Yale University
Department of Computer Science

Place Recognition Using Image Signatures

Sean P. Engelson

YALEU/DCS/TR-946
January 1993

This work was partially supported by the Defense Advanced Research Projects Agency, contract number DAAA15-87-K-0001, administered by the Ballistic Research Laboratory. The author is supported by a fellowship from the Fannie and John Hertz Foundation

Place Recognition Using Image Signatures

Sean P. Engelson

Abstract

For reliable navigation, a mobile robot needs to be able to recognize where it is in the world. We describe an efficient and effective image-based representation of perceptual information for place recognition. Each place is associated with a set of stored *image signatures*, each a matrix of numbers derived by evaluating some *measurement function* over large blocks of pixels. Measurements are chosen to be characteristic of a location yet reasonably invariant over different viewing conditions. Signature matching can be done quickly by element-wise comparison; greater stability is assured by offset matching. Multiple measurements are easily used in tandem for enhanced recognition accuracy. Even so, all image-based techniques are subject to the *recognition problem*, in that many scenes are inherently unrecognizable, due to view ambiguity and instability. We deal with this problem by using active methods to select the best signatures to use for recognition. We formulate heuristic *distinctiveness metrics* as functions on image signatures which are good predictors of view distinctiveness. These functions are used to direct the motion of the camera to find locally distinctive views. These views are also stable, since we use local optimization techniques. We evaluate the results of applying this method with a camera mounted on a pan-tilt platform.

This work was partially supported by the Defense Advanced Research Projects Agency, contract number DAAA15-87-K-0001, administered by the Ballistic Research Laboratory. The author is supported by a fellowship from the Fannie and John Hertz Foundation

Contents

List of Figures

List of Tables

1 Introduction

Mobile robots often need to register their current location in a stored map of the world. There are chiefly two types of robot maps used: metric and topological. Metric maps represent the geometrical shape of the world to one degree or another. Systems using such representations can thus make use of a more-or-less well-defined correspondence between map components and environmental structures. However, this approach becomes less useful and more cumbersome when sensor uncertainty is introduced; furthermore, information is not easily represented in a way which is useful for task performance.

The topological approach, pioneered by Kuipers [17, 18], avoids these problems by discretizing the world in a robot-relative manner. A topological map is a graph whose nodes represent places in the world, and whose arcs encode robot actions taking the robot from one place to another. Since topological representations focus on the structure of the robot's paths rather than that of its environment, they are immediately useful for navigation; in addition they appear to be more concise than the metric alternatives. In [9] we proposed a system for learning topological maps reliably in the presence of sensor and odometric error.

Any system which makes use of topological maps must be able to tell which place node represents the robot's current location (if any). In particular, a map-learning system must be able to tell if it is at a place it already knows about (whose description it should update) or an unknown place (and therefore augment its map). Thus, we posit that place representations must include a description of the place sufficient to support reliable recognition. Some of this burden can be laid on representations encoding the relative positions of places in the map [7, 28]. In [9], we introduced the *diktiometric* map representation, which integrates both topological and high-level geometric knowledge in a uniform framework. In any case, due to odometric uncertainty, perceptual cues must be used as well.

1.1 The basic idea

Image-based place recognition has several benefits over reconstructionist approaches. Recognition based on visual extraction of geometric features is often slow and error-prone. Representing uncertainty in a way which is efficient, complete, and consistent is also a difficult problem. We therefore propose to use an image-based matching technique, using arrays of measurements called image signatures. The method of image signatures is simple and easy to apply, generally applicable to all sorts of environments, computationally inexpensive, and is easily integrated with other recognition cues. When used with appropriate measurement functions, image signatures support efficient recognition over a range of viewer positions and orientations.

Signatures are matched against one another to confirm hypotheses of the robot's location, and can be used to efficiently index into a knowledge base to generate such hypotheses. Reduction of images to signatures can be very fast, since only simple image-based processing is required. Matching is also quite fast, as only a small amount of data is involved. By matching signatures at offsets and using signatures of differing resolution, approximate correction can be made for differences in viewing position. Databases of image signatures can also be indexed efficiently due to their simple structure.

An image signature is an array of measurement values. The input image is tesselated into a grid of subimages. Each measurement value in a signature is derived from a subimage by a measurement function. Measurement functions usually map from image regions to numbers or angles. The idea is that a measurement represents some coarse-scale physical property of a portion of the scene. By representing images coarsely, signatures achieve significant data reduction, and also affords stable recognition. Recognition is decided by element-wise matching. Signatures from several measurement functions can be used together for conjunctive match

filtering. Viewpoint rotation can be corrected for by matching signatures at an offset. Signatures can be indexed for hypothesis generation using standard techniques for multidimensional point indexing (cf. [26]). These techniques provide efficient search methods for finding good matches for an input image. Image signatures thus provide an simple, efficient, and effective basis for place recognition.

An inherent difficulty with any image-based recognition method is the *recognizability problem.* This problem consists of two factors. First, many scenes inherently give little information for localization (eg, blank walls). Second, there are too many scenes in the world to remember them all—which should be used for recognition? The image signature paradigm provides a direct approach to dealing with this problem. We develop the notion of distinctiveness metrics, real-valued functions of signatures such that high values indicate unambiguous scenes. As for recognition, distinctiveness values can be combined for different measurement types for added reliability. Optimization techniques are then used for directing camera motion to find local distinctiveness maxima. Such distinctive viewpoints will tend to be unambiguous. Furthermore, since there are few local maxima, relatively few scenes will ever be used for recognition. This further underscores the usefulness of the image signature method.

1.2 Place recognition

Nelson's work on image-based homing [23] used a coarse pattern-matching approach, which our framework generalizes. His 'patterns' are essentially image signatures based on measurements of dominant edge orientation; he used them to construct a reactive plan for homing a robot to a predetermined location. We generalize the notion to encompass matching based on multiple feature types, at different offsets, and at multiple scales. A similar approach is taken by Zetsche and Caelli for the problem of invariant pattern recognition [32]. They use oriented gaussian filters to derive a 4D translation-, scale-, and rotation-invariant representation of 2D input patterns, using cross-correlation for matching.

Another approach to place recognition for topological maps is to use local geometry. Sarachik [27] has developed a robust method for estimating the size and aspect ratio of rectangular rooms. In a non-homogeneous environment, this could be useful for place recognition. Kriegman [16] developed a system for instantiating generic place models (such as hallways), which can be used both for recognizing place categories, and for matching based on high-level place descriptions. Moravec and Elfes [22] used correlation of sonar-derived certainty grids for rough place recognition and robot localization. Braunegg [6] developed a method for matching place descriptions based on the relative 3D positions of vertical edges. In a similar vein, Leonard and Durrant-Whyte developed a sonar-based mapping system [19], which effectively performs localization using sparsely distributed geometric features.

Contextual information can also be directly integrated into the place recognition process. For example, Kirman, Basye, and Dean [14] use a Bayesian decision theoretic approach to classifying the robot's place. They incorporate a temporal component to take into account the robot's recent history, allowing them to disambiguate places based on topological considerations. Mataric uses a unique topological mapping scheme based on closed cycles [20], which makes contextual cues very useful for place recognition. Her system also uses place categorization for recognition, done by analysing the robot's behavior in different locations.

1.3 Paper overview

The remainder of this paper describes our results in applying the image signature method to the problem of place recognition[1]. We first develop the notion of image signature and

[1] Some aspects of this work were previously reported in [10] and [8].

stable matching, within the context of our robust map learning framework. We then turn to the question of measurement function design, exhibiting a sample of such functions, with some empirical comparisons. Experimental results testing image matching in a large database support our contention that image signatures can be effective at recognition tasks. This established, we describe a method based on measuring signature distinctiveness to deal with the problem of selecting good viewpoints for recognition. Experimental results using a pan-tilt platform support the use of local distinctiveness search for viewpoint selection and suggest directions for further research.

2 Image Signatures

We need a representation of an image which can be stably matched to representations of other images showing the same scene. It should also be insensitive to the presence of image noise. To satisfy these requirements, we define an *image signature* as an array of measurement values, each value derived from a portion of the original image. An input image is tesselated evenly on a grid (typically square), and a *measurement function* is applied to each subimage, giving a value for the corresponding position in the signature array. Such measurement functions map from image regions onto a low-dimensional codomain, typically either the reals or a discretized space. One example of a measurement function is the dominant edge orientation (DEO) for a region (see Figure 1); it is reasonable to expect that two image regions having a similar dominant edge direction share some physical similarity. Signature matching is done by comparing corresponding elements for similarity; there are different types of matching, discussed below. A large number of images can be stored this way, since typical signatures will take up no more than perhaps a hundred words in an optimized implementation.

In the remainder of this section, we develop the machinery necessary to use image signatures for place recognition. We begin with a discussion of how signatures are computed. This amounts to a discussion of measurement functions. In addition to describing the particular measurement functions we developed during this work, we also formulate evaluation criteria that can be used to design good sets of measurement functions in general. In the next section we describe methods for signature matching, including how we deal with viewpoint motion. We then describe how signatures derived from multiple measurements are used together for place recognition. We then conclude this section with discussions of scene ambiguity and hypothesis generation.

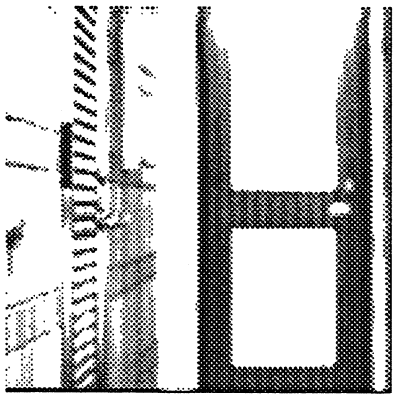
2.1 Measurement functions

A crucial part of a place recognition system based on image signatures is the choice of measurement functions. This choice is, to some extent, dependent upon the robot and its environment. Although there is, as yet, no well-founded theory to guide development of good measurement functions, we can state some heuristic principles that can be used to design and evaluate measurement functions. Also, measurement functions should be quickly computable. Measurement functions map image regions into values; without loss of generality, we speak of measurement functions applying to entire images.

When considering measurement functions, it is often the case that we need to distinguish between *sparse* and *dense* measurement types. Sparse functions measure an overall property of an image region by extracting a set of tokens from the region and computing an aggregate measurement for the token set. A sparse measurement value can occasionally change drastically if the viewpoint changes so that a token moves into or out of view. By contrast, dense measurements compute a value by averaging a pixel-based measurement over the image region. Thus small viewpoint changes produce small dense measurement changes.

2.1.1 Some measurement functions

We have developed a set of measurement functions which generally work reasonably well. We describe them and their implementations below, as well as their classification as sparse or dense. Examples of signatures derived using the various measures are shown in Figure 1. While we have no formal results describing the properties of these functions, we present empirical evidence for their stability and distinguishability in Sections 4.1.1 and 4.1.1.



