??, ??, 1–39 (??)

Statistical Approaches to Feature-Based Object Recognition

WILLIAM M. WELLS III *

sw@ai.mit.edu

Massachusetts Institute of Technology, Artificial Intelligence Laboratory , 545 Technology Square, Cambridge, Massachusetts 02139. The author is also affiliated with Harvard Medical School, Brigham and Women's Hospital, Department of Radiology.

Received ??. Revised ??.

Abstract. This paper examines statistical approaches to model-based object recognition.

Evidence is presented indicating that, in some domains, normal (Gaussian) distributions are more accurate than uniform distributions for modeling feature fluctuations. This motivates the development of new maximum-likelihood and MAP recognition formulations which are based on normal feature models. These formulations lead to an expression for the posterior probability of the pose and correspondences given an image.

Several avenues are explored for specifying a recognition hypothesis. In the first approach, correspondences are included as a part of the hypotheses. Search for solutions may be ordered as a combinatorial search in correspondence space, or as a search over pose space, where the same criterion can equivalently be viewed as a robust variant of chamfer matching. In the second approach, correspondences are not viewed as being a part of the hypotheses. This leads to a criterion that is a smooth function of pose that is amenable to local search by continuous optimization methods. The criteria is also suitable for optimization via the Expectation-Maximization (EM) algorithm, which alternates between pose refinement and re-estimation of correspondence probabilities until convergence is obtained.

Recognition experiments are described using the criteria with features derived from video images and from synthetic range images.

Keywords: Object recognition, expectation maximization algorithm, pose estimation

1. Introduction

1.1. *Summary*

This paper examines statistical approaches to model-based object recognition, in order to clarify the assumptions that are made in formulating the problem, and to explore alternative assumptions, formulations and their interconnections.

* This paper describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by an Office of Naval Research University Research Initiative grant under contract N00014-86-K-0685, and in part by the Advanced Projects Agency of the Department of Defense under Army contract number DACA76-85-C-0010 and under Office of Naval Research contract N00014-85-K-0124. Summer support was provided by Group 53 at the Massachusetts Institute of Technology, Lincoln Laboratory.

In model-based object recognition, details about some object(s) are known in advance, and the problem is that of determining information about the presence of objects in a scene. The information that is sought typically includes the position and orientation (pose) of the object, with respect to the camera, as well as an indication of the amount of evidence supporting the hypothesis.

Feature-based approaches, where compact geometrical entities (e.g. points, lines and curves) are used to represent both the object and the image, have been utilized for many years with the motivation that such representations of the salient aspects of the model and image may be employed to quickly search for and evaluate hypotheses. Recently, *indexing* methods have become available for fast hypothesis generation. These methods motivate the complementary mechanisms described here for the refinement and verification of indexed hypotheses. While correspondences established among object and image features are typically constructed during the search process, they can also be used for evaluating the evidence for a particular hypothesis. For example, a hypothesis might be supported if the hypothesized pose projects the object features in a manner that agrees with measured image features. The need to evaluate such feature agreement raises the issue of uncertainty.

The predictions of projected object features will seldom agree exactly with measured image features (e.g. in their image coordinates) because of sensor noise and artifacts introduced by image processing and feature extraction mechanisms. This difficulty has been frequently addressed by tolerating a limited amount of mismatch between predicted and detected features. Such methods have been called *bounded-error* approaches. In addition, some predicted features may fail to be detected due to occlusions.

The bounded-error recognition problem may be formulated as finding the pose such that the number N of object features that project to within the tolerance of some image feature is maximized. An alternative formulation is to search until finding a pose such that the number N is greater than a threshold.

The issue of uncertainty in feature location and detection suggests a statistical approach to model-based object recognition. With simple probabilistic modeling of the object recognition problem, it can be shown (see Section 6.3) that N may be interpreted as the log-posterior-probability of a pose given an image, and the first description of bounded-error matching listed above is equivalent to a Maximum A-Posteriori Probability (MAP) formulation of the problem. One of the assumptions that leads to this formulation is that image features are uniformly distributed (within a region bounded by the tolerance) about their predicted positions. To the extent that image features are uniformly distributed, the bounded-error formulations may enjoy the optimality properties traditionally associated with MAP estimators.

In this paper, evidence is presented indicating that, in some domains, normal (Gaussian) distributions are more accurate than uniform distributions for modeling feature fluctuations. This motivates the development of new maximum-likelihood and MAP recognition formulations which are based on normal feature models, since optimal estimators typically incorporate measurement models of the domain. Object and image features will be modeled as simple geometric entities that are ideally related by a camera model, and may have correspondences established among them. Probabilistic models (conditionally-normal for matched features) will be developed for the locations of image features. In addition, probabilistic models of feature correspondences will be developed that capture the level of clutter in a scene. These formulations will lead to an expression for the posterior probability of the pose and correspondences given an image.

At this point, several avenues are explored for specifying recognition hypotheses. The first, and most traditional approach is called MAP Model Matching (MMM). Here, correspondences are included as a part of the hypotheses. Search for solutions may be ordered as a combinatorial search in correspondence space, where tree-search methods have traditionally been used. When ordered so that the outer search is instead over pose space, the same criterion can also be viewed as a robust variant of chamfer matching.

In the second approach, Posterior Marginal Pose Estimation (PMPE), correspondences are not viewed as being a part of the hypotheses. Formally, the posterior probability of pose is obtained from the joint posterior probability on pose and correspondences by computing a marginal over correspondences. This leads to a criteria that is a smooth function of pose that is amenable to local search by continuous optimization methods. PMPE is also suitable for optimization via the Expectation-Maximization (EM) algorithm, which alternates between pose refinement and re-estimation of correspondence probabilities until convergence is obtained.

The pose-space search approaches, particularly PMPE, allow continuous local search in pose space for recognition hypotheses – an attractive alternative to combinatorial search, particularly when combined with indexing methods, which typically yield somewhat inaccurate pose estimates, since they are based on minimal sets of corresponding features.

1.2. The Problem

In this paper, the recognition problem is restricted to finding occurrences of a single object in scenes that may contain other unknown objects. Despite the simplification and years of research, the problem remains largely unsolved. Robust systems that can recognize smooth objects having six degrees of freedom of position, under varying conditions of illumination, occlusion, and background, are not commercially available. Much effort has been expended on this problem as is evident in the comprehensive reviews of research in computer-based object recognition by Besl and Jain (Besl and Jain, 1985), who cited 203 references, and Chin and Dyer (Chin and Dyer, 1986), who cited 155 references.

1.3. Statistical Approach

Model-based object recognition will be approached via the principles of Maximum Likelihood (ML) and Maximum A-Posteriori probability (MAP). These principles, along with specific probabilistic models of aspects of object recognition, will be used to derive objective functions for evaluating and refining recognition hypotheses. The ML and MAP criteria have a long history of successful application in formulating decisions and in making estimates from observed data. They have attractive properties of optimality and they are often useful in domains where measurement errors are significant.

In other areas of computer vision, statistics has proven useful as a theoretical framework. The work of Yuille, Geiger and Bülthoff on stereo (Yuille, Geiger and Bülthoff, 1990) is one example, while in image restoration the work of Geman and Geman (Geman and Geman, 1984), Marroquin (Marroquin, 1985), and Marroquin, Mitter and Poggio (Marroquin, Mitter and Poggio, 1987) are others. The statistical approach that is used in this paper converts the recognition problem into a well defined (although not necessarily easy) optimization problem. This has the advantage of providing an explicit characterization of the problem, while separating it from the description of the algorithms used to solve it.

Ad hoc objective functions have been profitably used in some areas of computer vision. Such an approach is used by Barnard in stereo matching (Barnard, 1987), Blake and Zisserman (Blake and Zisserman, 1987) in image restoration and Beveridge, Weiss and Riseman (Beveridge, Weiss and Riseman, 1989) in line segment based model matching. With this approach, plausible forms for components of the objective function are often combined using trade-off parameters that are determined empirically.

Objective functions that are derived from statistical theories have several potential advantages over ad hoc formulations. One is that the assumptions become explicit – the forms of the objective function components are clearly related to specific probabilistic models. If these models fit the domain then there is some expectation that the resulting criteria will perform well. A second advantage is that the trade-off parameters in the objective function can be derived from measurable statistics of the domain.

1.4. Alignment and Indexing Methods

Hypothesize-and-test, or *alignment* methods have proven useful in visual object recognition. Huttenlocher and Ullman (Huttenlocher and Ullman, 1988) used search over minimal sets of corresponding features to establish candidate hypotheses. In their work these hypotheses, or *alignments*, are tested by projecting the object model into the image using the pose (position and orientation) implied by the hypothesis, and then by performing a detailed comparison with the image. The basic strategy of the alignment method is to use separate mechanisms for generating and testing hypotheses.

Recently, feature-based *indexing* methods have become available for efficiently generating hypotheses in recognition. Such methods avoid a significant amount of search by using pre-computed tables for looking up object features that might correspond to a group of image features. The geometric hashing method of Lamdan and Wolfson (Lamdan and Wolfson, 1988) uses invariant properties of small groups of features under affine transformations as the look-up key. Clemens and Jacobs (Clemens and Jacobs, 1990) (Clemens and Jacobs, 1991), and Jacobs (Jacobs, 1992) described indexing methods that gain efficiency by using a feature grouping process to select small sets of image features that are likely to belong to one object in the scene.

The Hough transform (Hough, 1962) (Illingworth and Kittler, 1988) is another commonly used method for generating hypotheses in object recognition. In the Hough method, feature-based clustering is performed in *pose space*, the space of the transformations describing the possible motion of the object. This method was used by Grimson and Lozano-Pérez (Grimson and Lozano-Pérez, 1987) to localize the search in recognition.

These fast methods of hypothesis generation provide additional motivation for using the alignment approach. They are often most effective when used in conjunction with a verification technique. Verification is important because indexing methods can be susceptible to table collisions, while Hough methods sometimes generate false positives due to their aggregation of inconsistent evidence in pose space bins. This last point has been argued by Grimson and Huttenlocher (Grimson and Huttenlocher, 1990).

The usual alignment strategy may be summarized as *align, verify*. Alignment and verification place differing pressures on the choice of features for recognition. Mechanisms used for generating hypotheses typically have computational complexity that is polynomial in the number of features involved. Because of this, there is significant advantage to using low resolution features – there are fewer of them. Unfortunately, pose estimates based on coarse features tend to be less accurate than those based on high resolution features.

In contrast, verification may be more reliable with high resolution features, since this yields more detailed comparisons. These differing pressures may be accommodated by employing *coarse-fine* approaches. The coarse-fine strategy was utilized successfully in stereo by Grimson (Grimson, 1985). In the coarse-fine strategy, hypotheses derived from low-resolution features limit the search for hypotheses derived from high-resolution features.

Align, Refine, Verify The recognition approach described here may be summarized as *align, refine, verify*. The key observation is that local search in pose space may be used to refine the hypothesis from the alignment stage before verification is carried out. In hypothesize and test methods, the pose estimates of the initial hypotheses tend to be somewhat inaccurate, since they are based on minimal sets of corresponding features. Better pose estimates (hence, better verification) are likely to result from using all supporting image feature data, rather than a small subset. Section 8 describes a method that refines the pose estimate while simultaneously identifying and incorporating the constraints of all supporting image features.

This paper focuses on mechanisms that are designed for the refinement of pose estimates. There is a remaining problem that concerns detection, or formulating a decision about whether the object is actually present at the indicated pose. This issue has often been approached by likelihood-ratio testing (i.e. by setting thresholds on an objective function). While the same approach may be used with the criteria developed here, it is not addressed in this paper.

1.5. Structure of the Paper

Section 2 describes the probabilistic models of the correspondences between image features and features belonging to either the object or to the background. In Section 3, probabilistic models are developed that characterize feature fluctuations. Empirical evidence is described that supports the use of normal (Gaussian) feature fluctuation models.

Section 4 discusses the modeling of objects, while deterministic models of the projection of features into the image are discussed in Section 5. The projection methods used in the experiments reported here are linear in the parameters of the transformations. Methods for 2D and 3D are discussed, including the Linear Combination of Views method of Ullman and Basri (Ullman and Basri, 1989).

In Section 6 the above models are combined in a Bayesian framework to construct a criterion, *MAP Model Matching*, for evaluating hypotheses in object recognition. A connection between MAP Model Matching and a method of robust chamfer matching (Jiang, Robb and Holton, 1992) is described.

Building on the above, a second criterion is described in Section 7: *Posterior Marginal Pose Estimation* (PMPE). Here, the solution being sought is simply the pose of the object.

