i$, all other models $\mathcal{M}_j, i \neq j$ can be treated as $\mathcal{G}_j$, and the recognition interval $\mathcal{R}_i$ of the model $\mathcal{M}_i$ can be computed . We can then compute the total Bayes cost $\mathbf{C}_B$ using the error function. Given l models, there are $(l-1)$ mismatch models $\mathcal{G}_j$

for each model $\mathcal{M}_i$. Since we assume that $p(\mathcal{I} \mid \mathcal{M}_i)$ is a Gaussian distribution with mean $\bar{\eta}_i$ and variance $\bar{\sigma}_i^2$, the cost of missing the model $\mathcal{M}_i$ using the recognition interval (x_l, x_r) is:

$$\mathbf{F}_n = 1 - \int_{x_l}^{x_r} p(\mathcal{I} \mid \mathcal{M}_i) d\mathcal{I}$$

The cost of false positives (misidentifying something else as the model) is

$$\mathbf{F}_p = \frac{\mathcal{C}_{10}}{\mathcal{C}_{01}} \sum_{j=1}^{n-1} \int_{x_l}^{x_r} p(\mathcal{I} \mid \mathcal{G}_j) d\mathcal{I}.$$

Both $\mathbf{F}_n$ and $\mathbf{F}_p$ can be computed using the error function $\mathbf{erf}x = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy - \frac{1}{2}$. The Bayes cost for using $\mathcal{I}$ to recognize $\mathcal{M}$ with interval (x_l, x_r) is

$$\mathbf{C}_B = \mathbf{F}_p + \mathbf{F}_n.$$

The total Bayes cost $\mathbf{C}_B$ can be computed for each candidate model, and the models can be sorted according to these rates. Those models with lower average cost are more likely to be recognized from noisy image data and not confused from the majority of viewpoints. We thus select the model with the lowest Bayes cost as the most distinguishable landmark.

4 Implementation and examples

The presented approach to landmark selection has been prototyped in Common Lisp, and we now consider an example. Fig. 2 shows an overhead drawing of an office scene, and Figs. 3 and 4 show images of the scene. A subset of 12 vertical lines, indicated in Fig. 2, were considered features and used as input to the selection algorithm. The selection process was applied assuming that image measurements are corrupted by Gaussian noise with a standard deviation of one pixel in a 640 by 480 pixel image. To demonstrate the effectiveness of the selection technique, 24 images of an office scene were gathered from positions covering a quarter circle at three depths. The optimal landmark with the lowest Bayes cost was selected according to the procedure in Sect. 3, and the features comprising this landmark are indicated by darkened triangles in Fig. 2. In addition, a landmark with low Bayes cost was also selected, and the recognition performance for both landmarks in the 24 images will be compared. These experimental results are summarized in Tables 1 and 2. Finally, we also show that additional constraints can be imposed to reduce the matching combinatorics and improve recognition performance.

Note that there are two kinds of errors which can be made, the consequence of a false negative will often more important than a false positive since we do not want to miss the correct landmark. In our implementation, the cost factor $\frac{\mathcal{C}_{01}}{\mathcal{C}_{10}}$ was assigned to be close to the number of all hypothetical landmarks. For the optimal landmark, the mean value of

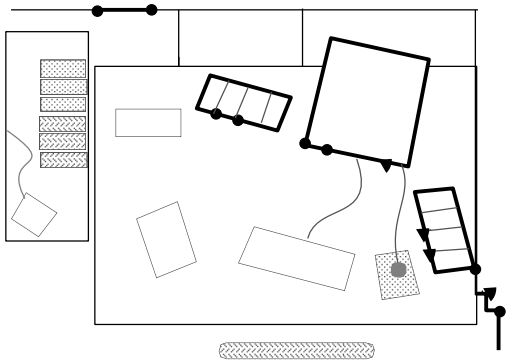
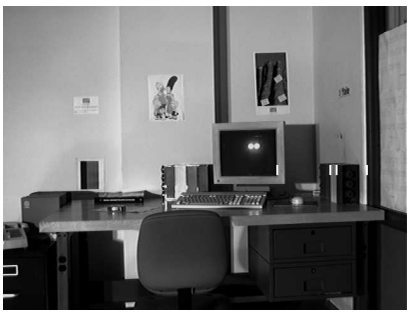


Figure 2: An overhead view of the office scene shown in Figs. 3 and 4; the optimal landmark (drawn as triangles) is selected from the set of vertical lines features (drawn as darkened points).