Image

$$\begin{bmatrix} 0° & 180° & 0° & 180° \\ 90° & 0° & 0° & 180° \\ 0° & 180° & 0° & 180° \\ 90° & 180° & 0° & 180° \end{bmatrix}$$
DEO

$$\begin{bmatrix} 3.5 & 1.4 & .55 & .57 \\ 3.5 & 1.5 & .55 & 1.45 \\ 2.5 & 1.9 & 1.5 & .57 \\ 2.5 & 2.1 & .57 & 1.4 \end{bmatrix}$$
SGD

$$\begin{bmatrix} .39 & .34 & .16 & .20 \\ .36 & .38 & .17 & .18 \\ .37 & .40 & .23 & .22 \\ .37 & .43 & .22 & .25 \end{bmatrix}$$
ZC

$$\begin{bmatrix} .30 & .29 & .08 & .13 \\ .32 & .18 & .11 & .15 \\ .23 & .14 & .11 & .18 \\ .20 & .11 & .14 & .19 \end{bmatrix}$$
TEX

$$\begin{bmatrix} 66 & 113 & 103 & 163 \\ 67 & 88 & 121 & 140 \\ 47 & 44 & 90 & 121 \\ 36 & 35 & 96 & 120 \end{bmatrix}$$
ESTR

$$\begin{bmatrix} .92 & 1.1 & 1.3 & 1.2 \\ .84 & 1.0 & 1.2 & 1.1 \\ .83 & .92 & 1.0 & 1.0 \\ .79 & .86 & 1.0 & 1.0 \end{bmatrix}$$
RI

Figure 1: Examples of image signatures for various measurement functions. All angles are given with 0° vertical. SGD is given in multiples of 45° (the scale that the matcher uses).

Dominant edge orientation

The first measurement function we discuss is the dominant edge orientation (DEO) in a region, suggested by Nelson [23] for image-based homing. The intuition behind this is that since edges are discrete and mostly derive from scene features (occlusions, albedo, and so on), the dominant orientation in a subimage will usually change little with motion, but will be useful for recognition. We calculate DEO by convolving the image with the two first-order 3×3 Prewitt

operators [2] to estimate the image gradient at each pixel. We then restrict attention to those pixels whose gradient magnitude is both a local maximum in the gradient direction, and which is above a threshhold, $\theta_{\text{grad}}^{\text{DEO}}$. The selected pixels are then classified into one of eight categories denoting the edge orientation implied by the gradient. The orientation to which most of the pixels in the subimage are assigned is the DEO value of the subimage. Since DEO extracts edges, it is a sparse measure.

Significant gradient direction

Another measurement, closely related to the DEO, is the significant gradient direction (SGD). The directional intensity derivatives at 0, 45, 90, and 135 degrees are estimated using rotated versions of the 5×3 Prewitt operator (at skew angles, the convolution masks are 5×5). The mean of all pixels above a threshhold, $\theta_{\text{der}}^{\text{SGD}}$, in each derivative subimage is computed, giving a 'response' for each direction. The average of the top two directions is computed, weighted by their responses; this is the SGD of the subimage. The SGD is similar to the DEO, except that (a) it takes into account the magnitude of contrast, and (b) it does not explicitly tokenize edges by non-maximum suppression. Voting on edge direction does, however, constitute a form of tokenization, so SGD is considered a sparse measurement function.

Edge strength

In addition to measuring the direction of the edges in the image, the strength of those edges can also be measured. The edge strength (ESTR) measure is computed as the average directional response over all filter directions (as computed for SGD). ESTR is simply an average 'edginess' for an image, and so is a dense measurement.

Edge density

A fourth edge-based measurement is the density of edges in the subimage (ZC). This gives a rough measure of the visual complexity of the scene. It can be computed by convolving the subimage with a Laplacian of Gaussian filter (we use a 3×3 mask), and then counting the number of vertical and horizontal pixel boundaries constitute zero-crossings of the convolved image. The fraction of such zero-crossing boundaries constitutes the edge density of the subimage. While fairly fine-grained, ZC relies on picking out edge fragments, and so is sparse.

Texturedness

We have also developed a simple measure which is more specifically sensitive to image texturedness. We call a pixel's 'texturedness' the number of neighboring pixels whose intensities differ from it by more than a fraction $\theta_{\text{diff}}^{\text{TEX}}$ of the (sub)image's mean. Then we calculate an overall texturedness for a subimage as the average of pixel texturedness on the subimage. This measure will not be much confused by edges, as areas will predominate by the number of pixels involved in the average. Texturedness averages a local measure over the entire subimage, so is a dense operation.

Relative intensity

We can also roughly estimate the average reflectance of surfaces in a scene by measuring the relative intensity of a subimage compared to the entire image. We simply compute the ratio of the intensity mean of each subimage to the intensity mean of the entire image. This normalizes

for the ambient illumination level. Naturally, this measure will not be stable if the light source direction is changed such that the shadow structure of the scene changes. In fact, this is the least useful of the measurement functions tested. Relative intensity is clearly a dense measurement.

Other measurements

In addition to the measurement functions described above, we tried some others that were found wanting. We describe them briefly here, and why we think they failed.

Vertical line count This measure extracted vertical lines that extended the full height of the subimage and counted them. The intuition was to take advantage of the strong vertical orientation of our office environment. It didn't work too well, though, probably because vertical lines tend to come in dense patches so that small motions could change the count considerably.

Intensity centroid We also tried using measurements based on the grey-level centroid for something like region segmentation. Unfortunately, these measurements are very sensitive to camera motion for obvious reasons, which are difficult to correct for.

There is also a great deal of previous work that can be applied to measurement function development. For example, detailed texture analysis (eg, [31]), color histograms [29], or shading analysis (as in [15, 25]) might provide good physically-based measurement functions.

2.1.2 Evaluating measurement functions

What makes some measurement functions useful and others not? We would like a succinct characterization of measurement 'goodness' which we could use when designing measurement functions for a given environment. We can derive criteria for good measurement functions by considering the properties demanded by place recognition. First of all, views from nearby configurations should map to close measurements, since views from a single place should look similar. This we call the *stability* criterion. Dominant edge direction, for example, satisfies this criterion, since the features being measured change slowly and rarely disappear from view after a small movement (this argument is made more formally by Nelson [23]). Another desirable sort of stability is invariance with respect to lighting conditions. Unfortunately, it is nearly impossible to achieve in general, so the robot may have to store different signatures for different lighting conditions. But if shadows are rare or small, they will not often affect results.

Another criterion, *distinguishability*, is that views from distant configurations should have different measurements, so that recognition can be reliably performed. More practically, this becomes a requirement that different physical layouts (geometry, reflectance, etc.) should look different under the measurement. Since the information in the image is being reduced severely, this criterion will not be completely satisfied; however, measurements that come close to describing inherent physical properties of the scene will be preferred. Thus, since contrast edges usually denote physical structures in the world (occlusions, surface markings, and so on), we expect dominant edge orientation to be a distinguishing measure.

2.2 Signature matching

2.2.1 Similarity metrics

There are different methods that could be used to determine if two signatures may match, depending on the properties of the measurement function used. Those we consider allow a

match if some 'similarity metric' is above a given threshhold. So, if the robot wishes to test the hypothesis that it is at a particular place, an allowable signature match permits the hypothesis to be accepted. If competing hypotheses must be disambiguated, the match with highest similarity is chosen.

Nelson, in [23], suggests using the fraction of identical elements in two signatures (which he calls 'patterns') as a similarity metric. He uses discrete measurements—for real-valued measurements we can define two values to be 'identical' if their difference is less than a threshhold. So, if s and t are $n \times n$ image signatures, we have the *pattern similarity* metric:

$$\mathbf{sim}_{\text{pat}}(s,t) = \frac{1}{n^2}\sum_{ij}\text{id}(s_{ij}, t_{ij})$$

where

$$\text{id}(v_1, v_2) = \begin{cases} 1 & \text{if } |v_1 - v_2| \leq \theta_{\text{id}} \\ 0 & \text{otherwise} \end{cases}$$

This metric corresponds to using a 0–1 loss for measurement estimation [4], and so is reasonable if measurement deviations are usually less than θ_{id}, but occasionally are completely unpredictable. This will typically be the case if the measurements are based on tokens representing features in the world which may suddenly disappear from view when the viewpoint changes. If features are sparsely distributed, few will move out of a given image region under small motions—allowing pattern similarity to match the derived signatures. A match for signatures s and t, with measurement function M, is *permitted* under pattern similarity if $\mathbf{sim}_{\text{pat}}(s,t) \geq \theta^M_{\text{pat}}$.

A second similarity metric is based on the *root-mean-square-difference*, or *cross-correlation*, of the signature array elements. To wit:

$$\mathbf{sim}_{\text{rms}}(s,t) = -\sqrt{\frac{1}{n^2}\sum_{ij}(s_{ij} - t_{ij})^2}$$

This similarity metric corresponds to using a squared-difference loss for measurement estimation [4], hence a Gaussian probability density on measurement deviation[2]. More generally, this metric is reasonable whenever the chance of a particular measurement deviation decreases smoothly with deviation's magnitude. In particular, dense measurement functions vary continuously with robot motion, so $\mathbf{sim}_{\text{rms}}$ will be appropriate. A match is permitted (as above) if $\mathbf{sim}_{\text{rms}}(s,t) \geq \theta^M_{\text{rms}}$.

2.2.2 Offset and cross-scale matching

Recall that one of our goals for matching is stability—signatures from nearby views should be 'close'. However, even with reasonably stable measurement functions, signatures will only match by the above criteria if they line up almost exactly. For example, with 5×5 signatures and a 25 degree field of view, a rotation of only 5 degrees will induce a signature match where no matching parts of the scene are compared. But if matching is done at a horizontal offset of one column, the signatures will match. Thus we see that matching signatures at horizontal offsets can correct for rotation about the vertical axis. Of course, under planar projection, the corresponding subimages will not be identical, but we have empirical evidence (see section 4.1.1) that the difference is tolerable.

[2] It is possible that matching could improved by reducing the influence of outliers, using robust estimation techniques [13]. This kind of approach has been applied to good effect for pixel-level matching, see [5] for example. We have not yet examined the effects of robustification on signature matching.

Translation can be corrected for similarly. Translation perpendicular to the optical axis results in simple image translation and can be dealt with by offset matching as above. If translation is purely along the optical axis, then the focus-of-expansion (FOE) is at the center of the image. Provided that objects in the scene are reasonably far from the observer and motion is small, the image change can be approximately modelled as pure expansion. If this is the case, then the center portion of a high-resolution signature at the original position should match a lower-resolution signature at the forward position. Other translations can be approximately corrected for by combining multiscale and offset matching.

2.3 Using multiple measures

Despite our best efforts, any measurement function we design will be far from perfect, due to the amount of information reduction inherent in the process. Hence we use additional cues to filter allowable matches and improve recognition. The most obvious is to use several measurement functions together for matching. The basic idea is quite simple. Several signature databases are kept, one for each of the different measurement functions used. A suggested image match is permitted if the several signature matches implied by the hypothesis are themselves permitted. Even though one particular measurement function may accidentally allow an invalid match, it is less likely that several will. This is only valid, of course, if the measurement functions are, in some sense, independent. The notion of independent measurements we discuss in more detail later.

More precisely, matching using multiple measurement functions is done as follows. Given a match hypothesis, corresponding signatures are checked to see if a match is allowed, based on the measurements' threshholded similarity metrics. Note that different measurements may use entirely different match criteria. If fewer than a given number of the signature matches are disallowed, then the match is allowed. If enough measurement functions are used, then allowing one or two mismatches avoids some false negatives without unduly hindering positive recognition.

Choosing the 'best' match among several permissible alternatives is more difficult, however, since signature similarities for different measurements are not generally commensurable. The strategy we currently use is to designate one measurement as 'special' and rank matches based on similarity with respect to that measurement. There are a host of other methods for producing orders using all an image's signatures, but this simple method seems to work reasonably well (for a good choice of special measurement).