Section 8 describes use of the the *Expectation-Maximization* (EM) algorithm (Dempster, Laird and Rubin, 1977) for finding local maxima of the PMPE objective function, and describes a limiting-case connection to the Iterative Closest Point (ICP) algorithm (Besl and McKay, 1992).

In Section 9, MMM PMPE and bounded error methods are compared and contrasted by formulating them in a common framework that is based on potential functions.

Recognition experiments are described in Section 10. The MMM objective function is explored via combinatorial search with features derived from video images. In addition, search of PMPE is carried out via the EM algorithm to refine and evaluate poses in 2D and 3D recognition. Initial hypotheses for the 2D experiments were generated by a simple form of 2D indexing, *Angle Pair Indexing* (Wells, 1992). Local search of PMPE is also demonstrated in an experiment using features derived from synthetic range images.

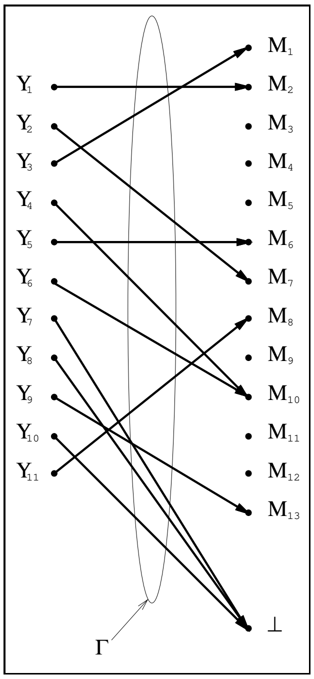

Fig. 1. Image Features (Y), Object Features (M), Background Feature ($\perp$), and Correspondences (Γ).

Related work and conclusions appear in Sections 11 and 12.

2. Modeling Feature Correspondence

This section is concerned with probabilistic models of feature correspondences. These models will serve as priors in the statistical theories of object recognition that are described in Sections 6 and 7. They are used to assess the probability that features correspond before the image data is compared to the object model. They capture the expectation that some features in an image are anticipated to be due to the object.

2.1. Features and Correspondences

Let the image that is to be analyzed be represented by a set of v-dimensional features

$$Y = \{Y_1, Y_2, \ldots, Y_n\} \ , \ \ Y_i \in R^v \ .$$

Similarly, the object to be recognized is also described by a set of features,

$$M = \{M_1, M_2, \ldots, M_m\} \ .$$

The features will usually be represented by real matrices. Additional details on image and object features will be found in later sections.

The interpretation of the features in an image will be represented by the variable Γ, which describes the mapping from image features to object features or the scene background. This is also referred to as the *correspondences*.

$$\Gamma = \{\Gamma_1, \Gamma_2, \ldots, \Gamma_n\} \ , \ \ \Gamma_i \in M \cup \{\perp\} \ .$$

An interpretation is illustrated in Figure 1.

In an interpretation, each image feature, Y_i, will be assigned either to some object feature M_j, or to the background, which is denoted by the symbol $\perp$. This symbol plays a role similar to that of the null character in the interpretation trees of

Grimson and Lozano-Pérez (Grimson and Lozano-Pérez, 1987). Γ is a collection of variables that is indexed in parallel with the image features. Each variable Γ_i represents the assignment of the corresponding image feature Y_i. It may take on as value any of the object features M_j, or the background, $\perp$. Thus, the meaning of the expression $\Gamma_5 = M_6$ is that image feature five is assigned to object feature six, likewise $\Gamma_7 = \perp$ means that image feature seven has been assigned to the background. In an interpretation each image feature is assigned, while some object features may not be. Additionally, several image features may be assigned to the same object feature. While this representation allows image interpretations that are implausible, they will ultimately be discouraged by other mechanisms that penalize metrical inconsistency.

2.2. *Independent Correspondence Model*

In this section a simple probabilistic model of correspondences is described. The intent is to capture some information bearing on correspondences (before the image is compared to the object), while making a reasonable compromise between simplicity and accuracy.

In this model, the correspondence status of differing image features are assumed to be independent, so that

$$p(\Gamma) = \prod_i p(\Gamma_i) \quad . \qquad (1)$$

Here, $p(\Gamma)$ is a probability mass function on the discrete variable Γ. While there is evidence against statistical independence here, for example, because occlusion is locally correlated, it is nevertheless used as an approximation that simplifies the resulting formulations of recognition. It may be justified by the good performance of the recognition experiments described in Section 10. Additionally, few recognition systems have used non-independent models of correspondence. Relaxations of this assumption are described in (Breuel, 1992) and (Wells, 1992).

The component probability function is designed to characterize the amount of clutter in the image, but to be otherwise as non-committal as possible:

$$p(\Gamma_i) = \begin{cases} B_i & \text{if } \Gamma_i = \perp \\ \frac{1-B_i}{m} & \text{otherwise} \end{cases} \quad . \qquad (2)$$

This model assigns the the feature to the image background with probability B_i, while the remaining probability is uniformly distributed on assignments to the model features.

The joint model $p(\Gamma)$ is the maximum entropy probability function that is consistent with the constraint that the probability of an image feature belonging to the background is B_i. E.T. Jaynes has argued that maximum entropy distributions are the most honest representation of a state of incomplete knowledge, because "Any other distribution would necessarily either assume information that we do not have, or contradict information that we do have" (Jaynes, 1982). B may be estimated by taking simple statistics on images from the domain. $B = .9$ would mean that 90% of image features are expected to be due to the background.

Having B constant during recognition is an approximation. The number of features due to the object will likely vary according to the size of the object in the scene. B could be estimated at recognition time by pre-processing mechanisms that evaluate image clutter, and factor in expectations about the size of the object. In practice, the approximation works well in controlled situations. In the experiments described Section 10 B is estimated by $\frac{N-M}{N}$, from the number of model features (M) and the number of image features (N) (using the assumption that one un-obscured object is present in the image).

3. Modeling Image Features

This section describes probabilistic models of image features, another component of the statistical theories of object recognition that are described in Sections 6 and 7.

The probability density function for the coordinates of image features, conditioned on correspondences and pose, is defined. The PDF has two important cases, depending on whether the image feature is assigned to the object, or to the background. Features matched to the object are modeled with normal densities, while uniform densities are used for background features.

Empirical evidence is provided to support the use of normal densities for matched features. Many recognition systems implicitly use uniform densities to model matched image features (*bounded-error* models). The empirical evidence of Section 3.2 indicates that the normal model may sometimes be better, and its use could lead to better performance in recognition.

Finally, a form of feature stationarity is described.

3.1. A Uniform Model for Background Features

The image features, Y_i, are v dimensional vectors. When assigned to the background, they are assumed to be uniformly distributed,

$$p(Y_i \mid \Gamma, \beta) = \frac{1}{W_1 \cdots W_v} \qquad \text{if } \Gamma_i = \perp \quad . \tag{3}$$

(The PDF is defined to be zero outside the coordinate space of the image features, which has extent W_i along dimension i.) Γ describes the correspondences from image features to object features, and β describes the position and orientation, or *pose* of the object. For example, if the image features are 2D points in a 640 by 480 image, then $p(Y_i \mid \perp, \beta) = \frac{1}{640*480}$, within the image. For Y_i, this probability function depends only on the i'th component of Γ.

Providing a satisfying probability density function for background features is problematic. Equation 3 describes the maximum entropy PDF consistent with the constraint that the coordinates of image features are always expected to lie within the coordinate space of the image features.

3.2. Normal Model for Matched Features

This section discusses the use of normal models for feature fluctuations.

Image features that are matched to object features are assumed to be normally distributed about their predicted position in the image,

$$p(Y_i \mid \Gamma, \beta) = \mathrm{G}_{\psi_{ij}}(Y_i - \mathcal{P}(M_j, \beta)) \qquad \text{if } \Gamma_i = M_j \quad . \tag{4}$$

Here Y_i, Γ, and β are defined as above, and the predicted coordinates of image feature Y_i are given by $\mathcal{P}(M_j, \beta)$, the projection of object feature j into the image with object pose β. $\mathrm{G}_{\psi_{ij}}$ is the v-dimensional normal probability density function with covariance matrix ψ_{ij},

$$\mathrm{G}_{\psi_{ij}}(x) \equiv (2\pi)^{-\frac{v}{2}} |\psi_{ij}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} x^T \psi_{ij}^{-1} x\right) \quad .$$

The covariance matrix ψ_{ij} is discussed more fully in Section 3.3, while projection and pose are discussed in more detail in Section 5.

Normal densities have the property of maximum entropy among densities of fixed mean and covariance.

Empirical Evidence for the Normal Model
This section describes some empirical evidence from the domain of video image edge features indicating that normal probability densities are good models of feature fluctuations, and that they can be better than uniform probability densities. The evidence is provided in the form of observed and fitted cumulative distributions and Kolmogorov-Smirnov tests. The model distributions were fitted to the data using the Maximum Likelihood method.

The data that is analyzed are the perpendicular and parallel deviations of coarse and fine edge features derived from video images. The fine features are shown in Figure 9.

The model features were selected by hand from features extracted from a video image where the object was viewed against a black background, while the image features were derived from an image of the object that had a cluttered background. The relationship of the camera and object was the same in the two images.

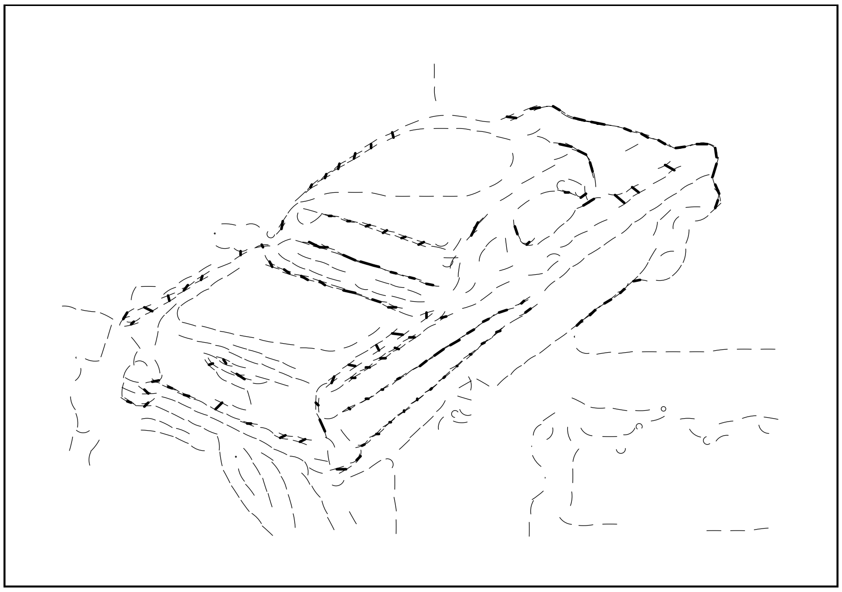

Fig. 2. Fine Feature Correspondences

For the analysis in this section, the feature data consists of the average of the x and y coordinates of the pixels from edge curve fragments – they are 2D point features. The features are displayed as circular arc fragments for clarity. The edge curves were broken into 10 and 20 pixel fragments for the fine and coarse features respectively, starting from one end of the underlying edge curve.

Correspondences from image features to model features were established by a neutral subject using a mouse. These correspondences are indicated by heavy lines in Figure 2 for the fine features. Perpendicular and parallel deviations of the corresponding features were calculated with respect to edge curve normals at the image features.

The model features were extracted using methods that are described in (Wells, 1992), the smoothing standard deviation used in the edge detection was 2.0 and 4.0 pixels respectively, for the fine and coarse features. These features were also used in the experiments reported in Section 10.2, and the correspondences were used there as data for estimating feature fluctuation statistics.

Figure 3 shows the cumulative distributions of the perpendicular and parallel deviations of the fine features. The cumulative distributions of fitted normal densities are plotted as dots over the observed distributions. The distributions were fitted to the data using the Maximum Likelihood method – the mean and variance of the normal density are set to the mean and variance of the data. These figures show good agreement between the observed distributions, and the fitted normal distributions. Similar agreement was obtained with the coarse features.

The observed cumulative distributions are shown again in Figure 4, this time with the cumulative distributions of fitted uniform densities over-plotted in dots. As before, the uniform densities were fitted to the data using the Maximum Likelihood method – in this case the uniform densities are adjusted to just include the extreme data. These figures show relatively poor agreement between the observed and fitted distributions, in comparison to the case of normal densities.