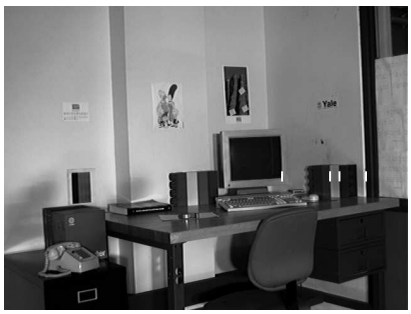
$\mathcal{I}(\mathbf{a})$ was -0.00825, and the recognition interval was $(-0.01722, -0.00632)$. The computed average cost for this landmark was 0.04951. Note that the worst hypothesized landmark had an average cost of 0.40276.

We then attempted to recognize the selected landmark in images. A one dimensional edge detector was applied across one row of the image to extract all of the vertical edges (between 13 and 21) crossing that line. The selection of an optimal landmark assumed that all features are visible and that there are no unmodelled vertical lines in the scene. To test the performance under these assumptions, the same 12 vertical edges were manually selected from the detected edges. Given 12 features, $\frac{12!}{(12-4)!} = 11,880$ groups of hypothetical landmarks can be formed. The recognition function for the optimal model was applied to 11,880 groups of features, and those falling within the recognition interval were taken as instances of landmark. There were 18 false positives and 1 false negative found in the 24 images. The resultant Bayes cost was 0.049 which is close to the theoretical one. Figure 3 shows three of the 24 images. The four highlighted vertical edges in the image indicate the correctly recognized landmark. Fig. 4.a shows another image with 17 detected edges. Two groups of four features were accepted as being instances of the landmark – only one (Fig. 4.b) is correct. Note that the interrelationship of the four edges in Figures 4.b and 4.c appears quite similar, indicating why both groups would be accepted as instances of the landmark.

Since the order of the vertical lines does not change in the 24 images, an ordering constraint can be used to reduce matching combinatorics. Given 12 features, $\binom{12}{4} = 495$ groups of hypothetical landmarks can be formed, and the same set of features as in Figure 2 had the lowest Bayes cost with the ordering constraint. The recognition function was applied to the 24 images, and the landmark was recognized in 23 out of 24 images. It was recognized as the only landmark in 18 of those 23 images. In the other 5 images, there were 1 to 4 false matches found in addition to the correct one with a total of 11 false positives. The landmark was missed in only one image, and there were two false



a. invariant=-0.016947
correctly and uniquely recognized.



b. invariant=-0.017228
correctly and uniquely recognized.

c. invariant=-0.015875
correctly and uniquely recognized.

Figure 3: The optimal landmark was correctly and uniquely found in these three images.

a. detected 17 edges
found 2 matches

b. invariant=-0.012793
correct match

c. invariant=-0.013129
false match

Figure 4: Seventeen edges were detected in the image in (a), and amongst the $\binom{17}{4} = 2380$ groups of four vertical lines, the invariant function evaluated to a value within the optimal recognition interval for the two groups shown in (b) and (c). The group in (c) is a false positive.

	Theoretical			*Selected Edges*			*All Edges*		
Constraints	$\mathbf{C}_B$	F_p	F_n	$\mathbf{C}_B$	N_{fp}	N_{fn}	$\mathbf{C}_B$	N_{fp}	N_{fn}
None	0.049	0.008	0.041	0.048	282	1	0.047	1346	1
Ordering	0.045	0.005	0.040	0.048	11	1	0.049	57	1
Ordering and Gradient				0.047	0	1	0.047	3	1

Table 1: Theoretical and experimental results for an optimal landmark. The theoretical performance can be compared to the experimental performance when the input is composed of the same twelve edges used during selection and when it includes all detected edges. The improvement gained by using additional constraints is also demonstrated. $\mathbf{C}_B$ denotes the Bayes cost, F_p is the false positive cost, F_n is the false negative cost, N_{fp} denotes the experimental number of false positives, and N_{fn} is the number of false negatives.

	Theoretical			*Selected Edges*			*All Edges*		
Constraints	$\mathbf{C}_B$	F_p	F_n	$\mathbf{C}_B$	N_{fp}	N_{fn}	$\mathbf{C}_B$	N_{fp}	N_{fn}
None	0.402	0.045	0.347	0.388	1832	8	0.387	10243	8
Ordering	0.384	0.031	0.353	0.391	116	8	0.387	437	8
Ordering and Gradient				0.384	41	8	0.381	102	8

Table 2: Experimental results for a bad landmark with high Bayes cost.

matches found in that image. In summary, the experimentally computed Bayes cost was 0.048.