2.3.1 Measurement set evaluation

For a set of measurement functions to be usefully combined, they should be 'independent' in some sense. If different measurement functions measure different image properties, then they will tend to disagree on accidental matches of different scenes, and hence the set will perform better than its elements.

Dependence

The most direct way to evaluate measurement independence is to estimate the probability that the measurements agree on an incorrect match. This gives the dependence of the measurements

$$D = P(\bigwedge_i m_i \mid I_1 \not\sim I_2 \wedge \bigvee_i m_i)$$

where I_j are images, m_i denotes the event that I_1 and I_2 are deemed to match under measurement function i, and $I_1 \sim I_2$ iff the images arise from the same scene. This can be simplified as follows:

$$\begin{aligned} D &= P(\textstyle\bigwedge_i m_i \mid (\bigvee_i m_i) \wedge I_1 \not\sim I_2) \\ &= \frac{P((\bigwedge_i m_i) \wedge (\bigvee_i m_i) | I_1 \not\sim I_2)}{P(\bigvee_i m_i | I_1 \not\sim I_2)} \\ &= \frac{P(\bigwedge_i m_i | I_1 \not\sim I_2)}{P(\bigvee_i m_i | I_1 \not\sim I_2)} \end{aligned}$$

This can be estimated from a representative sample of images S after computing sets of incorrect matches for each measurement function i, $B_i \subset S \times S$ by

$$\frac{|\bigcap_i B_i|}{|\bigcup_i B_i|}$$

This provides an easy way to approximately evaluate the independence of a set of measurement functions. In section 4.1.1 we evaluate a set of measurement functions by this criterion.

Measurement utility

What may be more useful in practice than the raw independence of a measurement set is to determine how useful it would be to add another measurement function to a given set. In decision-theoretic terms, this corresponds to calculating the expected risk of using the augmented set over the given set. If we ignore the cost of storing another signature database for the new measurement function, all we need to know is the cost of applying the new measurement function, C_a, and the cost of allowing an incorrect signature match C_m. In practice, this latter may be difficult to estimate. We can conceive, though, that it might be estimated experientially over a period of time by correlating image signature matching with other place recognition cues. The probability that the new measurement will rule out a bad match that would otherwise have been accepted is given by

$$\begin{aligned} P(\neg m_{\text{new}} \mid m_{\text{old}} \wedge I_1 \not\sim I_2) &= \frac{P(\neg m_{\text{new}} \wedge m_{\text{old}} | I_1 \not\sim I_2)}{P(m_{\text{old}} | I_1 \not\sim I_2)} \\ &= \frac{P(r_{\text{new}} \wedge m_{\text{old}} | I_1 \not\sim I_2)}{P(m_{\text{old}} | I_1 \not\sim I_2)} \\ &= \frac{|R_{\text{new}} \cap (\bigcap_i B_i)|}{|\bigcap_i B_i|} \quad \text{(estimated)} \end{aligned}$$

where m_{new} and m_{old} are the events that a match is allowed by the augmented and non-augmented measurement sets respectively, r_{new} is the event that the new measurement function rejects a match, and $R_{\text{new}} \subset S \times S$ is the set of matches from a sample set rejected by the new measurement function. If we call this probability p, then the expected utility of adding the new measurement is given by $C_a - pC_m$. This can then be used to evaluate new measurement functions, so that system design can be done incrementally (and perhaps automatically).

2.4 Ambiguity

Many scenes in the real world are inherently ambiguous. For example, if a robot is navigating in an office building, an image of a blank wall gives it almost no information[3], as blank walls are everywhere. Hence, ambiguous images should not be used for matching; if the robot is somewhere with an ambiguous view, it should look around for a less ambiguous one to use for matching. Ambiguity filtering of a signature database can easily be done by removing all signatures which match to more than a given fraction of the signatures in the database (say, 5% for a large database). Then, if an input image is ambiguous by this criterion, it is also ignored, and the robot should move a bit and try another view. This method is quite easy to implement, but not too efficient for large databases. In the sequel we address the problem of directly determining if an image is ambiguous and using this information to choose useful viewpoints for recognition.

2.5 Hypothesis generation

The above discussion on matching image signatures assumes that all signatures in the database are checked as possible matches. Since in practice the robot may need to store thousands of images, it must be able to find the relevant ones effectively. Such hypothesized matches can be found relatively easily if the robot has strong expectations about its current location. If the robot believes it is at one of a small set of places in its map, then it need only search signatures corresponding to those places to disambiguate its position. If it fails to rule out all but one possibility, then it can look around some more and repeat the process with another image.

If, on the other hand, none of the expected signatures match, or the robot has no expectations at all, the robot must figure out where it is from scratch. If it has knowledge about its geometric position from odometry, that may be used to generate candidate places the robot might be. If not, or if its positional information is not good enough, the robot must solve the ***kidnapped robot problem.*** It is as if the robot fell asleep and then was silently kidnapped and moved to some new location by gremlins. When the robot wakes up, it must come up with some idea of where it is based on purely perceptual cues (we ignore here the use of experimentation). Thus, we require an efficient way to index a signature database for effective retrieval of reasonable match hypotheses.

The solution is to index signatures using k-d trees (a k-dimensional generalization of binary search trees [26]). Direct search can be efficiently performed on a k-d tree for points falling within intervals of $\mathbb{R}^k$. A more flexible search strategy for our purposes, though, is a variety of *spiral search* [21]. Spiral search uses a heuristic strategy to find points in the tree progressively further and further from a given starting point. Note that k-d tree search can be efficiently implemented on massively parallel machines using marker-passing.

For our application, the simplest thing to do would be to treat $n \times n$ signatures as points in $\mathbb{R}^{n^2}$ and index them in the tree directly. However, this does not take into account the element-wise flexibility of our similarity measures (some elements may be very different) or offset/multi-scale matching. These problems can both be heuristically ameliorated by indexing on signature subarrays. If we index each signature under each of its columns, offset match hypotheses can be generated using markers. Each signature keeps a vector of markers, one for each possible offset. Every time a column of the input signature is 'matched' to an index column, the corresponding offset marker is incremented; when it goes above some fraction of n, a match hypothesis at that offset is proposed. This also eases the problem of some elements being far off, since only some columns must be nearby. This approach may be generalized by indexing based on arbitrary

[3] Of course, generic knowledge that the robot is looking at a blank wall may, combined with other cues, serve to disambiguate its location. We are concerned here, however, with highly-disambiguating perceptual cues.

sub-signatures—supporting generation of cross-scale match hypotheses. However, there are an exponential number of such possible indices, so only 'reasonable' ones should be used. A good way to filter indices is to only use those which are relatively unambiguous; if an index points to hundreds of signatures it doesn't constrain the hypothesis space much.

3 The Recognizability Problem

There are several difficulties with the naive approach to place recognition described above. First, if all images seen are stored, the signature database will grow enormously. Thus, some method is needed to choose which signatures are most useful for recognition, and should be stored. This brings us to the second point, which is that a large fraction of images in many environments are ambiguous, and tell us nothing about the viewer's location. Consider how many viewpoints in an office building give an image of a blank wall. In the experiments described above, upwards of 35% of images were found to be ambiguous. Measuring ambiguity by seeing how many signatures in a database are matched by an input image is expensive, in both time and space (since ambiguous signatures must be stored for testing). Hence, finding distinct images should not depend on direct examination of a signature database. A final consideration is the problem of *stability*. The difficulty here is that there are configurations where a change in viewpoint can radically change the view seen. Thus, if the naive approach is applied at an unstable viewpoint, many more views will need to be stored for a small region, to cover all the possible views. The question we address in this section is finding viewpoints which are distinctive and stable. We term such viewpoints (and the images they induce) *recognizable*.

To find recognizable viewpoints, we use the fact that place recognition is performed from a mobile platform, and hence the camera can be moved during recognition. Most work in vision that uses camera motion can be divided into two categories: *active vision* and *sensor planning*. Active vision techniques focus on temporal integration of visual information over a controlled trajectory to improve interpretation (see [1]). Usually, low-level intensity/feature models are used to determine camera motion (for, eg, fixation as in [3]). Visual servoing of some sort is often used for controlling camera motion, though arbitrary motion can also be helpful [24]. Sensor planning, on the other hand, addresses the problem of finding good viewpoints for object recognition and registration [30]. This is done by optimizing some criterion of image goodness, using measures such as number of visible useful features or expected accuracy of image measurements. A high-level model is used to determine the best viewpoint subject to kinematic constraints, given some previous estimate(s) of the target's pose. The problem can also be formulated as a decision-theoretic problem of choosing the viewpoint with highest expected utility [12]. This can be iterated to obtain successively better estimates (of target identity or pose).

The problem of finding recognizable viewpoints is similar to the classical sensor planning problem, in that we seek to optimize some measure of viewpoint 'goodness'. However, we do not have an accurate and complete geometrical model of the target, and so such approaches cannot be used. The approach we took uses a coarse sort of visual servoing to find distinctive viewpoints. This is done by performing local maximization of 'signature distinctiveness' to home in on distinctive images. This has an advantage that it does not require explicit modelling of environment structure, only a heuristic notion of signature 'distinctiveness'. It is still nontrivial to develop such notions, and to show that they correspond to something like true distinctiveness. We shall do this below.

3.1 Distinctiveness

Generally speaking, for recognition we prefer signatures which rarely match others, since a signature will always match itself. Given a representative signature database, then, a signature's distinctiveness can be directly measured, by seeing how many signatures in the database match it (this is how ambiguous images were filtered out for the recognition experiments described above). For batched experimentation this is satisfactory; however it becomes impractical when (quasi-) real time performance is needed (as on an autonomous robot). Hence we need *distinctiveness metrics* that can be computed directly from an image signature; a *distinctiveness value* can then be used to classify images as distinctive or ambiguous. This seems a plausible goal since we, as humans, can usually easily tell whether or not an image is distinctive.

Given a distinctiveness metric, the locally 'best' view is chosen by searching for a local maximum of the metric. This idea is similar to that used by Kuipers and Byun for place definition [18], though our purposes are different. Here, we deal only with rotational search, but the approach can be extended to include translation as well.

Now, while the main motivation for using locally distinctive signatures is to improve recognition by filtering out ambiguous views, there are other benefits as well. Most obvious is the reduction in the size of the necessary signature database. Assuming that distinctive images are relatively rare, substantially fewer signatures will be stored for each place, since only the most distinctive views will be recorded. A further benefit that we would hope to obtain would be greater recognition stability, since active search should compensate for differences in original viewpoint. Such stability is not *a priori* obvious though, and must be demonstrated. We do so empirically, as described below in section 4.2.

3.2 Search

Since it is impractical to examine all viewpoints in a neighborhood for a distinctive image, some sort of search is required. In this paper, we only consider rotational search, so we specify a single search granularity, $\epsilon_V{}^\circ$, such that a view orientation within $\epsilon_V{}^\circ$ of a true distinctiveness maximum is acceptable. We mainly consider two search algorithms: a linear hillclimbing method and an interval subdivision method. The hillclimbing method works by sampling images every $\epsilon_V{}^\circ$ (always turning clockwise, say), until a three-image window is found whose middle image is a distinctiveness maximum. If the expected distance to a maximum is $\bar{V}^\circ$, this algorithm will examine $O(\frac{\bar{V}}{\epsilon_V})$ images.