Kolmogorov-Smirnov Tests The Kolmogorov-Smirnov (KS) test (described in (Press et al., 1986)) is one way of analyzing the agreement between observed and fitted cumulative distributions, such as the ones in Figures 3 and 4. The KS test is computed on the magnitude of the largest difference between the observed and hypothesized (fitted) distributions. This will be referred to as D. The probability distribution on this distance, under the hypothesis that the data were drawn from the hypothesized distribution, can be calculated. An asymptotic formula is given by

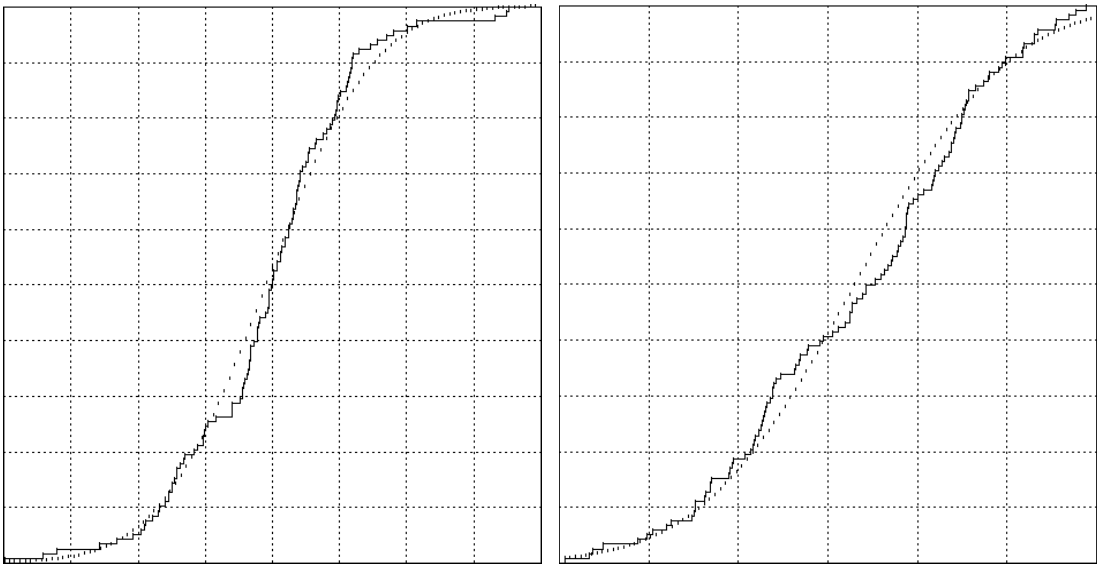

Fig. 3. Observed Cumulative Distributions and Fitted Normal Cumulative Distributions for Fine Features. Vertical Units are Tenths, Horizontal Units are Pixels. Left: Parallel Fluctuations, Right: Perpendicular Fluctuations

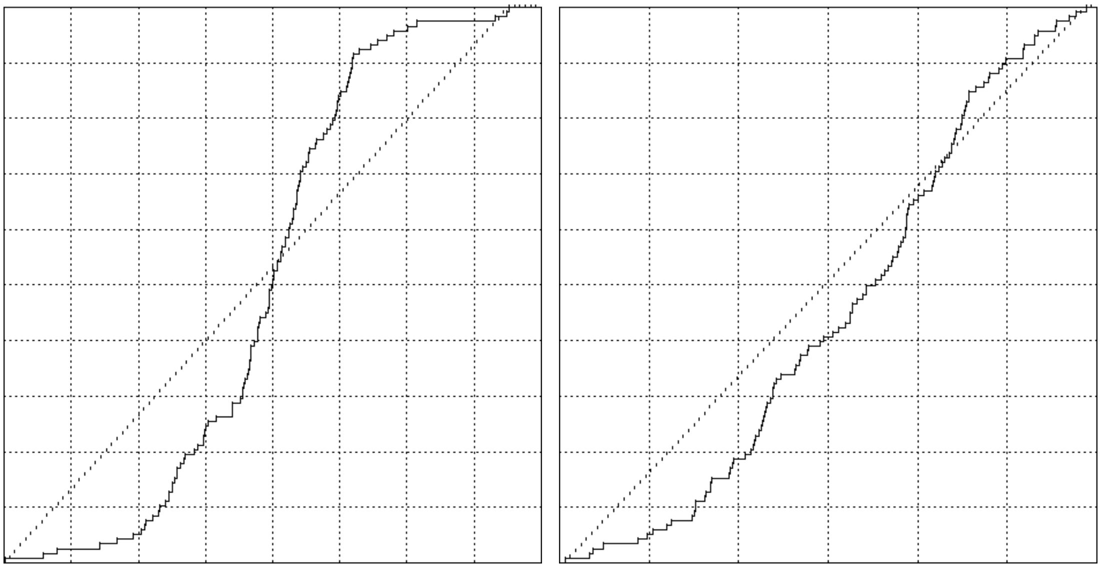

Fig. 4. Observed Cumulative Distributions and Fitted Uniform Cumulative Distributions for Fine Features. Vertical Units are Tenths, Horizontal Units are Pixels. Left: Parallel Fluctuations, Right: Perpendicular Fluctuations

$$P(D \geq D_o) = Q(\sqrt{N} D_o)$$

where

$$Q(x) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp\left(-2j^2 x^2\right) \quad ,$$

and D_o is the observed value of D.

The results of KS tests of the consistency of the data with fitted normal and uniform distributions are shown in Table 1. Low values of $P(D \geq D_o)$ suggest incompatibility between the data and the hypothesized distribution. In the cases of fine perpendicular and parallel deviations, and coarse perpendicular deviations, refutation of the uniform model is strongly indicated. Strong contradictions of the fitted normal models are not indicated in any of the cases.

If the subdivision of edge curves into features were un-correlated across images, a uniform distribution would be expected for parallel deviations. As implemented, the subdivision is started at the end of detected edge curves, and these ends are somewhat correlated. This correlation may may explain the peaked density that has been observed.

3.3. Oriented Stationary Statistics

The covariance matrix ψ_{ij} that appears in the model of matched image features in Equation 4 is allowed to depend on both the image feature and the object feature involved in the correspondence. Indexing on i allows dependence on the image feature detection process, while indexing in j allows dependence on the identity of the model feature. This is useful when some model features are known to be noisier than others. This flexibility is carried through the formalism of later sections. Although such flexibility can be useful, substantial simplification results by assuming that the features statistics are stationary in the image, i.e. $\psi_{ij} = \psi$, for all ij. This could be reasonable if the feature fluctuations were isotropic in the image, for example. In its strict form this assumption may be too limiting, however. This section outlines a compromise approach, oriented stationary statistics, that was used in the implementations described in Section 10.

Table 1. Kolmogorov-Smirnov Tests

		Normal Hypothesis		Uniform Hypothesis	
Deviate	N	D_o	$P(D \geq D_o)$	D_o	$P(D \geq D_o)$
Fine Perpendicular	118	.0824	.3996	.2244	.000014
Fine Parallel	118	.0771	.4845	.1596	.0049
Coarse Perpendicular	28	.1526	.5317	.2518	.0574
Coarse Parallel	28	.0948	.9628	.1543	.5172

The method attaches a coordinate system to each image feature. The coordinate system has its origin at the point location of the feature, and is oriented with respect to the direction of the underlying curve at the feature point. When (stationary) statistics on feature deviations are measured, they are taken relative to these coordinate systems.

Estimating the Parameters The experiments reported in Section 10 use the normal model and oriented stationary statistics for matched image features. After this choice of model, it is still necessary to supply the specific parameters for the model, namely the covariance matrices of the normal densities.

The parameters were estimated from observations on matches done by hand on sample images from the domain. Because of the stationarity assumption it is possible to estimate the common covariance, $\hat{\psi}$, by observing match data on one image. For this purpose, a match was done with a mouse between model features and a representative image from the domain. During this process, the pose of the object was the same in the two images. This produced a set of corresponding edge features. For the sake of example, the process will be described for 2D point features. The procedure has also been used with 2D point-radius features and 2D oriented-range features, that are described in Section 5.

Let the observed image features be described by Y_i, and the corresponding model features by $\hat{Y}_i$. The observed residuals between the "data" image features, and the model features are $\Delta_i = Y_i - \hat{Y}_i$.

The features are derived from edge data, and the underlying edge curve has an orientation angle in the image. These angles are used to define coordinate systems specific to each image feature Y_i. These coordinate systems define rotation matrices R_i that are used to transform the residuals into the coordinate systems of the features, in the following way: $\Delta'_i = R_i \Delta_i$.

The stationary covariance matrix of the matched feature fluctuations observed in the feature coordinate systems is then estimated using the Maximum Likelihood method, as follows, $\hat{\psi} = \frac{1}{n}\sum_i \Delta'_i \Delta'^T_i$ (here T denotes the matrix transpose operation).

The resulting covariance matrices typically indicate larger variance for deviations along the edge curve than perpendicular to it, as suggested by the data in Figure 3.

Specializing the Covariance At recognition time, it is necessary to specialize the constant covariance to each image feature. This is done by rotating it to orient it with respect to the image feature.

A covariance matrix transforms like the following product of residuals: $\Delta'_i \Delta'^T_i$. This may be transformed back to the image system as follows, $R_i^T \Delta'_i \Delta'^T_i R_i$. Thus the constant covariance is specialized to the image features in the following way, $\psi_{ij} = R_i^T \hat{\psi} R_i$.

4. Modeling Objects

This section briefly describes several approaches for obtaining object model features.

4.1. Monolithic 3D Object Models

For some objects, having a single 3D model seems a natural choice for a recognition system. If the object is polyhedral, and is represented by a list of 3D line segments, then predicting the features that will appear in a given high resolution view is a simple matter.

For other objects, such as smoothly curved objects, the situation is more complex. Predicting features becomes more elaborate. In video imagery, occluding edges (or *limbs*) are often important features. Calculating the limb of a smooth 3D surface is usually complicated. Ponce and Kriegman (Ponce and Kriegman, 1989) describe an approach for objects modeled by parametric surface patches. Algebraic elimination theory is used to relate image limbs to the model surfaces that generated them. Brooks' vision system, Acronym (Brooks, 1983), also recognized curved objects from image limbs. It used generalized cylinders to model objects. A drawback of this approach is that it is awkward to realistically model typical objects, like telephones or automobiles, with generalized cylinders.

Predicting reduced resolution image features is another difficulty with monolithic 3D models. This is a drawback because doing recognition with reduced resolution features is an attractive strategy: with fewer features less search will be needed. One solution would be to devise a way of smoothing 3D object models such that simple projection operations would accurately predict reduced resolution image features. No such method is known to the author.

Self occlusion is an additional complexity of the monolithic 3D model approach. In computer graphics there are several ways of dealing with this issue, among them hidden line and z-buffer methods. These methods are fairly expensive, at least in comparison to sparse point projections.

4.2. Image-Based Object Modeling

One approach to avoiding the difficulties discussed in the previous section is to use an image-based approach to object modeling. Ullman and Basri (Ullman and Basri, 1989) and Breuel (Breuel, 1992) have discussed such approaches. Some motivation for image-based approaches may be provided by the biological evidence that animal vision systems are attuned to specific views of faces (Perrett et al., 1985).

An important issue with image-based object modeling concerns how to predict image features in a way that covers the space of poses that the object may assume.

Bodies undergoing rigid motion in space have six degrees of freedom, three in translation, and three in rotation. This six parameter pose space may be split into two parts – the first part being translation and in image-plane rotations (four parameters) – the second part being out of image-plane rotations (two parameters: the "view sphere").

Synthesizing views of an object that span the first part of pose space can often be done using simple and efficient linear methods of translation, rotation, and scale in the plane. This approach can be precise under orthographic projection with scaling, and accurate enough in some domains with perspective projection. Perspective projection is often approximated in recognition systems by 3D rotation combined with orthographic projection and scaling. This has been called the *weak perspective* approximation (Thompson and Mundy, 1987).

The second part of pose space, out of plane rotation, is more complicated. One approach involves tesselating the view sphere around the object, and storing a view of the object for each vertex of the tesselation. Arbitrary views will then entail, at most, small out of plane rotations from stored views. These views may be synthesized using interpolation. The Linear Combination of Views method of Ullman and Basri (Ullman and Basri, 1989), works well for interpolating between nearby views.

5. Modeling Projection

5.1. Linear Projection Models

Pose determination is often a component of model-based object recognition systems, and it is frequently framed as an optimization problem. Pose determination may be significantly simplified if the feature projection model is linear in the pose parameters. The experiments described in this paper use projection models having this property, this enables solving the embedded optimization problem using least squares methods. Such techniques are advantageous because unique solutions may be obtained easily in closed form. This is a significant advantage, since the embedded optimization problem may be solved many times during the course of a search for an object in a scene.