Since the sign of the gradient of the image intensity across the four edges in the model and measured features must be consistent, this gradient constraint can be used in the recognition process to improve performance. In the current form of the selection process, we cannot exploit this constraint, however it can be used during on-line recognition. With ordering and gradient constraints and the recognition interval computed above, the experimental the number of false positive matches was reduced from 11 to 0.

We consider the performance when the recognition algorithm is applied to all detected vertical edges. Given k features, $\frac{k!}{(k-4)!} = (17160 \sim 143640)$ groups of hypothetical landmarks can be formed. With no constraints, the recognition function was applied to $17,160 \sim 143,640$ groups of features. For the optimal landmark, the experimental Bayes cost was 0.048 for the 24 images. The process was also applied using both the ordering and gradient constraints, and the results are summarized in Table 1.

The same set of experiments was applied to a group of four features which had a high Bayes cost (an order of magnitude larger than for the optimal landmark). It is expected that many more mistakes are likely. This hypothesis is experimentally validated as shown in table 2. After applying the ordering and gradient constraints, there were still many false positives and false negatives for this bad landmark. The lower Bayes cost of the selected landmark resulted in many more correct matches than for the bad landmark in all cases; that is, the selected landmark is much more distinguishable than the bad landmark.

5 Discussion

The method described is a starting point for a Bayesian approach to landmark selection. In the process, a number of assumptions and simplifications were made. Further empirical investigation is needed to determine the validity of this model for landmark selection. There are a number of issues, improvements and extensions to this basic scheme.

- Though the close match of theoretical and experimental recognition performance demonstrates the validity of the assumptions and simplifications, further study is required to either relax some of these assumptions and empirically verify the simplifications.
- When hypothesizing possible landmarks, all $\frac{n!}{(n-4)!}$ hypothetical groups of features were considered. This is an explosive number, and so principled means of reducing the number of hypothetical landmarks, such as the ordering constraint, must be developed. The combinatorial issue was addressed for the related robot localization problem by Sugihara [10].
- Finally, it is obvious that inclusion of additional features in the landmark (e.g. five or more vertical lines) would greatly improve recognition performance. The explosion in the number landmarks becomes even more crucial, and perhaps recognition strategies such as constrained search (interpretation trees) would be necessary.

Acknowledgements

Many thanks to C.J. Taylor whose work on landmark-based navigation started us on this path.

References

[1] S. Atiya and G. Hager. Real-time vision-based robot localization. *IEEE Trans. on Robotics and Automation*, 9:785–800, 1993.

[2] R. A. Brooks. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2(1):14–23, Mar. 1986.

[3] J. Craig. *Introduction to Robotics: Mechanics and Control*. Addison-Wesley, New York, 1989.

[4] A. Kosaka and A. Kak. Fast vision-guided mobile robot navigation using model-based reasoning and prediction of uncertainties. *CVGIP: Image Understanding*, 56(3):271–329, 1993.

[5] D. J. Kriegman, E. Triendl, and T. O. Binford. Stereo vision and navigation in buildings for mobile robots. *IEEE Trans. on Robotics and Automation*, 5(6):792–803, Dec. 1989.

[6] K.-D. Kuhnert. Fusing dynamic vision and landmark navigation for autonomous driving. In *IEEE Int. Workshop on Intelligent Robots and Systems*, pages 113–119, July 1990.

[7] B. Kuipers and Y. Byun. A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations. *Robotics and Autonomous Systems*, 8:47–63, 1981.

[8] Lazanas and Latombe. Landmark-based robot navigation. In *Proc. Am. Assoc. Art. Intell.*, 1992.

[9] T. Levitt, D. Lawton, D. Chelberg, and P. Nelson. Qualitative navigation. In *Proc. Image Understanding Workshop*, pages 447–465, 1987.

[10] K. Sugihara. Some location problems for robot navigation using a single camera. *Comp. Vision, Graphics, and Image Proces.*, 42:112–129, 1988.

[11] C. Taylor and D. Kriegman. Vision-based motion planning and exploration algorithms for mobile robots. In K. Goldberg, D. Halperin, J. Latombe, and R. Wilson, editors, *The Algorithmic Foundations of Robotics*. A. K. Peters, Boston, MA, 1995. Proceedings from the workshop held in February, 1994.

[12] H. L. V. Trees. *Detection, Estimation, and Modulation Theory*. John Wiley and Sons, 1968.

[13] D. Weinshall. Model-based invariants in 3-D vision. *Int. J. Computer Vision*, 10(1):27–42, 1993.

[14] E. Yeh and D. J. Kriegman. Toward selecting and recognizing natural landmarks. Technical Report 9503, Yale University, 1995. Available from anonymous ftp at daneel.eng.yale.edu.