For interval subdivision, we use a form of bisection line search, since no derivative information is available. Also, rather than assume that we have a maximum bracketed, we search for the local maximum within a predetermined interval. The algorithm is given a starting search size $V_0{}^\circ$. The starting interval is then that of size $V_0{}^\circ$ clockwise (say) of the starting orientation. The current interval is iteratively trisected[4], and the side containing the most distinctive signature is chosen as the next interval. This continues until the search interval is smaller than $\epsilon_V{}^\circ$. The trisection algorithm processes 2 images on each trisection, and performs $O(\log(\frac{V_0}{\epsilon_V}))$ trisections. Clearly, if $\bar{V}$ is known, choosing $V_0 = \bar{V}$ makes trisection more efficient than hillclimbing, but this is hard to do in general. However, in practice we found that trisection was more efficient with only a rough guess of a good V_0.

[4] Golden section search could be similarly used with a further efficiency improvement; trisection search was used due to ease of implementation. Also, some results seem to show that trisection is more stable, probably since more data are used.

3.3 Using multiple measurements

As demonstrated above, using multiple measurement functions can dramatically improve recognition. However, finding distinctive images based on multiple measurement types is not simple, since distinctiveness is not comparable across measurement types. Rather than try to guess at a combination function (which we have no principled method for), we make use of the fact that we are trying to maximize distinctiveness. This means that we only need a comparison relation giving a partial order (preferably nearly total). To compare two signature sets S_1 and S_2, we can count up the number of measurements for which each corresponding signature in S_1 has higher distinctiveness than that in S_2 and vice versa. The set with a higher count is preferred. This is extended for trisection search by also counting the number of distinctiveness preferences of each internal view to the interval endpoints. In general, preference comparison leads to the same results as distinctiveness comparison, though there are cases in which preferences cannot distinguish between possibilities. Note also that the preferred viewpoint may not be distinctive with respect to all measurements; a higher preference is still indicative of a better viewpoint.

3.4 Distinctiveness metrics

Due to the unpredictable complexity of the world, we designed distinctiveness metrics heuristically and evaluated them empirically. The basic idea is to ask what sorts of views tend to be ambiguous, and what signature features predict those views. We look at ambiguous views since they seem to be a more homogeneous class than distinctive views. Naturally, different measurement functions will require (in general) different distinctiveness metrics. Also, since views considered ambiguous in one environment may not be in another (consider a view of a tree in a forest, or in an office building), different environments may require different distinctiveness metrics. Discussion of the issues of operation in multiple environments is beyond the scope of this paper, however. The environment which we consider is the ubiquitous indoor office environment. This means that the main class of ambiguous images will be blank walls. Below we describe the distinctiveness metrics we developed for the measurement functions we developed (see section 2.1.1 above). Since relative intensity gave indifferent recognition results, it was omitted from consideration in this part of the work.

If we assume that most ambiguous views will be of blank walls (or others similarly homogeneous), we can design distinctiveness metrics for each measurement function to test for homogeneity[5]. We developed a number of possible metrics, only some of which checked out; due to space limitations we only discuss the good ones here. The simplest way to measure the featurelessness of a signature is to estimate its homogeneity. For real-valued measurements such as TEX and ZC, the homogeneity of a signature can be taken as inversely proportional to the variance of the signature array values, $\frac{1}{n^2}\sum_{i,j<n}(v_{ij}-\bar{v})^2$. Note that the signature variance metric (SigVar) is applicable to signature subarrays as well.

SigVar is not applicable to DEO, however, since its values are angles (and so variance is not well-defined). A similar idea can be applied, however, by calculating the estimated entropy of the signature values, taking them to come from a single probability distribution. It must be emphasized, though, that we only use this as a useful measure of homogeneity of values—if all values are the same, entropy is minimized; if they are equally distributed, entropy is maximised. If the possible values are indexed by i, and the fraction of signature elements with values i is p_i, the signature entropy (SigEnt) is computed as $\sum_i p_i \log p_i$.

As it turns out, neither of the metrics above works well for SGD. SigVar is again inapplicable due to SGD being angle-valued, and SigEnt has trouble, it seems, because lighting can cause

[5]Note that signatures are homogeneous, not image intensity. For example, wallpaper with a repetitive but highly textured pattern would be judged to be homogeneous, since it is self similar on a large scale, despite being visually interesting.

noticable contrast variations over large areas of homogeneous regions. However, a metric that does give satisfactory results for SGD is the fraction of array elements for which a significant gradient was registered. When there is any significant environmental structure, a significant intensity gradient will usually be generated; empirically, the converse seems to be true as well. Hence, the noticable gradient metric (NoGrad) will usually be low only for ambiguous featureless views.

4 Results

4.1 Place recognition

We conducted experiments on image matching using the measurement functions described above in section 2.1.1. Images were taken by a CCD camera with a 25mm lens, giving slightly more than 12 degrees of view. The images were reduced by pixel averaging from 640×480 to 60×45 before signatures were computed. The bulk of the experiments described here were performed on a corpus of 168 images taken from 8 different locations in our building. At each place a sequence of 21 images was taken, with a rotation of around 6 degrees between adjacent images. Thus a total of about 130 degrees were covered at each place. Note that adjacent images overlap by about half and non-adjacent images barely, if at all. This gives us an easy way to automatically check ground truth for matching.

Except as otherwise noted below, the system parameters were as follows. Signatures were 8×8 grids. Mismatches were generally disallowed. Matching threshholds were: $\theta_{\text{pat}}^{\text{DEO}} = 0.6$, $\theta_{\text{pat}}^{\text{SGD}} = 0.6$, $\theta_{\text{pat}}^{\text{ZC}} = 0.5$, $\theta_{\text{rms}}^{\text{ESTR}} = 25$, $\theta_{\text{rms}}^{\text{TEX}} = 0.5$, and $\theta_{\text{rms}}^{\text{RI}} = 0.1$. Measurement parameters were set as $\theta_{\text{grad}}^{\text{DEO}} = 70$, $\theta_{\text{der}}^{\text{SGD}} = 50$, and $\theta_{\text{diff}}^{\text{TEX}} = 0.05$. When filtered, ambiguous images were taken to be those matching more than 5% of the database.

4.1.1 Measurement evaluation

Stability

We first evaluate individual measurements for their stability. Stability was estimated separately for rotation and translation. Rotational stability was evaluated by plotting the match similarity change with rotation at intervals of about 2 degrees. Translation was evaluated similarly for forward and transverse movement at intervals of about 5cm. The image sequences are shown in figure 2. We calculated both full signature similarity and best offset similarity between the first image and all other images in each series. The similarity is plotted against image number in figures 3, 4, and 5. The solid lines represent matching without offset, the dashed lines matching at best offset, and the thick horizontal lines show the matching threshhold we use.

Most of the plots show a reasonable degree of stability for non-offset matching; the exception are the rotation matches, which is understandable since a 2 degree rotation is an offset of just over one grid cell (at 8×8). The usefulness of offset matching is immediately apparent; most measures allow offset matching to a distance of about 25cm forward, 35cm sideways, and 9 degrees of rotation. The main exception is relative intensity, which doesn't match at all for any rotation. It appear that this is due to the greatly changing global illumination in that sequence; it confuses RI's normalization. In our experience, these plots seems representative of the measurements' behavior. The stability region implied, 25cm×35cm, is probably made smaller by the fact that our camera has a fairly narrow field of view.

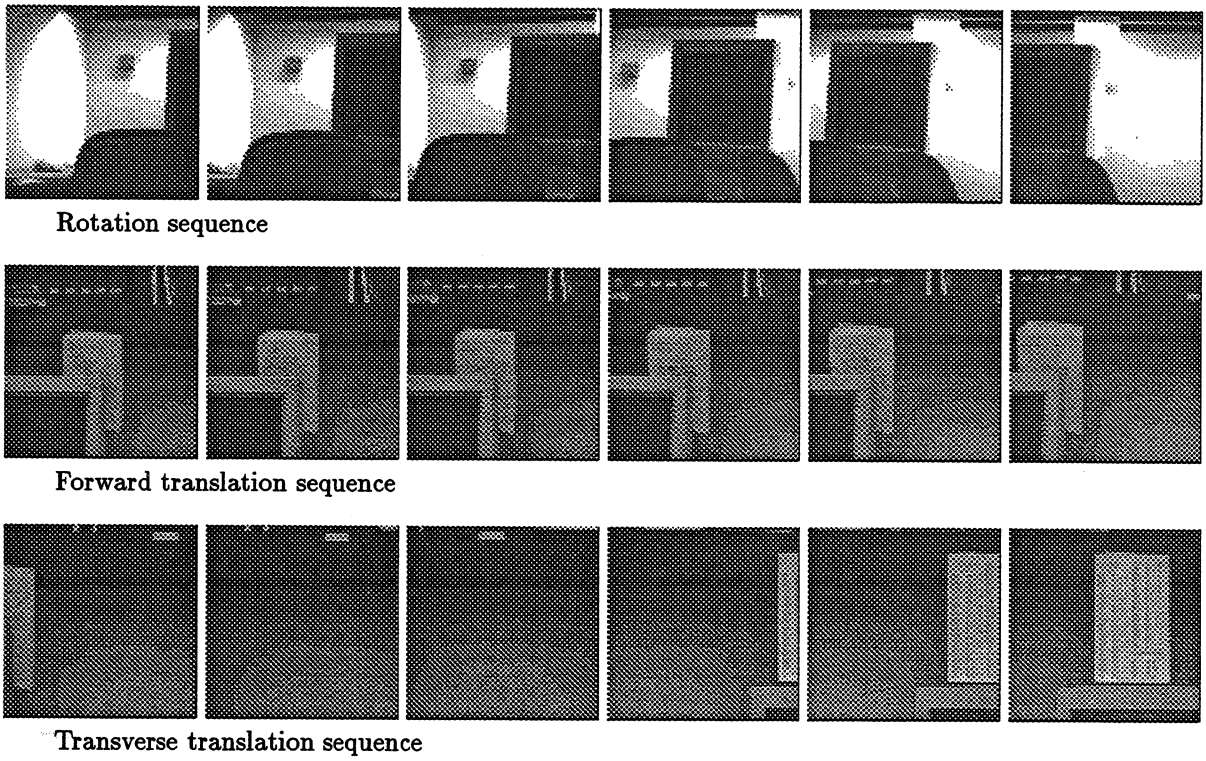


Figure 2: Image sequences for measurement stability evaluation (every second image shown).

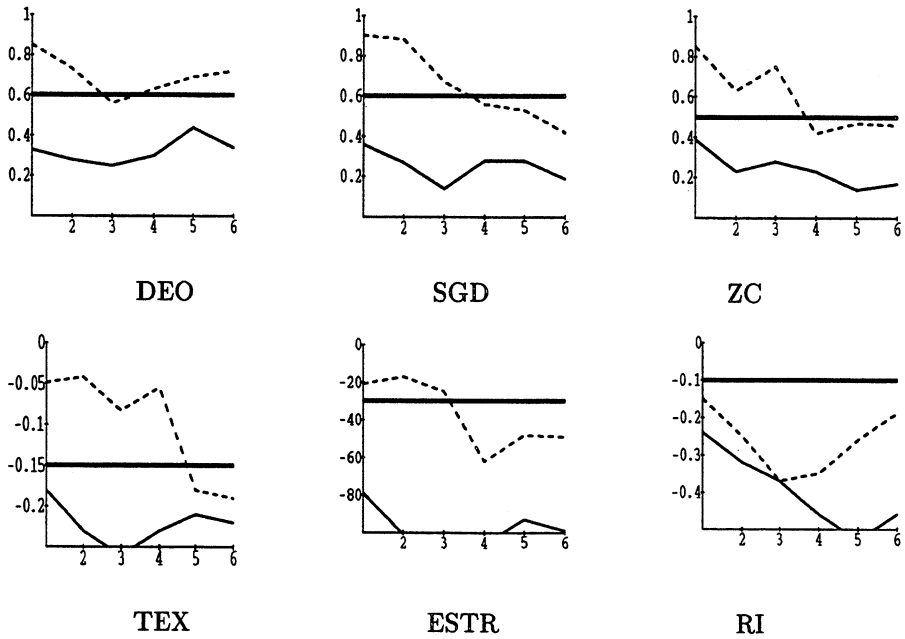


Figure 3: Rotation sequence stability plots. Signature similarity is plotted versus angular distance (in units of 2°).