All of the formulations of projection described below are linear in the parameters of the transformation. Because of this they may be written in the following form:

$$\eta_i = \mathcal{P}(M_i, \beta) = M_i \beta \ . \tag{5}$$

The pose of the object is represented by β, a column vector, and the object model feature by M_i, a matrix. The projection of the model feature into the image by pose β, is a column vector, η_i.

5.2. 2D Point-Radius Feature Model

There are advantages in using rich features in recognition – they provide more constraints, and can lead to space and time efficiencies. These potential advantages must be weighed against the practicality of detecting such features. For example, there is incentive to construct features incorporating higher derivative information at a point on a curve; however, measuring higher derivatives of curves derived from video imagery may be impractical, because each derivative magnifies the noise present in the data.

The feature described here is a compromise between richness and detectability. It is defined as follows $\eta_i = M_i \beta$, where

$$\eta_i = \begin{bmatrix} p'_{ix} \\ p'_{iy} \\ c'_{ix} \\ c'_{iy} \end{bmatrix} \qquad M_i = \begin{bmatrix} p_{ix} & -p_{iy} & 1 & 0 \\ p_{iy} & p_{ix} & 0 & 1 \\ c_{ix} & -c_{iy} & 0 & 0 \\ c_{iy} & c_{ix} & 0 & 0 \end{bmatrix}$$

$$\text{and} \qquad \beta = \begin{bmatrix} \mu \\ \nu \\ t_x \\ t_y \end{bmatrix} \qquad .$$

The coordinates of object model point i are p_{ix} and p_{iy}, while c_{ix} and c_{iy} represent the radius vector of the curve's osculating circle that touches the point on the curve, as illustrated in Figure 5. This vector is normal to the curve, and its length is the inverse of the curvature at the point.

The coordinates of the model point i, projected into the image by pose β, are p'_{ix} and p'_{iy}. This transformation is equivalent to rotation by θ, scaling by s, and translation by T, where

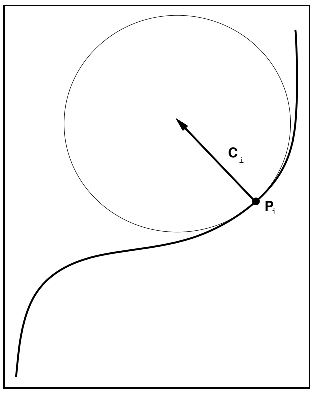

Fig. 5. Edge Curve, Osculating Circle, and Radius Vector

$$T = \begin{bmatrix} t_x \\ t_y \end{bmatrix} \qquad s = \sqrt{\mu^2 + \nu^2} \qquad \theta = \arctan\left(\frac{\nu}{\mu}\right) \tag{6}$$

. In this model, rotation and scale are effected by analogy to the multiplication of complex numbers, which induces transformations of rotation and scale in the complex plane. This analogy may be made complete by noting that the algebra of complex numbers $a + ib$ is isomorphic with that of matrices of the form

$$\begin{bmatrix} a & b \\ -b & a \end{bmatrix} \quad .$$

The image counterparts of c_{ix} and c_{iy} are given by c'_{ix} and c'_{iy}. With this model, the radius vector c rotates and scales as do the coordinates p, but it does not translate. Thus, the aggregate feature translates, rotates and scales correctly.

This feature model is used in the experiments described in Sections 10.1 and 10.2. When the underlying curvature goes to zero, the length of the radius vector diverges, and the direction becomes unstable. This has been accommodated in the experiments by truncating c.

This model is an extension of a model for 2D points that was used by Ayache and Faugeras with their vision system HYPER (Ayache and Faugeras, 1986).

5.3. 2-D Oriented-Range Feature Model

The following describes a related feature model that has been used in experiments where the imagery consists of dense range data (Section 10.4).

In similar fashion to the point-radius feature model, we may define a 2-D projection and feature model that incorporates local information about the coordinates, normal, and range at a point along a curve of range discontinuity as follows,

$$\eta_i = \begin{bmatrix} p'_{ix} \\ p'_{iy} \\ c'_{ix} \\ c'_{iy} \end{bmatrix} \qquad M_i = \begin{bmatrix} p_{ix} & -p_{iy} & 1 & 0 \\ p_{iy} & p_{ix} & 0 & 1 \\ c_{ix} & -c_{iy} & 0 & 0 \\ c_{iy} & c_{ix} & 0 & 0 \end{bmatrix}$$

$$\text{and} \qquad \beta = \begin{bmatrix} \mu \\ \nu \\ t_x \\ t_y \end{bmatrix} \qquad .$$

The coordinates of model point i are p_{ix} and p_{iy}. The coordinates of the model point i, projected into the image by pose β, are p'_{ix} and p'_{iy}. c_{ix} and c_{iy} are chosen to describe a vector who's direction is perpendicular to the range discontinuity and that points away from the discontinuity, while the length of the vector is the inverse of the range at the discontinuity. The counterparts in the image are given by c'_{ix} and c'_{iy}.

This transformation is equivalent to rotation by θ, scaling by s, and translation by T as in Equation 6.

The aggregate feature translates, rotates and scales correctly when used with imaging models where the object features scale according to the inverse of the distance to the object. This holds under perspective projection with attached range data when the object is small compared to the distance to the object.

5.4. Linear Combination of Views

The technique used in the above methods for synthesizing rotation and scale are equivalent to making linear combinations of the object model with a copy of itself that has been rotated 90 degrees in the plane.

(Ullman and Basri, 1989) describe a scheme for synthesizing views under 3D orthography with rotation and scale that has a linear parameterization. They show that the space of images of an object is a subspace of a linear space that is spanned by the components of a few images of an object. They discuss variants of their formulation that are based on two views, and on three and more views. Recovering conventional pose parameters from the linear combination coefficients is described in (Rivlin and Basri, 1992).

The following is a brief explanation of the two-view method. Point projection from 3D to 2D under orthography, rotation, and scale is a linear transformation. If two (2D) views are available, along with the transformations that produced them, then there is enough data to invert the transformations and solve for the 3D coordinates (as in stereo vision)(three equations are needed, four are available). The resulting expression for the 3D coordinates will be a linear equation in the components of the two views. New 2D views may then be synthesized from the 3D coordinates by yet another linear transformation. Compounding these linear operations yields an expression for new 2D views that is linear in the components of the original two views. There is a quadratic constraint on the 3D to 2D transformations, due to the ortho-normality of rotation matrices. The usual Linear Combination of Views approach makes use of the above linearity property while synthesizing new views with general linear transformations (without the constraints). This practice leads to two extra parameters that control stretching transformations in the synthesized image. It also reduces the need to deal with camera calibrations – the pixel aspect ratio may be accommodated in the stretching transformations.

The following projection model uses a two view variant of the Linear Combination of Views method to synthesize views with limited 3D rotation and scale. Additionally, translation has been added in a straightforward way. Model features are projected into the image as $\eta_i = M_i \beta$, where

$$\eta_i = \begin{bmatrix} \eta_{ix} \\ \eta_{iy} \end{bmatrix}$$

$$M_i = \begin{bmatrix} p_{ix} & 0 & q_{ix} & 0 & p_{iy} & 0 & 1 & 0 \\ 0 & p_{iy} & 0 & q_{iy} & 0 & p_{ix} & 0 & 1 \end{bmatrix}$$

and

$$\beta = \left[\, \beta_0 \beta_1 \beta_2 \beta_3 \beta_4 \beta_5 \beta_6 \beta_7 \, \right]^T \quad .$$

The coordinates of the i'th point in one view are p_{ix} and p_{iy}; in the other view they are q_{ix} and q_{iy}.

When this projection model is used, β does not in general describe rigid transformation, but it is nevertheless called the pose vector for notational consistency.

This method is used in the experiment described in Section 10.3.

6. MAP Model Matching

MAP Model Matching [1] (MMM) is the first of two statistical formulations of object recognition to be discussed in this paper. It builds on the models of features and correspondences, objects, and projection that are described in the previous sections. MMM evaluates joint hypotheses of match and pose in terms of their posterior probability, given an image. MMM is the starting point for the second formulation of object recognition, Posterior Marginal Pose Estimation (PMPE), which is described in Section 7.

The MMM objective function is amenable to search in correspondence space, the space of all possible assignments from image features to model and background features. This style of search has been used in many recognition systems, and it is used in a recognition experiment involving low resolution edge features in Section 10.

It will be shown that under certain conditions, searching in pose space for maxima of the MMM objective function is equivalent to robust methods of chamfer matching (Jiang, Robb and Holton, 1992).

6.1. *Objective Function for Pose and Correspondences*

In this section an objective function for evaluating joint hypotheses of match and pose using the MAP criterion will be derived.

Briefly, probability densities of image features, conditioned on the parameters of match and pose ("the parameters"), are combined with prior probabilities on the parameters using Bayes' rule. The result is a posterior probability density on the parameters, given an observed image. An estimate of the parameters is then formulated by choosing them so as to maximize their a-posteriori probability. (Hence the term *MAP*. See Beck and Arnold's textbook (Beck and Arnold, 1977) for a discussion of MAP estimation.) MAP estimators are especially practical when used with normal probability densities.

In more detail, the probabilistic models of image features described in Section 3 may be written as follows:

$$p(Y_i \mid \Gamma, \beta) = \begin{cases} \frac{1}{W_1 W_2 \cdots W_v} & \text{if } \Gamma_i = \perp \\ G_{\psi_{ij}}(Y_i - \mathcal{P}(M_j, \beta)) & \text{if } \Gamma_i = M_j \end{cases} \quad . \tag{7}$$

Here ψ_{ij} is the covariance matrix associated with image feature i and object model feature j. Thus image features arising from the background are uniformly distributed over the image feature coordinate space (the extent of the image feature coordinate space along dimension i is given by W_i), and matched image features are normally distributed about their predicted locations in the image. In some applications ψ_{ij} could be a constant that is independent of i and j – an assumption that the feature statistics are stationary in the image, or ψ_{ij} may depend only on i, the image feature index. The latter is the case when the oriented stationary statistics model is used (see Section 3.3).

Assuming independent features, the joint probability density on image feature coordinates may be written as follows

$$\begin{aligned} p(Y \mid \Gamma, \beta) &= \prod_i p(Y_i \mid \Gamma, \beta) \qquad (8) \\ &= \prod_{i:\Gamma_i = \perp} \frac{1}{W_1 W_2 \cdots W_v} \cdot \\ &\quad \prod_{ij:\Gamma_i = M_j} G_{\psi_{ij}}(Y_i - \mathcal{P}(M_j, \beta)) \ . \end{aligned}$$

This independence assumption is pragmatically motivated, although it typically holds in domains where sensor noise dominates in feature fluctuations, such as laser radar range imagery.

The next step in the derivation is the construction of a joint prior on correspondences and pose. Prior information on the pose is assumed to be supplied as a normal density,

$$p(\beta) = G_{\psi_\beta}(\beta - \beta_0)$$

where

$$G_{\psi_\beta}(x) = (2\pi)^{\frac{-z}{2}} |\psi_\beta|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} x^T \psi_\beta^{-1} x\right) \ .$$

Here ψ_β is the covariance matrix of the pose prior and z is the dimensionality of the pose vector, β. With the combination of normal pose priors and linear projection models the system is closed in the sense that the resulting pose estimate will also be normal. This is convenient for coarse-fine approaches, as discussed in Section 7. If little is known about the pose a-priori, the prior may be made quite broad. If nothing is known about the pose beforehand, the pose prior may be left out. In that case the resulting criterion for evaluating hypotheses will be based on Maximum Likelihood for pose, and on MAP for correspondences.

Assuming independence of the correspondences and the pose (before the image is compared to the object model), a mixed joint probability function may be written as follows, using Equations 1 and 2,

$$p(\Gamma, \beta) = G_{\psi_\beta}(\beta - \beta_0) \prod_{i:\Gamma_i = \perp} B_i \prod_{i:\Gamma_i \neq \perp} \frac{1 - B_i}{m} \ ,$$

(B_i is the probability that image feature i is due to the background). This a good assumption when view-based approaches to object modeling are used (these are discussed in Section 4 and used in the experiments described in Section 10). (With general 3D rotation it is inaccurate, as the visibility of features depends on the orientation of the object.) This probability function on match and pose is now used with Bayes' rule as a prior for obtaining the posterior probability of Γ and β:

$$p(\Gamma, \beta \mid Y) = \frac{p(Y \mid \Gamma, \beta) p(\Gamma, \beta)}{p(Y)} \ , \tag{9}$$

where $p(Y) = \sum_\Gamma \int d\beta \; p(Y \mid \Gamma, \beta) p(\Gamma, \beta)$ is a normalization factor that is formally the probability of the image. It is a constant with respect to Γ and β, the parameters being estimated.