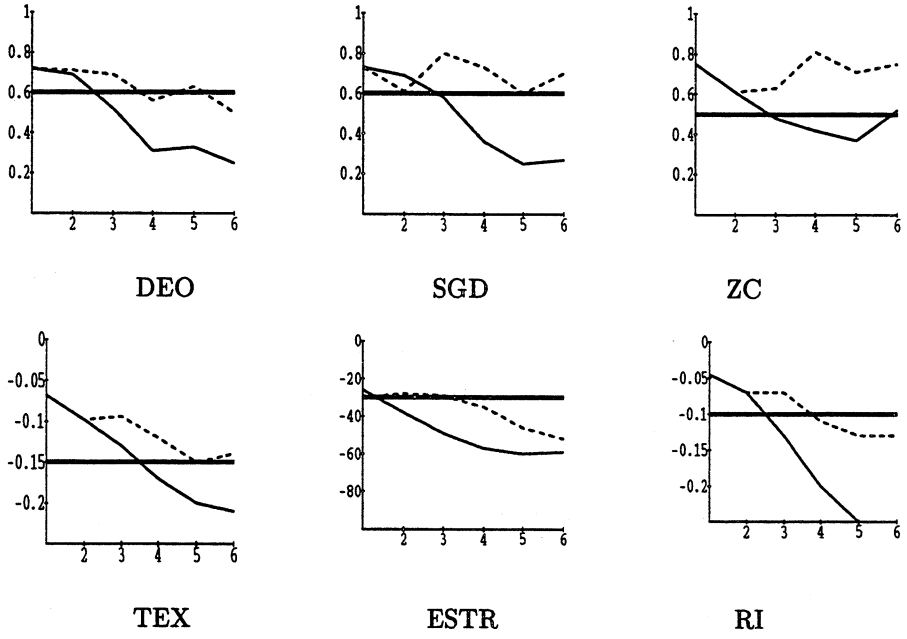


Figure 4: Forward sequence stability plots. Signature similarity is plotted versus distance (in units of 5cm).

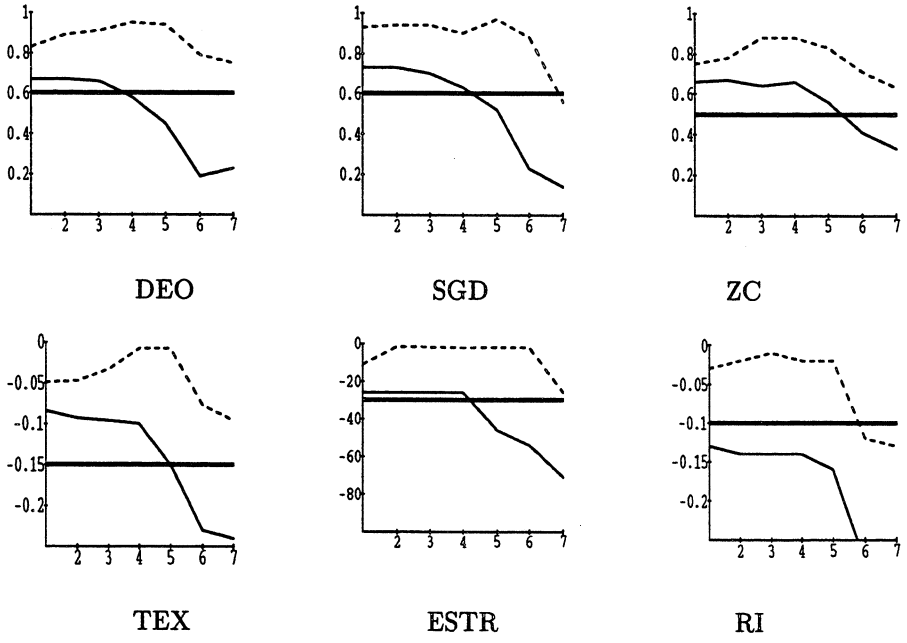


Figure 5: Transverse sequence stability plots. Signature similarity is plotted versus distance (in units of 5cm).

Measurement(s)	Average number of matches
DEO	1.66
SGD	2.26
ZC	3.43
TEX	3.39
ESTR	1.68
RI	2.74
DEO,SGD	2.08
DEO,TEX	2.19
TEX,ESTR	1.59
ZC,RI	2.14
DEO,SGD,TEX,ESTR,ZC	1.74

Table 1: Estimated measurement function distinguishability.

Distinguishability

The distinguishability of measurement functions was evaluated, based on our image corpus, by computing, for each measurement function, the average number of matches found per image for which a match is found. Since there are two correct matches for nearly every image in the corpus (its predecessor and successor), we should expect this average to be something less than 2. The greater the number, the less distinguishable the measurement. To avoid being confounded by inherently ambiguous images, this average was computed while filtering ambiguities out. Table 1 summarizes our results; we also evaluated some sets of measurement functions working in tandem. The most distinguishable measurements seem to be DEO and ESTR, and the least ZC and TEX. As expected, distinguishability increases when measurement functions are combined.

Independence

Recall that in section 2.3.1 we came up with a way of evaluating the dependence of sets of measurement functions in the context of signature matching. Since efficiency dictates that as few measurements should be used as possible, effort should be taken to choose a reasonably independent set. We evaluated pairs of measurement functions described above in terms of the approximate dependence formula derived above. In our case, the numerator counts the number of non-adjacent image pairs in our sample which matched according to both measurements being evaluated, and the denominator those which matched one of the two. Table 2 shows the results of applying this dependency metric to pairs of measurement functions over the image corpus described above.

Some of the results in the table immediately make sense. For example, the high degree of dependence between TEX and ZC—since TEX measures texture by large intensity variations between neighbors, it is to be expected that those same pixels would tend to neighbor on Laplacian zero-crossings. Some results are somewhat surprising, though. The high degree of dependence between RI and ESTR seems strange at first; but it is probably due to the fact that image regions with high edge strength will have large amounts of intensity variation, and hence the average intensity will tend to not be extreme. This would tend to correlate the two measures. A similar argument explains the lesser, though significant, dependence of RI with DEO and SGD. The result that is perhaps most surprising is the lack of dependence between DEO and SGD. Since these were both intended, in a way, to measure 'edge direction', one would

	SGD	ESTR	ZC	TEX	RI
DEO	0.08	0.10	0.05	0.06	0.17
SGD	–	0.17	0.05	0.06	0.10
ESTR	0.17	–	0.10	0.08	0.22
ZC	0.05	0.10	–	0.26	0.09
TEX	0.06	0.08	0.26	–	0.05
RI	0.10	0.22	0.09	0.05	–

Table 2: Matching dependencies for pairs of measurement functions.

expect their matches to be highly correlated. However, since SGD does not extract edge tokens, but merely looks at high-magnitude gradients, it also picks up curved and specular surfaces, as well as surfaces near a light source. These results support our dependence measure.

4.1.2 Recognition

Finally, we tested the recognition capabilities of the system on our image corpus. The images were each processed to produce a set of 6 8×8 signatures, one for each combination of image and measurement function. The following experimental procedure was performed twice, once as is, and again filtering out ambiguous images (those matching more than 5% of the corpus). For different combinations of measurement functions, each image's signatures were matched at all offsets, up to 3/4 of the signature, against all other images' signatures. Several statistics were accumulated:

- The number of ambiguous images found (presented above),
- The number of images for which a match was found,
- The average number of matches per image with one (presented above),
- The number of images for which the highest-ranked match was correct,
- The number of images for which a correct match was among the 4 highest ranked. This gives us an idea of how well we could do with geometric and contextual filtering.

The results for various combinations of measurement functions are summarized in Table 3.

The first thing to note is that we can get accuracy of around 90% when filtering ambiguities. When combined with geometric and contextual matching in a full mapping system, we can expect near-perfect accuracy. Also, most of the time adding more measurement functions increases accuracy. In those few places where percentage accuracy decreases, this is because fewer images are being classified as ambiguous, and the raw number of correct matches does increase. The relatively poor results that we got without ambiguity filtering are mainly due to the high degree of inherent ambiguity of some of our images—the average number of matches found for each image was typically over 40 without ambiguity filtering. We also note that distinguishability seems to correlate with performance without ambiguity filtering, as we would expect. For example, ZC is the least distinguishable, and it gives the worst results. On the other hand, DEO and ESTR, which are much more distinguishable, give noticeably better results. Thus, these empirical results also support our evaluation metrics.

	Ambiguity filtered			Not ambiguity filtered		
Measures	M	BC	4C	M	BC	4C
DEO	41 (24%)	20 (49%)	24 (59%)	158 (94%)	18 (11%)	58 (39%)
SGD	39 (23%)	26 (67%)	30 (77%)	166 (99%)	53 (32%)	66 (40%)
ZC	14 (8%)	8 (57%)	9 (64%)	168 (100%)	3 (2%)	9 (5%)
TEX	23 (14%)	8 (35%)	21 (91%)	168 (100%)	13 (8%)	24 (14%)
ESTR	25 (15%)	21 (84%)	24 (96%)	163 (97%)	24 (15%)	30 (18%)
RI	38 (23%)	17 (45%)	24 (63%)	165 (98%)	13 (8%)	27 (16%)
DEO, SGD	51 (30%)	39 (76%)	46 (90%)	139 (83%)	49 (35%)	62 (45%)
DEO, TEX	63 (38%)	39 (62%)	51 (81%)	148 (88%)	39 (26%)	39 (26%)
TEX, ESTR	29 (17%)	26 (90%)	28 (97%)	163 (97%)	28 (17%)	33 (20%)
ZC, RI	44 (26%)	25 (57%)	29 (66%)	159 (95%)	26 (16%)	34 (21%)
DEO, SGD, TEX, ESTR, ZC	54 (32%)	43 (80%)	50 (93%)	128 (76%)	47 (37%)	61 (48%)
DEO, SGD, TEX, ESTR, ZC, RI	50 (30%)	37 (74%)	45 (90%)	116 (69%)	37 (32%)	49 (42%)

Table 3: Recognition statistics both with and without ambiguity filtering. M = images for which matches were found, BC = preferred matches which were correct, and 4C = correct match in the top 4. With ambiguity filtering, ambiguous images were neither checked for matches nor matched against. Percentages shown in parentheses are as follows: images matched vs. images checked, best match correct vs. images matched, correct match in top 4 vs. images matched.

4.2 Distinctiveness search

We performed a number of experiments to evaluate the performance of the active recognition methods described above. Experiments were performed using a CCD camera at a height of 110cm. Images were reduced by pixel averaging from 640×480 to 60×45 before signatures were computed. Some of the experiments below use a corpus of 291 images (a superset of that used in section 4.1) taken at 11 different positions in our building using a 25mm lens. Those experiments not involving the image corpus used an 8mm lens, for a larger field of view. At each position a set of images was taken with the camera at different orientations. Some others of our experiments were performed using a pan-tilt platform in our laboratory. A drawing of the laboratory layout during the experiments is shown in Figure 6. The robot arm in the upper left and the electronics workbench were the most visually interesting things in the room.

Due to the difficulty of performing exhaustive and systematic tests using a mobile platform, two sets of experiments were performed to evaluate the signature distinctiveness method. The first are static evaluations, using the image corpus and rotational image sequences to evaluate ambiguity prediction and density of distinctiveness maxima. A second series of experiments was performed using a pan-tilt head to perform rotational distinctiveness maximization, and the results were evaluated. The details of the experiments and analysis are described below.

4.2.1 Distinctiveness and ambiguity

Due to the difficulty of obtaining large quantities of representative data using dynamic, on-line techniques, we evaluated the use of distinctiveness metrics for predicting ambiguity statically, using the image corpus. This gave us a large, reasonably representative data set (for our office building). For each measurement function, we computed signatures at several resolutions from each image in the corpus. For each signature we then found all other signatures that matched it, and computed the signature's distinctiveness using the appropriate metric. This then gave

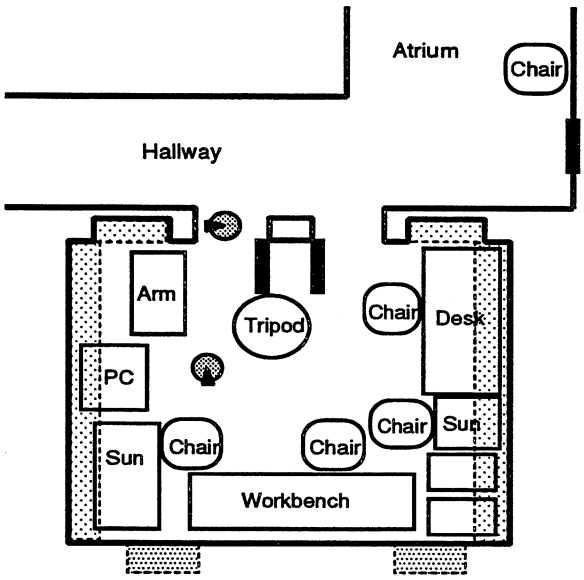


Figure 6: Approximate floor-plan of our laboratory. The two camera positions ("Door" and "InLab") used for rotational experiments are shown as grey circles, with the zero points for the static stability experiments (section 4.2.2) shown as black wedges.

us a pair of numbers for each image i, a number of matching images M_i and a calculated 'distinctiveness' D_i.

The task is then to evaluate the hypothesis that local distinctiveness maxima are probably also unambiguous. The data for 8×8 signatures are plotted in Figure 7. From the shape of the plots it is clear that there is a strong relationship between the number of images matched (ie, ambiguity) and distinctiveness values. However, to properly evaluate the results, a more precise evaluation of the data is required. Due to the impossibility of modelling, even approximately, the underlying environmental structure that gave rise to the observations, we approach this problem indirectly. We first established a threshhold θ on the number of images matched for calling an image ambiguous or not; this threshhold was 5% of the database (14.6 images). We can then think of the distinctiveness values for 'ambiguous' images ($M_i \geq 15$) and those for 'unambiguous' images ($M_i \leq 14$) as two random variables, call them A and U. We then wish to compare the two distributions to evaluate the reasonableness of our hypothesis. We did this using four tests. The first two provide an intuitive feel for the overall properties of the data, and the latter two statistically test the hypothesis that distinctiveness values are unrelated to ambiguity.

- We calculate the ratio of the sample means $\bar{U}/\bar{A}$; if it is far from 1, then on average unambiguous images have distinctiveness noticeably different from ambiguous images.
- We estimate the probability of a random unambiguous image having greater distinctiveness than a random ambiguous image, $Pr(D_i > D_j \mid M_i < \theta \wedge M_j > \theta)$. This is evaluated as the fraction of image pairs with one unambiguous and the other ambiguous, satisfying the former condition. If this probability is high (say, >90%), a strong argument can be made that unambiguous images have significantly greater distinctiveness than ambiguous ones.
- We test the hypothesis that A and U come from distributions with the same median (which we would expect if distinctiveness was unrelated to ambiguity). Since we have no

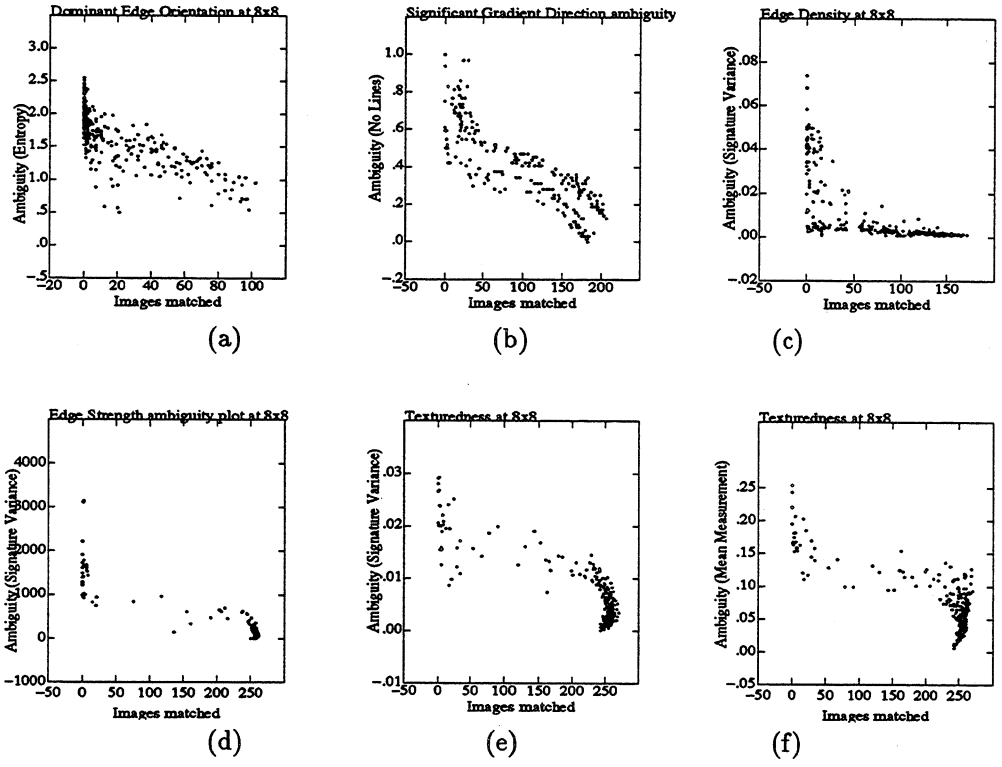


Figure 7: Representative plots of distinctiveness estimates (y axis) versus number of database images matched (x axis), using 8×8 signatures. (a) DEO signatures using the SigEnt metric. (b) SGD signatures using NoGrad. (c) ZC signatures using SigVar. (d) ESTR signatures using SigVar (e) TEX signatures using SigVar. (f) TEX signatures using MeanVal; note the similarity to (e).

good statistical model for the data, we use the Mann-Whitney median U non-parametric test (described in [11]). It uses rank sums to test the hypothesis that two samples of different sizes come from distributions with the same median. The procedure provides a z score that can be checked for significance against tables of the normal distribution. If the hypothesis of equal medians can be rejected at a reasonable level of significance (say, 1%), then we can conclude that a distinctiveness metric operates differently on ambiguous and unambiguous images.

- Finally, we use the Kruskal-Wallis H test to test the hypothesis that the two samples of distinctiveness values come from the same population. The H test tests whether sample rank sums are equidistributed as would be expected if the samples came from the same population. The procedure provides a number, H, which is used to perform a χ^2 significance test with 1 degree of freedom (since we have two samples). Again, if the hypothesis (of a single population) can be rejected with reasonable significance, we can say that distinctiveness and ambiguity are related.

While no one of these tests is sufficient to establish our main hypothesis, if a distinctiveness metric satisfies all of them, the evidence is strongly suggestive that distinctiveness maxima are generally unambiguous. The results of these tests for the various measurement functions/distinctiveness metrics at signature resolutions of 5, 6, and 8 are given in Table 4.

Measure	Res.	R	P	U sig.	H sig.	Measure	Res.	R	P	U sig.	H sig.
DEO	5×5	1.7	0.94	0.01	0.01	ESTR	5×5	39	1.0	0.01	0.01
(SigEnt)	6×6	1.6	0.93	0.01	0.01	(SigVar)	6×6	19	0.99	0.01	0.01
	8×8	1.5	0.93	0.01	0.01		8×8	13	0.99	0.01	0.01
SGD	5×5	1.3	0.66	0.01	0.01	TEX	5×5	5.3	0.99	0.01	0.01
(NoGrad)	6×6	1.6	0.74	0.01	0.01	(SigVar)	6×6	5.2	0.99	0.01	0.01
	8×8	1.9	0.87	0.01	0.01		8×8	4.1	0.99	0.01	0.01
ZC	5×5	8.9	0.94	0.01	0.01	TEX	5×5	3.2	0.99	0.01	0.01
(SigVar)	6×6	9.3	0.94	0.01	0.01	(MeanVal)	6×6	6.7	0.99	0.01	0.01
	8×8	7.5	0.94	0.01	0.01		8×8	3.0	0.99	0.01	0.01

Table 4: Summary of hypothesis test results on the image corpus. R is the ratio of distinctiveness means; $P = Pr(D_i > D_j \mid M_i < \theta \wedge M_j > \theta)$; U significance is the probability of erroneously rejecting the hypothesis of equal medians, and H significance of the hypothesis of the same population. See text for a more detailed description of the tests.

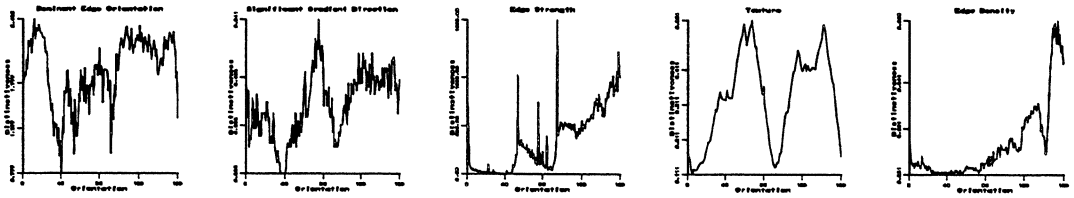
When the results are examined, we see that the metrics presented here satisfy all the tests, confirming the impressions gotten from the plots in Figure 7. The larger distinctiveness mean is at least 1.3 times the smaller, which indicates that a difference is noticeable. For all measurement functions but SGD, $P > 0.9$ which indicates that it is overwhelmingly probable that unambiguous views can be distinguished from ambiguous ones. For SGD, the resolution of the signatures seems to play a significant role; one explanation is that at lower resolution, the regions that are averaged over are larger, and hence more likely to contain significant contrast edges, obscuring the existence of large-ish regions with little contrast. In any case, while $P > 0.7$ is not overwhelmingly large, it may still be reasonable if SGD is used with other measurement functions.