The MAP strategy is used to obtain estimates of the correspondences and pose by maximizing their posterior probability with respect to Γ and β, as follows

$$\widehat{\Gamma, \beta} = \arg\max_{\Gamma,\beta} p(\Gamma, \beta \mid Y) \ .$$

For convenience, an objective function, L, is introduced that is a scaled logarithm of $p(\Gamma, \beta \mid Y)$. The same estimates will result if the maximization is instead carried out over L.

$$\widehat{\Gamma, \beta} = \arg\max_{\Gamma,\beta} L(\Gamma, \beta)$$

where

$$L(\Gamma, \beta) \equiv \ln\left(\frac{p(\Gamma, \beta \mid Y)}{C}\right) \ . \qquad (10)$$

The denominator in Equation 10 is a constant that has been chosen to cancel constants from the numerator. Its value, which is independent of Γ and β is

$$C = \frac{B_1 B_2 \cdots B_n}{(W_1 W_2 \cdots W_v)^n} (2\pi)^{\frac{-z}{2}} |\psi_\beta|^{\frac{-1}{2}} \frac{1}{p(Y)} \ .$$

After some manipulation the objective function may be expressed as

$$\begin{aligned} L(\Gamma, \beta) = &-\frac{1}{2}(\beta - \beta_o)^T \psi_\beta^{-1} (\beta - \beta_0) \qquad (11) \\ &+ \textstyle\sum_{ij:\Gamma_i = M_j} [\lambda_{ij} - \frac{1}{2}(Y_i - \mathcal{P}(M_j, \beta))^T \psi_{ij}^{-1} \\ &\cdot (Y_i - \mathcal{P}(M_j, \beta))] \ , \end{aligned}$$

where

$$\lambda_{ij} = \ln\left(\frac{1}{(2\pi)^{\frac{v}{2}} m} \frac{(1 - B_i)}{B_i} \frac{W_1 W_2 \cdots W_v}{|\psi_{ij}|^{\frac{1}{2}}}\right) \ . \quad (12)$$

When a linear projection model is used, $\mathcal{P}(M_j, \beta) = M_j \beta$ (as discussed in Section 5). In this case, the objective function takes the following form

$$\begin{aligned} L(\Gamma, \beta) = &-\frac{1}{2}(\beta - \beta_o)^T \psi_\beta^{-1} (\beta - \beta_0) \qquad (13) \\ &+ \textstyle\sum_{ij:\Gamma_i = M_j} [\lambda_{ij} - \frac{1}{2}(Y_i - M_j\beta)^T \psi_{ij}^{-1} (Y_i - M_j\beta)] \end{aligned}$$

.

When the background probability, B_i, is constant, and when the feature covariance matrix determinant is constant (as when oriented stationary statistics are used), the formulas simplify further –

$$\lambda = \ln\left(\frac{1}{(2\pi)^{\frac{v}{2}} m} \frac{(1 - B)}{B} \frac{W_1 W_2 \cdots W_v}{|\hat{\psi}|^{\frac{1}{2}}}\right) \ , \quad (14)$$

and

$$\begin{aligned} L(\Gamma, \beta) = &-\frac{1}{2}(\beta - \beta_o)^T \psi_\beta^{-1} (\beta - \beta_0) \qquad (15) \\ &+ \textstyle\sum_{ij:\Gamma_i = M_j} [\lambda - \frac{1}{2}(Y_i - M_j\beta)^T \psi_{ij}^{-1} (Y_i - M_j\beta)] \ . \end{aligned}$$

Here, $\hat{\psi}$ is the stationary feature covariance matrix, and ψ_{ij} is the specialized feature covariance matrix as discussed in Section 3.3.

The first term of the objective function of Equation 13 expresses the influence of the prior on the pose. As discussed above, when a useful pose prior is not available, this term may be dropped.

The second term has a simple interpretation. It is a sum taken over those image features that are matched to object model features. The λ_{ij} are fixed rewards for making correspondences, while the quadratic forms are penalties for deviations of observed image features from their expected positions in the image. Thus the objective function evaluates the amount of the image explained in terms of the object, with penalties for mismatch. This objective function is particularly simple in terms of β. When Γ is constant, β and its (posterior) covariance may be estimated by weighted least squares. When using an algorithm based on search in correspondence space, the estimate of β can be cheaply updated by using the techniques of sequential parameter estimation. (See Beck and Arnold (Beck and Arnold, 1977).) The λ_{ij} describe the relative value of a match component or extension in a way that allows direct comparison to the entailed mismatch penalty. The values of these trade-off parameter(s) are supplied by Equation 12 in terms of measurable domain statistics.

The form of the objective function suggests an optimization strategy: make correspondences to object features in order to accumulate correspondence rewards while avoiding penalties for mismatch. It is important that the λ_{ij} be positive, otherwise a winning strategy is be to make no matches to the object at all. This condition defines a critical level of image clutter, beyond which the MAP criteria assigns the feature to the background. The parameter λ_{ij} describes the dependence of the value of matches on the amount of background clutter. If background features are scarce, then correspondences to object features become more important.

This objective function provides a simple and uniform way to evaluate match and pose hypotheses. It captures important aspects of recognition: the amount of image explained in terms of the object, as well as the metrical consistency of the hypothesis; and it trades them off in a rational way based on domain statistics. Most previous approaches have not made use of both criteria simultaneously in evaluating hypotheses, at the potential expense of some robustness.

6.2. Search in Pose Space – Robust Chamfer Matching

This section will explore searching the MMM objective function in pose space, and describe the equivalence to a robust extension of chamfer matching.

A pose estimate is sought by ordering the search for maxima of the MMM objective function as follows,

$$\hat{\beta} = \arg\max_{\beta} \max_{\Gamma} L(\Gamma, \beta) \quad .$$

Substituting the objective function from Equation 11 yields

$$\hat{\beta} = \arg\max_{\beta} \max_{\Gamma} \sum_{ij:\Gamma_i = M_j} [\lambda_{ij} - \tfrac{1}{2}(Y_i - \mathcal{P}(M_j, \beta))^T \psi_{ij}^{-1} \cdot (Y_i - \mathcal{P}(M_j, \beta))] \quad .$$

The pose prior term has been dropped in the interest of clarity.

This equation may be simplified with the following definition,

$$D_{ij}(x) \equiv \frac{1}{2} x^T \psi_{ij}^{-1} x \quad .$$

$D_{ij}(x)$ is a generalized squared distance between observed and predicted features. It is called the squared Mahalanobis distance (Duda and Hart, 1973).

The pose estimator may now be written as

$$\hat{\beta} = \arg\max_{\beta} \max_{\Gamma} \sum_{ij:\Gamma_i = M_j} [\lambda_{ij} - D_{ij}(Y_i - \mathcal{P}(M_j, \beta))] \quad ,$$

or equivalently, as a minimization rather that maximization,

$$\hat{\beta} = \arg\min_{\beta} \min_{\Gamma} \sum_{ij:\Gamma_i = M_j} [D_{ij}(Y_i - \mathcal{P}(M_j, \beta)) - \lambda_{ij}] \quad .$$

The sum is taken over those image features that are assigned to model features (not the background) in the match. It may be re-written in the following way,

$$\hat{\beta} = \arg\min_{\beta} \sum_i \min_{\Gamma_i} \begin{cases} 0 & \text{if } \Gamma_i = \perp \\ D_{ij}(Y_i - \mathcal{P}(M_j, \beta)) - \lambda_{ij} & \text{if } \Gamma_i = M_j \end{cases}$$

, or as

$$\hat{\beta} = \arg\min_{\beta} \sum_i \min(0, \ \min_j D_{ij}(Y_i - \mathcal{P}(M_j, \beta)) - \lambda_{ij}) \quad .$$

If the correspondence reward is independent of the model feature (this holds when oriented stationary statistics are used), $\lambda_{ij} = \lambda_i$. In this case, λ_i may be added to each term in the sum without affecting the minimizing pose, yielding the following form for the pose estimator,

$$\hat{\beta} = \arg\min_{\beta} \sum_i \min(\lambda_i, \ \min_j D_{ij}(Y_i - \mathcal{P}(M_j, \beta))) \quad . \tag{16}$$

This objective function is easily interpreted – it is the sum, taken over image features of a saturated penalty. The penalty (before saturation) is the smallest generalized squared distance from the observed image feature to some projected model feature. The penalty, $\min_j D_{ij}(x - \mathcal{P}(M_j, \beta))$, has the form of a Voronoi surface, as described by (Huttenlocher et al., 1991). They describe a measure of similarity on image patterns, the Hausdorff distance, that is the upper envelope (maximum) of Voronoi surfaces. The measure used here differs in being saturated, and by using the sum of Voronoi surfaces, rather than the upper envelope. In their work, the upper envelope offers some reduction in the complexity of the measure, and facilitates the use of computational geometry methods for explicitly computing the measure in 2 and 3 dimensional spaces.

Computational geometry methods might be useful for computing the objective function of Equation 16. In higher dimensional pose spaces (4 or 6, for example) KD-tree methods may be the only such techniques currently available.

Next a connection will be shown between MMM search in pose space and a method of robust chamfer matching. First, the domain of MMM is simplified in the following way. Full stationarity of feature fluctuations is assumed (as described in Section 3.3). Further, the feature covariance is assumed to be isotropic. With these assumptions we have $\psi_{ij} = \sigma^2 I$, and $D_{ij}(x) = \frac{1}{2\sigma^2}|x|^2$. Additionally, assuming constant background probability, we have $\lambda_{ij} = \lambda$. The pose estimator of Equation 16 may now be written in the following simplified form,

$$\hat{\beta} = \arg\min_{\beta} \sum_i \min(\lambda,\ \min_j(\frac{1}{2\sigma^2}\ |Y_i - \mathcal{P}(M_j, \beta)|^2)) \ .$$

When the projection function is linear, invertible, and distance preserving, (2D and 3D rigid transformations satisfy these properties), the estimator may be expressed as follows,

$$\hat{\beta} = \arg\min_{\beta} \sum_i \min(\lambda,\ \min_j(\frac{1}{2\sigma^2}\ |\mathcal{P}^{-1}(Y_i, \beta) - M_j|^2)) \ . \tag{17}$$

This may be further simplified to

$$\hat{\beta} = \arg\min_{\beta} \sum_i \min(\lambda,\ d^2(\mathcal{P}^{-1}(Y_i, \beta))) \ , \tag{18}$$

by using the following definition of a minimum distance function.

$$d(x) \equiv \frac{1}{\sqrt{2}\sigma}\ \min_j |x - M_j| \ . \tag{19}$$

This function is proportional to the distance between its argument (a coordinate in the model system) and the nearest model feature.

For efficiency, approximations of $d^2(x)$ may be stored in an image-like array that is indexed by pixel coordinates. Such approaches have been called *chamfering*. Chamfer-based approaches to image registration problems use the array to facilitate fast evaluation of pose objective functions. Barrow et al. (Barrow et al., 1977) describe an early method where the objective function is the sum over model features of the distance from the projected model feature to the nearest image feature. Borgefors (Borgefors, 1988) recommends the use of RMS distance rather than summed distance in the objective function.

Recently, Jiang et al. (Jiang, Robb and Holton, 1992) described a method of robust chamfer matching. In order to make the method less susceptible to disturbance by outliers and occlusions, they incorporated saturation into the RMS objective function of Borgefors. Their objective function has the following form

$$\frac{1}{3}(\frac{1}{n}\sum_j \min(t^2, d_j^2))^{\frac{1}{2}} \ ,$$

where d_j^2 is the squared distance from the j'th projected model point to the nearest image point. Aside from the constants and square root, which don't affect the minimizing pose, this objective function is equivalent to Equation 18 if the role of image and model features is reversed, and the sense of the projection function is inverted. Jiang et al. have shown impressive results using robust chamfer matching to register multi-modal 3D medical imagery.

6.3. Relation to Bounded-Error Methods

The approach taken for construction of the MMM objective function in Section 6 can also be easily used for constructing an objective function based on uniform rather than normal models of feature deviations. When this is done, the part of the objective function that is quadratic in matched feature deviations is replaced by an expression that is a constant if the predicted feature is consistent (within the bounds) with the model feature, and otherwise is an infinite penalty (from the logarithm of the zero of the probability density that occurs in the "outside" of the uniform densities). The winning optimization strategy would then be to make as many consistent matches as possible, for a given pose. Thus bounded-error (BE) criteria have a probabilistic interpretation that is based on uniform models of feature deviations.

7. Posterior Marginal Pose Estimation

In the previous section recognition hypotheses consisted of a description of the correspondences between image and object features, as well as the pose of the object.

The formulation of recognition that is described in this section, Posterior Marginal Pose Estimation [2] (PMPE), builds on MAP Model Matching. It provides a smooth objective function for evaluating the pose of the object – without commitment to a particular match. The pose may be the most important aspect of the problem, in the sense that knowing the pose enables grasping or other interaction with the object.

The objective function on pose may be searched by standard methods of continuous optimization. An alternative is described in Section 8, where the Expectation – Maximization (EM) algorithm is discussed as a means of searching for local maxima of the objective.

Experiments in object recognition using the PMPE objective function are described in Section 10. There, the EM algorithm is used in conjunction with an indexing method that generates initial hypotheses.

7.1. Objective Function for Pose

The following method was motivated by the observation that in heuristic searches over correspondences with the objective function of MAP Model Matching, hypotheses having implausible matches scored poorly (i.e., with low posterior probability) in the objective function. The implication was that summing posterior probability over all the matches (at a specific pose) might provide a good pose evaluator. It is not tied to specific matches – it is perhaps in keeping with Marr's recommendation that computational theories of vision should try to satisfy a principle of least commitment (Marr, 1982).

Additional motivation was provided by the work by Yuille, Geiger and Bülthoff on stereo (Yuille, Geiger and Bülthoff, 1990). They discussed computing disparities in a statistical theory of stereo where a marginal is computed over matches.

In MAP Model Matching, joint hypotheses of match and pose were evaluated by their posterior probability given an image, $p(\Gamma, \beta \mid Y)$. Γ and β stand for correspondences and pose, respectively, and Y for the image features. The posterior probability was built from specific models of features and correspondences, objects, and projection that were described in the previous sections.

Here the same strategy is used for evaluating object poses: they are evaluated by their posterior probability, given an image: $p(\beta \mid Y)$. The posterior probability density of the pose may be computed from the joint posterior probability on pose and match that was formulated in the previous section, by formally taking the marginal over possible matches:

$$p(\beta \mid Y) = \sum_{\Gamma} p(\Gamma, \beta \mid Y) \ .$$

In Section 6.1, Equation 9, $p(\Gamma, \beta \mid Y)$ was obtained via Bayes' rule from probabilistic models of image features, correspondences, and the pose. Substituting for $p(\Gamma, \beta \mid Y)$, the posterior marginal may be written as

$$p(\beta \mid Y) = \sum_{\Gamma} \frac{p(Y \mid \Gamma, \beta) p(\Gamma, \beta)}{p(Y)} \ . \quad (20)$$

Using equations 1 (the independent correspondence model) and 8, we may express the posterior marginal of β in terms of the component densities:

$$p(\beta \mid Y) = \frac{1}{p(Y)} \sum_{\Gamma_1} \cdots \sum_{\Gamma_n} \prod_i p(Y_i \mid \Gamma, \beta) \prod_i p(\Gamma_i) p(\beta)$$

or

$$p(\beta \mid Y) = \frac{p(\beta)}{p(Y)} \sum_{\Gamma_1} \cdots \sum_{\Gamma_n} \prod_i [p(Y_i \mid \Gamma_i, \beta) p(\Gamma_i)] \ .$$

Breaking one factor out of the product gives

$$p(\beta \mid Y) = \frac{p(\beta)}{p(Y)} \sum_{\Gamma_1} \cdots \sum_{\Gamma_n} \left[\prod_{i=1}^{n-1} [\, p(Y_i \mid \Gamma_i, \beta) p(\Gamma_i)] \right] p(Y_n \mid \Gamma_n, \beta) p(\Gamma_n) \ ,$$

or

$$p(\beta \mid Y) = \frac{p(\beta)}{p(Y)} \sum_{\Gamma_1} \cdots \sum_{\Gamma_{n-1}} \left[\prod_{i=1}^{n-1} [\, p(Y_i \mid \Gamma_i, \beta) p(\Gamma_i)] \right] \left[\sum_{\Gamma_n} p(Y_n \mid \Gamma_n, \beta) p(\Gamma_n) \right]$$

. Continuing in similar fashion yields

$$p(\beta \mid Y) = \frac{p(\beta)}{p(Y)} \prod_i \left[\sum_{\Gamma_i} p(Y_i \mid \Gamma_i, \beta) p(\Gamma_i) \right] \ .$$

This may be written as

$$p(\beta \mid Y) = \frac{p(\beta)}{p(Y)} \prod_i p(Y_i \mid \beta) \ , \qquad (21)$$

since

$$p(Y_i \mid \beta) = \sum_{\Gamma_i} p(Y_i \mid \Gamma_i, \beta) p(\Gamma_i) \ . \qquad (22)$$

Splitting the Γ_i sum into its cases gives,

$$p(Y_i \mid \beta) = p(Y_i \mid \Gamma_i = \perp, \beta) p(\Gamma_i = \perp) + \sum_{M_j} p(Y_i \mid \Gamma_i = M_j, \beta) p(\Gamma_i = M_j) \ .$$

Substituting the densities assumed in the model of Section 6.1 in Equations 7 and 2 then yields

$$p(Y_i \mid \beta) = \frac{1}{W_1 \cdots W_v} B_i + \frac{1 - B_i}{m} \sum_{M_j} \mathrm{G}_{\psi_{ij}}(Y_i - \mathcal{P}(M_j, \beta)) \ . \qquad (23)$$

Installing this into Equation 21 leads to

$$p(\beta \mid Y) = \frac{B_1 B_2 \cdots B_n}{(W_1 W_2 \cdots W_v)^n} \frac{p(\beta)}{p(Y)} \prod_i \left[1 + \sum_{M_j} \frac{W_1 \cdots W_v}{m} \frac{1 - B_i}{B_i} \mathrm{G}_{\psi_{ij}}(Y_i - \mathcal{P}(M_j, \beta)) \right]$$

.

As in Section 6.1 the objective function for Posterior Marginal Pose Estimation is defined as the scaled logarithm of the posterior marginal probability of the pose,

$$L(\beta) \equiv \ln \left[\frac{p(\beta \mid Y)}{C} \right] \ ,$$

where, as before,

$$C = \frac{B_1 B_2 \cdots B_n}{(W_1 W_2 \cdots W_v)^n} (2\pi)^{\frac{-z}{2}} |\psi_\beta|^{\frac{-1}{2}} \frac{1}{p(Y)} \ .$$

This leads to the following expression for the objective function (use of a normal pose prior is assumed)

$$L(\beta) = -\frac{1}{2}(\beta - \beta_o)^T \psi_\beta^{-1} (\beta - \beta_0) + \textstyle\sum_i \ln \left[1 + \sum_{M_j} \frac{W_1 \cdots W_v}{m} \frac{1 - B_i}{B_i} \mathrm{G}_{\psi_{ij}}(Y_i - \mathcal{P}(M_j, \beta)) \right] \ . \qquad (24)$$

This objective function for evaluating pose hypotheses is a smooth function of the pose. The smoothness originates in the use of normal models of feature deviations, and in not committing to specific matches.

Methods of continuous optimization can be used to search for local maxima of this objective function.

The first term in the PMPE objective function (Equation 24) is due to the pose prior. It is a quadratic penalty for deviations from the nominal pose. The second term measures the degree of alignment of the object model with the image. It is a sum taken over image features of a smooth non-linear function that peaks up positively when the pose brings object features into alignment with the image feature in question. The logarithmic term will be near zero if there are no model features close to the image feature in question.

In a straightforward implementation of the objective function, the cost of evaluating a pose is $O(mn)$, since it is essentially a non-linear double sum over image and model features.

8. Expectation – Maximization Algorithm

The Expectation – Maximization (EM) algorithm was introduced in its general form by Dempster, Rubin and Laird (Dempster, Laird and Rubin, 1977). It is often useful for computing estimates in domains having two sample spaces, where the events in one are unions over events in the other. This situation holds among the sample spaces of Posterior Marginal Pose Estimation (PMPE) and MAP Model Matching. In the original paper, the wide generality of the EM algorithm is discussed, along with several previous appearances in special cases, and convergence results are described.

In this section, a specific form of the EM algorithm is described for use with PMPE. It is used for hypothesis refinement in the recognition experiments that are described in Section 10, and a connection to the Iterative Closest Point (ICP) algorithm (Besl and McKay, 1992) is described as a limiting case.

8.1. *Definition of EM Iteration*

In PMPE, the pose of an object, β, is estimated by maximizing its posterior probability, given an image:

$$\hat{\beta} = \arg\max_{\beta} p(\beta \mid Y) \quad .$$

A necessary condition for the maximum is that the gradient of the posterior probability with respect to the pose be zero, or equivalently, that the gradient of the logarithm of the posterior probability be zero:

$$\mathbf{0} = \nabla_\beta \ln p(\hat{\beta} \mid Y) \quad . \tag{25}$$

Imposing the condition of Equation 25 on the formula for the posterior probability given in Equation 21 yields the following,

$$\mathbf{0} = \nabla_\beta \left[\ln \frac{1}{p(Y)} + \ln p(\hat{\beta}) + \sum_i \ln p(Y_i \mid \hat{\beta}) \right]$$

or

$$\mathbf{0} = \frac{\nabla_\beta p(\hat{\beta})}{p(\hat{\beta})} + \sum_i \frac{\nabla_\beta p(Y_i \mid \hat{\beta})}{p(Y_i \mid \hat{\beta})} \quad . \tag{26}$$

Using Equation 23, the zero gradient condition of Equation 26 may now be expressed as follows,

$$\mathbf{0} = \frac{\nabla_\beta p(\hat{\beta})}{p(\hat{\beta})} + \sum_i \frac{\frac{1-B_i}{m} \sum_j \nabla_\beta \mathrm{G}_{\psi_{ij}}(Y_i - M_j \hat{\beta})}{\frac{B_i}{W_1 W_2 \cdots W_v} + \frac{1-B_i}{m} \sum_j \mathrm{G}_{\psi_{ij}}(Y_i - M_j \hat{\beta})} \quad .$$

With a normal pose prior,

$$p(\beta) = \mathrm{G}_{\psi_\beta}(\beta - \beta_0) \quad ,$$

and

$$\nabla_\beta p(\beta) = -p(\beta) \psi_\beta^{-1} (\beta - \beta_0) \quad .$$

The gradient of the other normal density is

$$\nabla_\beta \mathrm{G}_{\psi_{ij}}(Y_i - M_j \beta) = -\mathrm{G}_{\psi_{ij}}(Y_i - M_j \beta)\, M_j^T \psi_{ij}^{-1} (Y_i - M_j \beta) \quad . \tag{27}$$

Returning to the gradient condition, and using these expressions,

$$\mathbf{0} = \psi_\beta^{-1}(\hat{\beta} - \beta_0) + \sum_i \frac{\frac{1-B_i}{m} \sum_j \mathrm{G}_{\psi_{ij}}(Y_i - M_j \hat{\beta})\, M_j^T \psi_{ij}^{-1}(Y_i - M_j \hat{\beta})}{\frac{B_i}{W_1 W_2 \cdots W_v} + \frac{1-B_i}{m} \sum_j \mathrm{G}_{\psi_{ij}}(Y_i - M_j \hat{\beta})}$$

. Finally, the zero gradient condition may be expressed compactly as follows,

$$\mathbf{0} = \psi_\beta^{-1}(\hat{\beta} - \beta_0) + \sum_{ij} W_{ij} M_j^T \psi_{ij}^{-1} (Y_i - M_j \hat{\beta}) \quad , \tag{28}$$

with the following definition:

$$W_{ij} = \frac{\mathrm{G}_{\psi_{ij}}(Y_i - M_j \hat{\beta})}{\frac{B_i}{1-B_i} \frac{m}{W_1 W_2 \cdots W_v} + \sum_j \mathrm{G}_{\psi_{ij}}(Y_i - M_j \hat{\beta})} \quad . \tag{29}$$

Equation 28 has the appearance of being a linear equation for the pose estimate $\hat{\beta}$ that satisfies the zero gradient condition. Unfortunately, it is not a linear equation, because W_{ij} (the "weights") are not constants, they are functions of $\hat{\beta}$. To find solutions to Equation 28, the EM algorithm iterates the following two steps:

- Using the most recent pose estimate $\hat{\beta}$, re-evaluate the weights, W_{ij}, according to Equation 29. This is referred to as the E step.
- Treating the weights, W_{ij} as constants, solve Equation 28 as a linear equation for a new pose estimate $\hat{\beta}$. This is referred to as the M step.