The significance of the median and population tests is quite striking. So striking, in fact, that the figures were cross-checked in two ways: a random partitioning of the data into 'pseudo-ambiguous' and 'pseudo-unambiguous' was used and the tests were performed, and the results of the tests on non-meaningful distinctiveness metrics (eg, SigVar for DEO) were examined. Random partitioning produced, as would be expected, no hypothesis rejection at any reasonable significance level. Non-meaningful distinctiveness metrics occasionally checked out on one of the two tests, but (a) not as significantly as those we accepted, (b) almost never passed both tests, and (c) never passed all four tests. This leads us to conclude that the distinctiveness metrics presented here are good predictors of scene ambiguity and hence, local distinctiveness maxima should be unambiguous. This claim is tested more directly below.

4.2.2 Stability

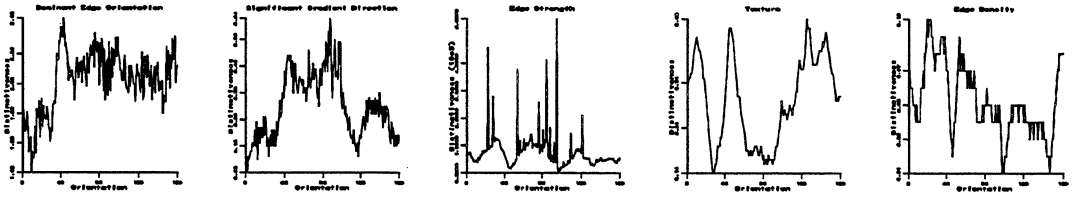
One way to evaluate the inherent stability of a distinctiveness metric is to plot distinctiveness versus viewpoint. If large maxima exist, then it can be assumed that stability can be achieved. Also, such plots help us get a feel for what distinctiveness actually measures. Plots for two positions in our laboratory are shown in Figures 8 and 9. The first salient feature of these plots is that they form large, wide humps on a large scale, implying large stability regions. This is mostly not the case for DEO, which is largely constant, but since it doesn't jump about randomly it works well in conjunction with other measurements. Another feature of these graphs is that the locations of their maxima correlate well across different measurement types. These two facts lead us to conclude that our distinctiveness metrics contain enough information for viewpoint stabilization.

If we look more closely at the plots and refer back to Figure 6, we can make correspondences between computed distinctiveness maxima and the physical contents of our lab. First, consider



(a) (b) (c) (d) (e)

Figure 8: Distinctiveness (y-axis) plotted against camera orientation (x-axis) from the lab doorway. (a) Signature entropy of DEO. (b) Noticable gradient on SGD. (c) Signature variance of ESTR. (d) Signature variance of TEX. (e) Signature variance of ZC.



(a) (b) (c) (d) (e)

Figure 9: Distinctiveness (y-axis) plotted against camera orientation (x-axis) from the interior of our lab. (a) Signature entropy of DEO. (b) Noticable gradient on SGD. (c) Signature variance of ESTR. (d) Signature variance of TEX. (e) Signature variance of ZC (jagginess due to plot quantization).

the doorway camera position. There are two large maxima here: at about 160° and at about 350°. The first corresponds to the camera looking towards the chair and door in the atrium, a region of considerable interest compared to the blank walls surrounding it. The second has the camera looking at the robot arm, also a distinctive feature of the lab. The interior position gives, as we would expect, more distinctive viewpoints. We can pick out several: at 10°, at 100°, at 190°, and at 300°. The first corresponds to the Sun workstation[6] and workbench in the lower left. The second is the robot arm again. The third is a combination of the open door and the camera tripod. The last is the constellation of chairs and workstations in the lower right. Thus we see that signature distinctiveness corresponds quite well to our intuitive notions of 'distinctive views'.

4.2.3 Dynamic experiments

We conducted experiments to measure several features of the distinctiveness maximizing algorithm described above in Section 3.2. The camera was mounted on an experimental pan-tilt platform (manufactured by Zebra Robotics). The platform was positioned about 1 meter off the ground. For all experiments but one (as noted below), an 8mm lens was used. Two main positions were used, one in the middle of our lab (InLab) and the other in one of the lab doorways (Door). Most of the experiments consist of a rotational test suite, each test performed starting with a different camera orientation, at 15° intervals. The search parameters were $V_0 = 100°$ and $\epsilon_V = 5°$.

[6] Sun is a registered trademark of Sun Microsystems Incorporated.

InLab					Door				
DEO	SGD ($\times 10^2$)	ESTR	TEX ($\times 10^2$)	ZC ($\times 10^3$)	DEO	SGD ($\times 10^2$)	ESTR	TEX ($\times 10^2$)	ZC ($\times 10^3$)
2.1	31	911	32	24	1.7	14	1129	25	25
2.0	50	277	23	25	1.8	22	1305	32	17
2.1	55	434	22	8	1.8	25	1421	31	17
2.1	47	1060	27	32	1.9	2	58	10	4
2.0	25	1031	33	32	1.9	3	83	8	3
1.8	27	1455	36	21	1.9	11	56	16	3
2.2	31	861	30	27	1.9	13	90	20	4
2.1	34	999	38	13	2.1	17	101	17	5
1.7	27	1203	32	29	1.7	28	104	15	5
2.0	12	1259	27	40	1.9	8	72	23	5
1.9	23	1146	25	33	2.3	26	467	20	8
2.2	34	1422	26	13	2.2	30	622	25	7
2.2	36	1291	28	14	2.3	34	692	26	9
					2.3	30	986	37	6
					2.3	38	1108	41	6
Average					**Average**				
2.1	35	1045	29	24	2.0	20	552	22	11

Table 5: Results of the rotation test experiment for InLab and Door, using trisection. The calculated distinctiveness of attractor views are shown for the five measurement functions. Averages were computed over all starting orientations.

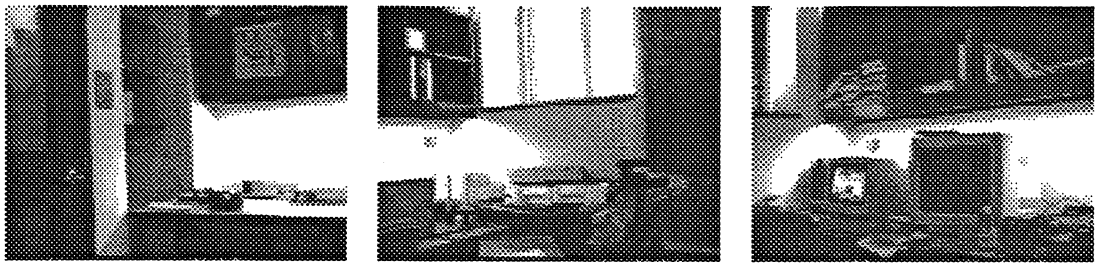
Distinctiveness and rotational stability

We first checked the distinctiveness values of the views found by trisection search. At each of the test positions the algorithm, using all five measurement functions, was used to find a local distinctiveness maximum. We then recorded the distinctiveness of the final viewpoint (an *attractive viewpoint*) with respect to all five measures. This was repeated for all starting orientations 15° apart.

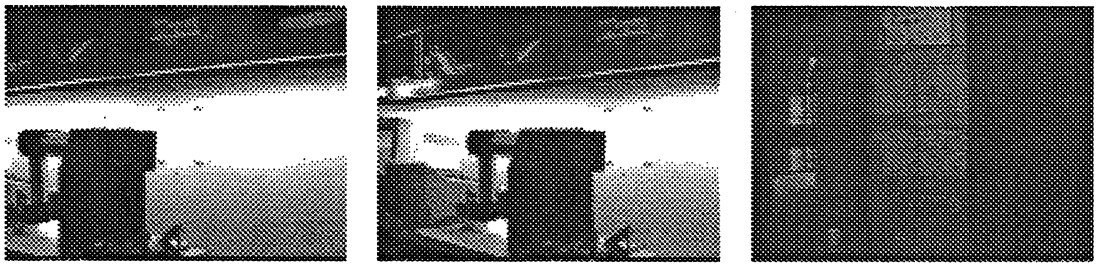
A similar method was used to evaluate rotational stability. The same procedures were used, and the final orientations (*attractors*) after each search were recorded. If the method is rotationally stable, we expect the camera to home in on one of a small set of distinctiveness maxima, regardless of starting orientation. For our purposes, if two final orientations were within ϵ_V of each other, they were considered the same. We measure stability in two ways. First, we look at the total number of attractors found (A); second, we check the minimum number of attractors needed to account for half of the starting orientations (M). For a very stable method, both of these numbers should be small, relative to the number of tests run (36). But if M is small,

Search		InLab	Door
Trisection	Attractors	13	15
	Main attractors	4	5
	Average steps	9	9
Hillclimbing	Attractors	14	16
	Main attractors	4	4
	Average steps	12	11

Table 6: Rotational stability and efficiency results. Main attractors are the minimal set attracting most viewpoints.



InLab



Door

Figure 10: The images at the three most frequent trisection-attractive viewpoints for InLab and Door in the rotation test.

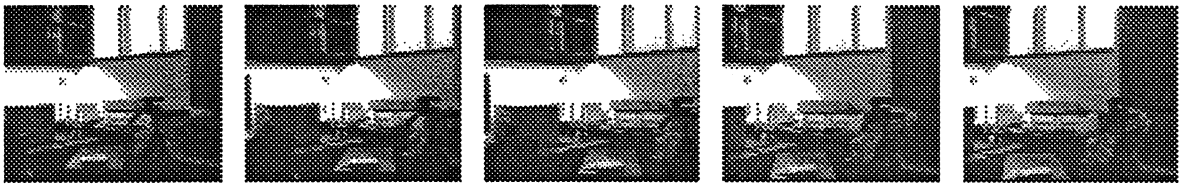
then even if a few starting orientations lead to anomalous attractors, the method can still be considered usually stable.

Tables 5 and 6 summarize the results of the distinctiveness and rotational stability tests. The stability results show a reasonable level of stability, in that the majority of starting orientations go to one of 4 or 5 attractors. On the other hand, the total number of attractors is over a third of the number of tests performed. This can probably be attributed to two features of our testing methods. First, only attractive orientations within 10° were considered the same. However, with the camera's 36° field of view, viewpoints up to 18° apart will be matchable (using offset matching, see section 2.2.2); the 'within ϵ_V' criterion, while justifiable, is perhaps too strict. Second, we would expect larger values of V_0 and smaller values of ϵ_V to produce more stable attractors, since (a) it is less likely for an attractor to be cut off from the initial interval, and (b) local maxima are found more precisely. There is clearly a tradeoff between search time required and quality of the solution.

In terms of distinctiveness, the first feature of the data which is noticed is the significant difference between distinctiveness of InLab and Door attractors. This can be explained by the fact that the doorway looks out on a hallway, thus many of the attractors (being *local* maxima) are non-descript views of the hall. If the values are compared against the plots in Figure 7, the distinctiveness of most attractors are clearly reasonable, and often excellent. This argues that the local maxima found by our method tend to be truly distinctive. Another thing to notice is that the distinctiveness measures are sometimes complementary, in that attractors distinctive in, say, SGD, are less distinctive in ESTR, and vice versa.