The E step is so named because calculating the weights W_{ij} corresponds to taking the expectation of correspondence "indicator" random variables, given the image data, and belief in the most recent pose estimate. These random variables have value one if the i'th image feature corresponds to the j'th object feature, and zero otherwise. Thus the weights provide continuous-valued estimates of feature correspondences that vary between zero and one. W_{ij} will be maximal if the projection of model feature j is substantially closer to image feature i than is any other model feature, and it will approach zero if some other model feature is substantially closer to image feature i. Distances here are in terms of the Mahalanobis distance associated with ψ_{ij}. The weights may be interpreted as a "soft", or smooth measure of the relative proximity of the various model features to the image features.

The M step is so named because, in the exposition of the algorithm in (Dempster, Laird and Rubin, 1977), it corresponds to a Maximum Likelihood estimate. As discussed there, the algorithm is also amenable to use in MAP formulations, such as the present one. Here the M step corresponds to a MAP estimate of the pose, given that the current estimate of the weights is correct.

It seems somewhat ironic that, having abandoned the correspondences as being part of the hypothesis in the formulation of PMPE, a good estimate of them has re-appeared as a byproduct of a method for search in pose space. This estimate, the posterior expectation, is the minimum variance estimator.

Being essentially a local method of non-linear optimization, the EM algorithm needs good starting values in order to converge to the right local maximum. It may be started on either step. If it is started on the E step, an initial pose estimate is required. When started on the M step, an initial set of weights is needed.

An initial set of weights can be obtained from a partial hypothesis of correspondences in a simple manner. The weights associated with each set of corresponding features in the hypothesis are set to one, the rest to zero. Indexing methods are one source of such hypotheses. In Section 10, Angle Pair Indexing is used to generate candidate hypotheses. In this scenario, indexing provides initial alignments, these are refined using the EM algorithm, then they are verified by examining the value of the peak of the PMPE objective function that the refinement step found.

Connection to the Iterative Closest Point (ICP) Algorithm The alternating algorithm described above bears some similarity to the Iterative Closest Point (ICP) algorithm. As described in (Besl and McKay, 1992), the ICP algorithm is used to register curve and surface data to curves and surfaces respectively. Paraphrased into the present situation relating model and image features, the ICP algorithm may be summarized as follows.

- Using the most recent pose estimate, find for each projected model point the closest point in the image.
- Perform a least-squares pose adjustment in order to minimize the sum of the squared distances between the most recently identified pairs of image and projected model points.

There is clearly a close analogy between ICP and the EM algorithm applied to PMPE optimization as described above. Both algorithms alternate between a step that evaluates the current proximity of model and image features and a step that re-estimates the pose based on a weighted least squares adjustment where the weights are controlled by the proximity evaluation.

In the case of ICP, the weighting scheme is "hard" and assigns the values one and zero based on relative proximity ranking. The EM approach described above uses a "softer" scheme for assigning the weights to values between one and zero.

It is interesting to note that the EM scheme described above can be made to approach the ICP algorithm in a limiting case by reducing $|\psi_{ij}|$ towards zero by scaling ψ_{ij}. As this reduction occurs, the behavior of the EM weights will become "harder" in evaluating the relative proximity of features, and in the limit they will be equivalent to the weighting used in ICP, namely, the weights on nearest features will approach one, while the others approach zero. This occurs because the ratio of equal-variance Gaussians of differing residuals scale in proportion to an inverse power of $|\psi_{ij}|$.

The ICP algorithm depends on other mechanisms to avoid difficulties with gross outliers (which in the present context can be caused by partial occlusion of the model in the image). The formulations described here have some built-in robustness to this occurrence because the penalty incurred for the lack of image features in the vicinity of a projected model feature, while initially quadratic, becomes saturated when the distance to the nearest image feature grows.

8.2. Convergence

In the original reference (Dempster, Laird and Rubin, 1977), the EM algorithm was shown to have good (local) convergence properties under fairly general circumstances. It is shown that the likelihood sequence produced by the algorithm is monotonic, i.e., the algorithm never reduces the value of the objective function (or in this case, the posterior probability) from one step to the next. (Wu, 1983) claims that the convergence proof in the original EM reference is flawed, and provides another proof, as well as a thorough discussion. It is possible that the iteration will wander along a ridge, or become stuck in a saddle point.

In the recognition experiments reported in Section 10 the algorithm typically converges in 10 – 40 iterations.

8.3. Implementation Issues

Some thresholding methods were used speed up the computation of the E and M steps in the experiments reported in the Section 10.

The weights W_{ij} provide a measure of feature correspondence. As the algorithm operates, most of the weights have values close to zero, since most pairs of image and object feature don't align well for a given pose. In the computation of the M step, most terms were left out of the sum, based on a threshold for W_{ij}.

In the E step, most of the work is in evaluating the normal functions, which have quadratic forms in them. For the reason stated above, most of these expressions have values very close to zero. The evaluation of these expressions was made conditional on a threshold test applied to the residuals $Y_i - M_j\beta$. When the (x,y) part of the residual exceeded a certain length, zero was substituted for the value of the normal expression. Tables indexed by image coordinates might provide another effective way of implementing the thresholding here.

The value of the PMPE objective function is computed as a byproduct of the E step for little additional cost.

9. Potential-Function Formulations

This section discusses and contrasts Bounded Error (BE), Map Model Matching (MMM) and Posterior Marginal Pose Estimation (PMPE) by expressing them in a common framework based on "potential functions". Under certain conditions these criteria may be expressed as pose space minimization problems having the following form,

$$\sum_i \psi(\mathcal{P}^{-1}(Y_i, \beta)) \quad .$$

This objective function may be viewed as a sum over a "potential function" ψ evaluated on the image features after they have been "back-projected" into the model coordinate system.

The asymmetry of the role of model and image features in this formulation is due to the way correspondences are formulated (in Section 2) as a mapping from image features to model features and the background. The potential functions ψ provide a reward (by tending towards the negative direction) for image features that "backproject" near model features.

If the following conditions are met:

- the projection function $\mathcal{P}()$ is invertible and distance preserving
- pose priors are not present
- the feature covariance is isotropic: $\psi_{ij} = \sigma^2 I$
- the background probability is constant: $B_i = B$

then the MMM and PMPE potentials may be obtained as follows, from Equations 17 and 24 respectively:

$$\psi_{\mathrm{MMM}}(x) \equiv \min(\lambda,\ \min_j(\frac{1}{2\sigma^2}\ |x - M_j|^2)) \ ,$$

$$\psi_{\mathrm{PMPE}}(x) \equiv -\ln\left[1 + \tau \sum_j \mathrm{G}_\sigma(|x - M_j|)\right] \ ,$$

where $\tau \equiv \frac{W_1 \cdots W_V}{m}\ \frac{1-B}{B}$.

The MMM potential, ψ_{MMM} is obtained by taking the minimum of a constant function and quadratic functions that are centered on the model features. It may be interpreted as a robust chamfer function (as discussed in Section 6.2) constructed from the model.

The PMPE potential, ψ_{PMPE}, is obtained by taking the (negative) logarithm of one and a sum of normal functions that are centered on the model features.

An analogous potential function may be constructed for a bounded-error criterion (see Section 6.3),

$$\psi_{\mathrm{BE}}(x) \equiv \begin{cases} 0 & \text{if } \min_j |x - M_j| < \sigma \\ 1 & \text{otherwise} \end{cases} \ .$$

As discussed above, it can be advantageous to refine the pose estimates that result from the use of indexing methods for hypothesis generation in recognition, since such pose estimates are typically based on minimal sets of features. One attractive alternative to search in correspondence space uses local methods of continuous optimization, e.g. Powell's method, on an objective function in pose space. Since ψ_{BE} is not continuous, it is not suitable for local optimization methods. In practice, ψ_{PMPE} may be more suitable for local optimization than ψ_{MMM} because it has continuous derivatives, and in some experiments ψ_{PMPE} has demonstrated somewhat larger "capture range" than ψ_{MMM}.

10. Recognition Experiments

10.1. Combinatorial Search of MMM

In this section an experiment demonstrating use of the MMM objective function is described. The intent is to demonstrate the utility of the objective function in a domain of features that have significant fluctuations. The features are derived from video images. The domain is matching among features from low-resolution edge images. The point-radius feature model discussed in Section 5.2 is used. Oriented stationary statistics, as described in Section 3.3, are used to model the feature fluctuations.

Good solutions of the objective function of Equation 15 are sought by a search in correspondence space. Search over the whole exponential space is avoided by heuristic pruning.

An objective function that evaluates a configuration of correspondences, or match (described by Γ), may be obtained as follows:

$$\mathcal{L}(\Gamma) = \max_\beta L(\Gamma, \beta) \ .$$

This optimization is quadratic in β and is carried out by least squares. Sequential techniques are used to obtain constant cost for extending partial matches by one correspondence.

The space of correspondences may be organized as a directed-acyclic-graph (DAG) by the following parent-child relation on matches. A point in correspondence space, or *match* is a child of another match if there is some i such that $\Gamma_i = \perp$ in the parent, and $\Gamma_i = M_j$, for some j, in the child, and they are otherwise the same. Thus, the child has one more assignment to the model than the parent does. This DAG is rooted in the match where all assignments are to the background. All possible matches are reachable from the root.

Heuristic beam search, as described in (Shapiro, 1987), is used to search over matches for good solutions of $\mathcal{L}$. Success depends on the heuristic that there are not many impostors in the image. An impostor is a set of image features that scores well but is not a subset of the optimum match implied by the objective function. Another way of stating the heuristic is that the best match to $n+1$ object features is likely to contain the best match to n object features (for n beyond minimal values).

The search method used in the experiments employs a bootstrapping mechanism based on distinguished features. Object features 1, 2 and 3 are special, and must be detected. The scheme could be made robust by considering more initial triples of object features. Alternatively, indexing methods could be used as an efficient and robust means to initiate the search. Indexing methods are described by Clemens and Jacobs (Clemens and Jacobs, 1990).

The algorithm that was used is outlined below.

```
BEAM-SEARCH(M, Y)
    CURRENT ← {Γ: exactly one image feature is matched to each of M1 M2 and M3}
    ;; the rest are assigned to the background.
    Prune CURRENT according to ℒ. Keep 50 best.
    Iterate to Fixpoint:
        Add to CURRENT all children of members of CURRENT
        Prune CURRENT according to ℒ. Keep N best.
        ;; N is reduced from 20 to 5 as the search proceeds.
    Return(CURRENT)
```

The sizes of CURRENT were determined empirically. Sometimes an extension of a match will produce one that is already in CURRENT, that was reached in a different sequence of extensions. When this happens, the matches are coalesced. This condition is efficiently detected by testing for near equality of the scores of the items in CURRENT. Because the features are derived from observations containing some random noise, it is very unlikely that two hypotheses having differing matches will achieve the same score, since the score is partly based on summed squared errors.

The maximum depth that this search can achieve is M, the number of model features. The complexity of the initial construction of CURRENT is $O(N^3)$, and the worst case complexity of the iteration is $O(M^2N)$. With the values of M and N for this experiment, the initial construction used about 2 million evaluations, and the worst case usage of the iteration is about 200K evaluations.

The search method described in the previous section was used to obtain good matches in a domain of features that have significant fluctuations. The features were derived from real images. A linear projection model was used.

The object model was derived from a set of 16 images. In this set, only the light source position varied.

The features used for matching were derived from edge maps obtained using the Canny edge detector (Canny, 1986). The smoothing standard deviation is eight pixels – these are low resolution edge maps. The object model edges were derived from a set of 16 edge maps, corresponding to the 16 images described above. The object model edges are essentially the mean edges with respect to fluctuations induced by variations in lighting. (Low resolution edges are sensitive to lighting.)

The features used in matching are shown in Figure 6. These are point-radius features, as described in Section 5.2. The point coordinates of the features are indicated by dots, while the normal vector and curvature are illustrated by arc fragments. Each feature represents 30 edge pixels. The 40 object features appear in the left image, the 125 image features in the right image. The distinguished features used in the bootstrap of the search are indicated with circles. The object features have been transformed to a new pose to insure generality.