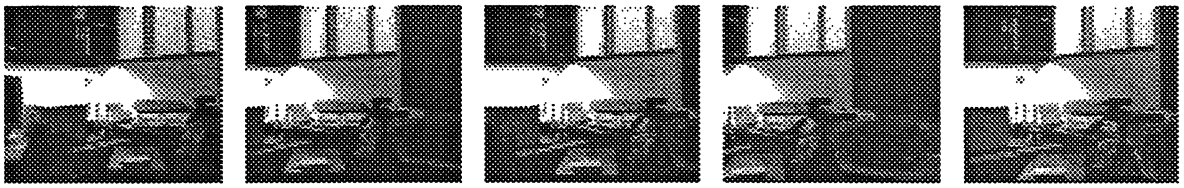
Search	Direction	0cm	10cm	20cm	30cm	40cm
Trisection	Forward	46°	47°	48°	60°	62°
	Transverse	47°	61°	48°	56°	45°
Hillclimbing	Forward	65°	50°	-161°	75°	—
	Transverse	95°	30°	30°	180°	—

Table 7: Orientations of the attractors in the translational stability test.



0cm 10cm 20cm 30cm 40cm

(a)



0cm 10cm 20cm 30cm 40cm

(b)

Figure 11: Images of the attractive viewpoints for the translational stability test, using trisection. (a) Forward sequence. (b) Transverse sequence.

Translational stability

The other sort of stability is with respect to viewpoint translation. We tested stability for forward and transverse translation (with respect to the starting viewing direction) separately. For each direction, we ran the trisection algorithm at 5 positions at 10cm intervals. A translationally stable method should give rise to similar attractive viewpoints and matchable views at different positions, in the same way as rotational stability implies for different angles. The results are summarized in Table 7. The data show a fair measure of attractor stability for trisection over a translation range of 40cm. While two attractors seem to have emerged, looking at Table 7, the strictness of the 'same attractor' criterion can be seen to have an effect, as the images in Figure 11 show. The images in each sequence, while not identical, show nearly the same scene; nearly enough to match at a signature offset. Thus trisection appears to produce reasonable translational stability. However, hillclimbing does not fare nearly as well, as is clear from Table 7. This is because the method uses a narrow window for deciding when a point is a maximum, and hence is more sensitive to low-amplitude distinctiveness oscillation (visible in Figures 8 and 9). This means that a small change in starting orientation can cause the method to miss a previously-found maximum. This reflects a more general problem, that of *distinctiveness smoothing*. Ideally, we would like to perform a low-pass filter on distinctiveness to produce a truly stable method; however, this is difficult to achieve on-line. This problem definitely requires further study.

	Maximum		Average		Not amb.
Starting	110	38%	20	7%	14
Attractive	21	7%	2.7	1%	17

Table 8: Results of the ambiguity test. Shown are the numbers of images matched in the corpus by images at all starting orientations (at 20° intervals) and at the resultant attractive orientations. We show the maximum number of images matched in each set, the average over the set, and the number of views in the set (18 images) not judged ambiguous (by the $< 5\%$ criterion). We also give the percentage of the image corpus (291 images) matched.

Ambiguity

We also tested trisection attractors' ambiguity directly, using the image corpus. The camera was configured with a 25mm lens and set at a height of 110cm (to match the corpus). The camera on the pan-tilt head was placed at Door, which was also a position used in the generation of the image corpus. A rotational test suite at intervals of 20° was performed, and the attractive viewpoints were matched to the corpus, using all five measurement functions, to directly evaluate their ambiguity. The results are given in Table 8. The difference is quite striking; the average regular viewpoint is ambiguous, while the average attractor is not. In fact, only one attractor was judged to be ambiguous; at least one is to be expected in a doorway, as noted above. This reinforces the conclusion in section 4.2.1 that our distinctiveness metrics are good predictors of ambiguity.

5 Discussion

Image-based place recognition has several benefits over reconstructionist approaches. In particular, the method of image signatures is simple and easy to apply, generally applicable to all sorts of environments, computationally inexpensive, and is easily integrated with other recognition cues. When used with appropriate measurement functions, image signatures support image-based recognition over a range of viewer positions and orientations. The qualitative performance of different measurement functions can be evaluated by the twin metrics of stability and distinguishability, so that appropriate measures can be found for particular environments. Signatures derived from different measurements can be used in combination for matching, to improve correctness; ambiguity filtering can also be applied.

We have also identified a difficulty for any image-based recognition framework—the *recognition problem*, which consists of two coupled subproblems: the ambiguity problem and the stability problem. We deal with this by problem using viewpoint motion. Scene ambiguity, at least in our indoor environment, appears to be reasonably well predictable from image structure by simple distinctiveness metrics. Using such metrics, a local search procedure can be used to efficiently find locally distinctive viewpoints. Most of the time, these locally distinctive viewpoints are also unambiguous, and hence useful for recognition. It also turns out that there are not many such attractive viewpoints, and so the method ensures a certain amount of viewpoint stability. This improves recognition in that the same viewpoint will usually be found each time a place is visited, easing the necessity of storing very large numbers of image signatures.

Future work

One area which needs to be addressed in the future is the development of a comprehensive basis for measurement function design, particularly in the analysis of measurement stability,

distinguishability, and independence. More principled methods of combining information from different measurement types would also be desirable. Decision theoretic notions may be useful for trading off the cost of matching against multiple measurements and the gain accrued.

Our methods have not yet been applied directly to mobile robot mapping and place recognition. In the future, we intend to integrate signature-based place recognition with robust map-learning [9] on a mobile robot. In addition to place recognition, we believe that image signatures can be used for image-based homing, as a generalization of Nelson's work [23]. If the robot can recognize a place from afar, it should be able to use perceptual feedback to home to it. In practice, efficient hypothesis generation is crucial, so effort needs to be put into developing efficient indexing methods for finding plausible matches.

There are also some important problems here which we have not yet addressed. Chief among these is the fact that image-based place recognition assumes that the world is visually unchanging. There is some robustness to small changes through coarse representation and imprecise matching, but the real world does change on a large scale. The best way in which to address this problem is to integrate different forms of perceptual representation with image signatures. If this is done, then inconsistencies between different representations of a scene indicates things that have changed; the system can then ignore or compensate for the changed portion of the scene. Another area for further work is extension of the method to deal with multiple environments. The measurement functions we developed are appropriate for some indoor environments (particularly office buildings), but will probably not be as useful in outdoor environments or even very different indoor environments (eg, malls). Hence there must be work done on developing a larger set of measurement functions, and determination of the conditions under which each is applicable. In addition, we would like a truly autonomous robot to be able to decide which measurement functions to use in a given situation, based on its sensor readings. This may be doable by using a comparative analysis of the signatures produced by different measurement functions.

Acknowledgements

Discussions with Drew McDermott, Greg Hager, Michael Beetz, and Michael Black helped greatly in the development of these ideas and this paper.

References

[1] Dana H. Ballard. Reference frames for animate vision. In *Proc. Int'l Joint Conference on Artificial Intelligence*, 1989.

[2] Dana H. Ballard and Christopher M. Brown. *Computer Vision.* Prentice-Hall, 1982.

[3] Dana H. Ballard and Altan Ozcandarli. Eye fixation and early vision: Kinetic depth. In *Proc. Int'l Conf. on Computer Vision*, 1988.

[4] James O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, 1985.

[5] Michael J. Black. *Robust Incremental Optical Flow.* PhD thesis, Yale University, September 1992. Technical Report 923.

[6] David J. Braunegg. *MARVEL: A System for Recognizing World Locations with Stereo Vision.* PhD thesis, MIT, 1990.

[7] Ernest Davis. *Representing and Acquiring Geographic Knowledge.* PhD thesis, Yale University Department of Computer Science, 1984.

[8] Sean P. Engelson. Active place recognition using image signatures. In *Proceedings of SPIE Symposium on Intelligent Robotic Systems, Sensor Fusion V*, 1992.

[9] Sean P. Engelson and Drew McDermott. Error correction in mobile robot map learning. In *Proc. Int'l Conf. on Robotics and Automation*, Nice, France, May 1992.

[10] Sean P. Engelson and Drew V. McDermott. Image signatures for place recognition and map construction. In *Proceedings of SPIE Symposium on Intelligent Robotic Systems, Sensor Fusion IV*, 1991.

[11] H. C. Fryer. *Concepts and Methods of Experimental Statistics.* Allyn and Bacon, Inc., Boston, 1966.

[12] Gregory D. Hager. *Task-Directed Sensor Fusion and Planning.* Kluwer, Boston, MA, 1990.

[13] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions.* John Wiley and Sons, New York, NY, 1986.

[14] Jak Kirman, Kenneth Basye, and Thomas Dean. Sensor abstraction for control of navigation. In *Proc. Int'l Conf. on Robotics and Automation*, Sacramento, CA, 1991.

[15] Jan J. Koenderink and Andrea J. van Doorn. Photometric invariants related to solid shape. In Berthold K. P. Horn and Michael J. Brooks, editors, *Shape From Shading*, pages 301–322. MIT Press, 1989.

[16] David J. Kriegman. *Object Classes and Image Contours in Model-Based Vision.* PhD thesis, Stanford University, 1989.

[17] Benjamin Kuipers. Modeling spatial knowledge. *Cognitive Science*, 2:129–153, 1978.

[18] Benjamin Kuipers and Yung-Tai Byun. A robust qualitative method for robot spatial reasoning. In *Proc. National Conference on Artificial Intelligence*, pages 774–779, 1988.

[19] John J. Leonard, Hugh F. Durrant-Whyte, and Ingemar J. Cox. Dynamic map building for an autonomous mobile robot. *International Journal of Robotics Research*, 11(4):286–298, 1992.

[20] Maja J. Mataric. A distributed model for mobile robot environment-learning and navigation. Technical Report 1228, MIT Artificial Intelligence Laboratory, 1990.

[21] Drew McDermott. Finding objects with given spatial properties. Technical Report 195, Yale University Department of Computer Science, 1981.

[22] Hans P. Moravec and Alberto Elfes. High resolution maps from wide angle sonar. In *Proc. Int'l Conf. on Robotics and Automation*, pages 138–145, 1985.

[23] Randall Nelson. *Visual Navigation.* PhD thesis, Computer Vision Laboratory, University of Maryland, 1989.

[24] Randall Nelson and John Aloimonos. Obstacle avoidance: Towards qualitative vision. In *Proc. Int'l Conf. on Computer Vision*, 1988.

[25] Alex P. Pentland. Local shading analysis. In Berthold K. P. Horn and Michael J. Brooks, editors, *Shape From Shading*, pages 443–488. MIT Press, 1989.

[26] Franco Preparata and Michael Ian Shamos. *Computational Geometry: An Introduction.* Springer-Verlag, New York, 2nd edition, 1988.

[27] Karen B. Sarachik. Visual navigation: Constructing and utilizing simple maps of an indoor environment. Technical Report 1113, MIT Artificial Intelligence Laboratory, 1989.

[28] Randall Smith, Matthew Self, and Peter Cheeseman. Estimating uncertain spatial relationships in robotics. In *Proceedings of the Second Workshop on Uncertainty in Artificial Intelligence*, Philadelphia, PA, 1986.

[29] Michael J. Swain. *Color Indexing.* PhD thesis, University of Rochester, 1990.

[30] Konstantinos Tarabanis, Roger Y. Tsai, and Peter K. Allen. Automated sensor planning and modeling for robotic vision tasks. Technical Report CUCS 045-90, Columbia University, New York, NY, 1990.

[31] Andrew P. Witkin. Recovering surface shape and orientation from texture. *Artificial Intelligence*, 17:17–47, 1981.

[32] Christoph Zetsche and Terry Caelli. Invariant pattern recognition using multiple filter image representations. *Computer Vision, Graphics, and Image Processing*, 45:251–262, 1989.