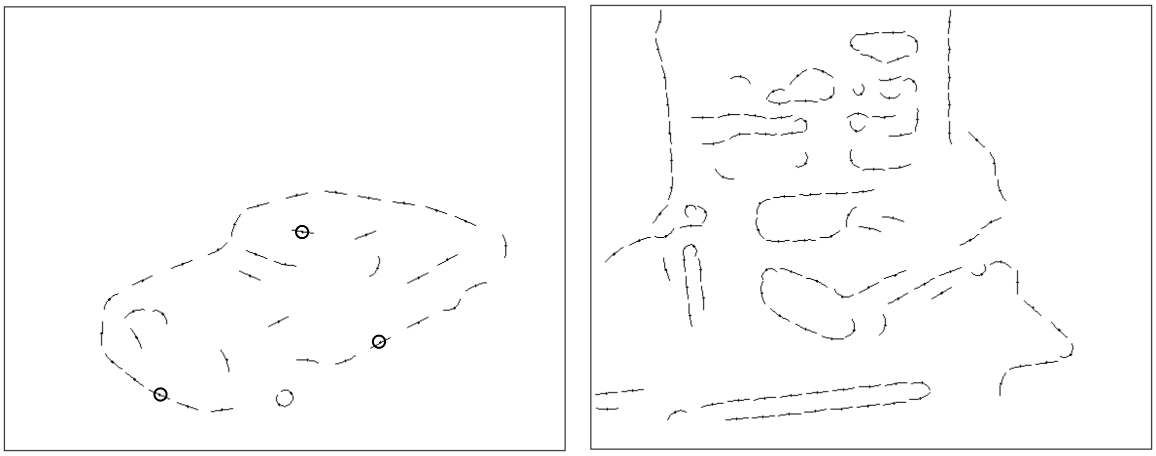

Fig. 6. Point-Radius Features used for Matching: Distinguished Features are Circled

The parameters that appear in the objective function are: B, the background probability and $\hat{\psi}$, the stationary feature covariance. These were derived from a match done by hand in the example domain. The oriented stationary statistics model of Section 3.3 was used here. (A normal model of feature fluctuations is implicit in the objective function of Equation 13. This was found to be a good model in this domain.)

A loose pose prior was used. This pose prior is illustrated in Figure 7. The prior places the object in the upper left corner of the image. The one standard deviation intervals of position and angle are illustrated. The one standard deviation variation of scale is 30 percent. The actual pose of the object is within the indicated one standard deviation bounds. This prior was chosen to demonstrate that the method works well despite a loose pose prior.

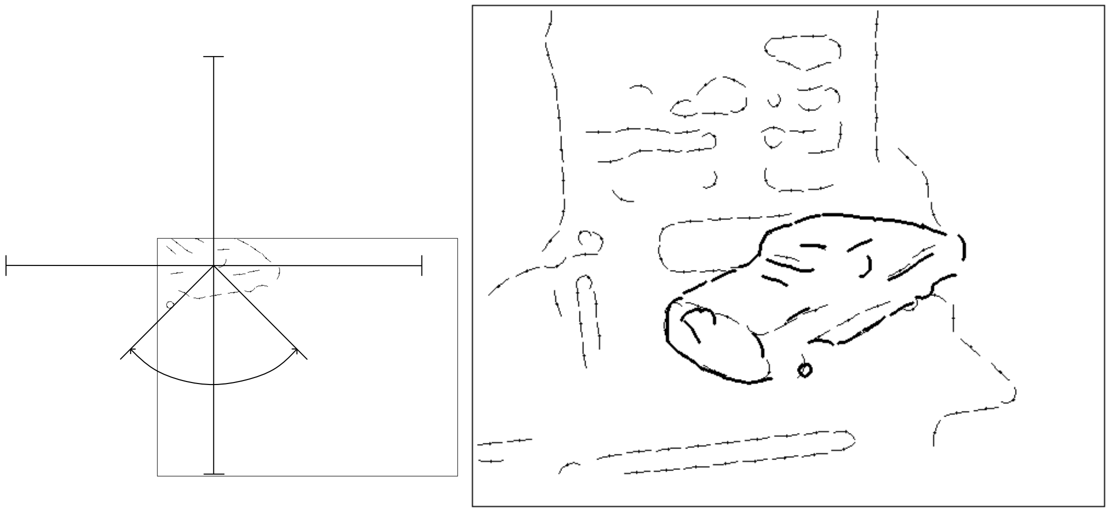

Fig. 7. Left: Pose Prior used in Search – The vertical and horizontal bars represent the standard deviation in position about the mean, and the arc represents the standard deviation in angle. Right:Best Match Results

The best results of the beam search appear in Figure 7. The object features are delineated with heavy lines. They are located according to the pose associated with the best match. There were 18 correspondences in the best match.

Extensions MAP Model Matching performs well on low resolution imagery in which feature uncertainty is significant. It could be used to bootstrap a coarse-fine approach to model matching, yielding good results with reasonable running times. Coarse-fine approaches have proven successful in stereo matching applications. (See Grimson (Grimson, 1985) and Barnard (Barnard, 1987).) A coarse-fine strategy is straightforward in the framework described here. In a hierarchy, the pose estimate from solving the objective function at one scale is used as a prior for the estimation at the next. Having a good prior on the pose will reduce the amount of searching required at high resolution.

10.2. PMPE Experiments

2D Recognition Experiments The experiments described in this section use the EM algorithm to carry out local searches in pose space of the PMPE objective function. This is used for evaluating and refining alignments that are generated by Angle Pair Indexing, (Wells, 1992). A coarse – fine approach is used in refining the alignments produced by Angle Pair Indexing. To this end, two sets of features are used, coarse features and fine features.

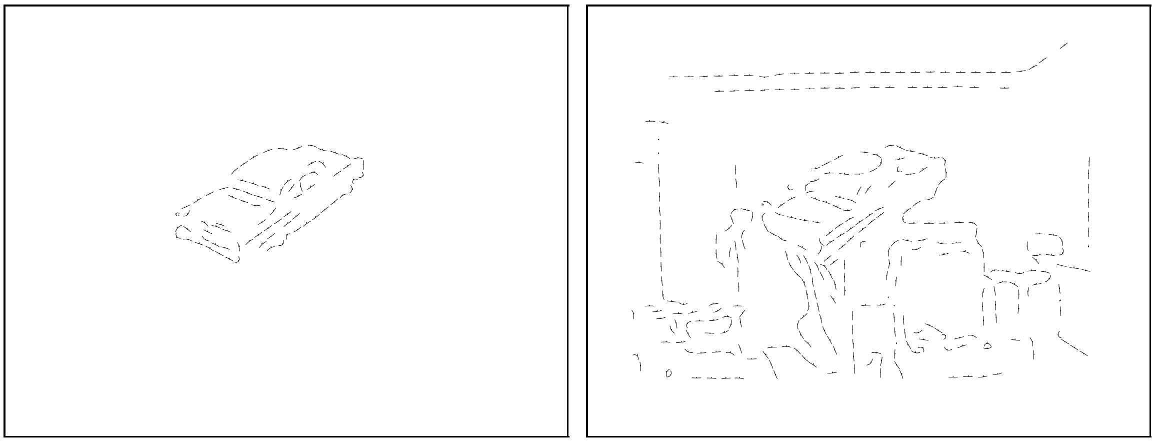

Fig. 8. Coarse Model and Image Features

The model features were derived from Mean Edge Images, (Wells, 1992). They are essentially edges averaged over a series of images in which the illumination varied. The standard deviation of the smoothing that was used in preparing the model and image edge maps was 3.97 for the coarse features, and 1.93 for the fine features. The edge curves were broken arbitrarily every 20 pixels for the coarse features, and every 10 pixels for the fine features. Point-radius features were fitted to the edge curve fragments, as described in Section 5.2. The coarse model and image features appear in Figure 8, the fine model and image features appear in Figure 9. There are 81 coarse model features, 334 coarse image features, 246 fine model features, and 1063 fine image features. The features are similar to, but not the same as those used in the previous section.

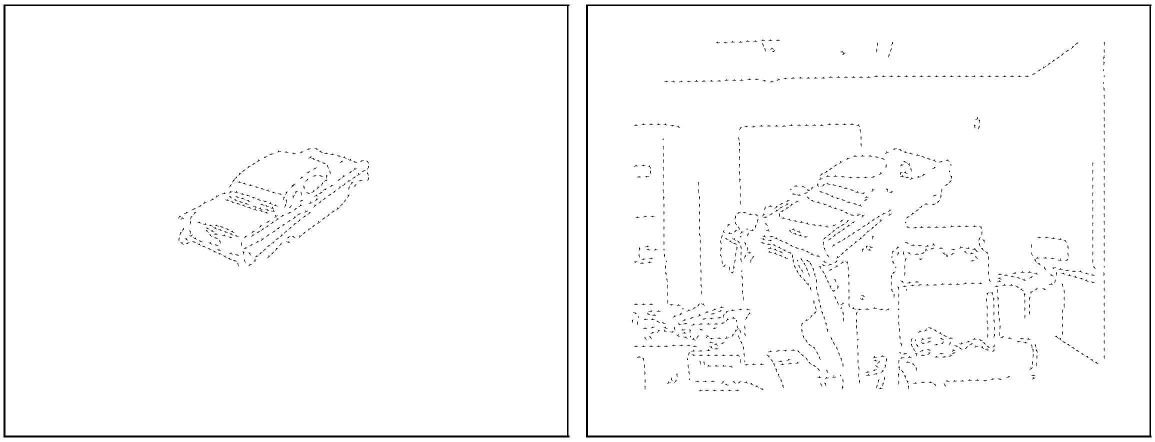

Fig. 9. Fine Model and Image Features

The oriented stationary statistics model of feature fluctuations was used. The parameters (statistics) that appear in the PMPE objective function, the background probability and the covariance matrix for the oriented stationary statistics, were derived from matches that were done by hand. These training matches were also used in the empirical study of the goodness of the normal model for feature fluctuations discussed in Section 3.2, and they are described there.

Generating Alignments Initial alignments were generated using Angle Pair Indexing (Wells, 1992) on the coarse features. This method precomputes a table that is consulted at recognition time. Figure 10 illustrates two of the candidate alignments by superimposing the object in the images at the pose associated with the initial alignment implied by the pairs of feature correspondences.

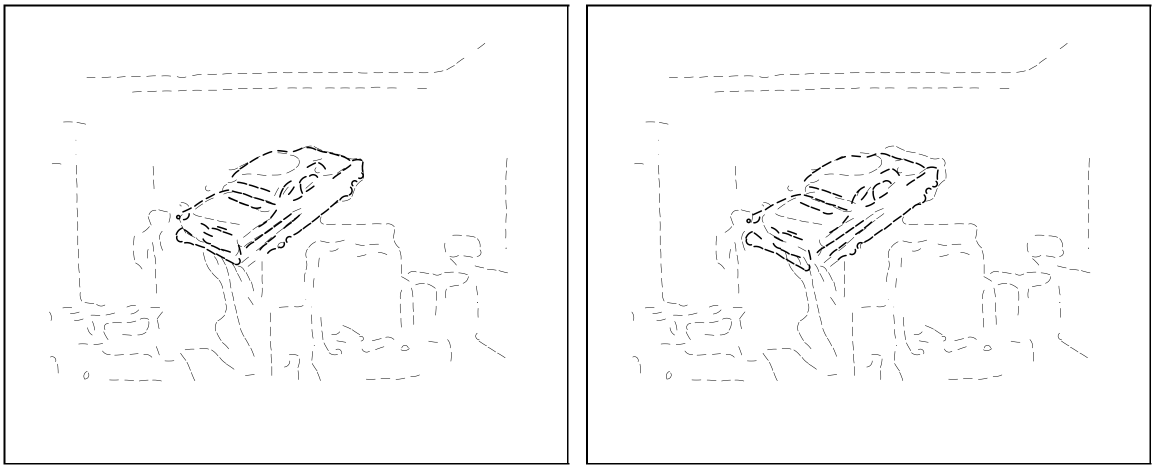

Fig. 10. Two Indexed Hypotheses

Scoring Indexer Alignments The initial alignments were evaluated in the following way. The indexing process produces hypotheses consisting of a pair of correspondences from image features to object features. These pairs of correspondences were converted into an initial weight matrix for the EM algorithm. The M step of the algorithm was run, producing a rough alignment pose. The pose was then evaluated using the E step of the EM algorithm, which computes the value of the objective function as a side effect (in addition to a new estimate of the weights). Thus, running one cycle of the EM algorithm, initialized by the pair of correspondences, generates a rough alignment, and evaluates the PMPE objective function for that alignment.

Refining Alignments Figure 11 shows a closer view of the the results of a coarse-fine alignment. The top image displays the best scoring initial alignment from Angle Pair Indexing. The roof appears to be well aligned (too well, it turns out). Additionally, many of the model features are systematically lower than their counterparts in the image features. The initial alignment was refined by running the EM algorithm to convergence using the coarse features and statistics. The coarse refinement was refined further by running the EM algorithm to convergence with the fine features and statistics. The result of this fine refinement is shown in the lower image, and over the video image in Figure 13. The overall agreement is better, and the roof features have pulled away properly.

Ground truth for the pose is available in this experiment, as the true pose is the null pose. The pose before refinement is

$$[.99595, -0.0084747, -0.37902, 5.0048]^T \ ,$$

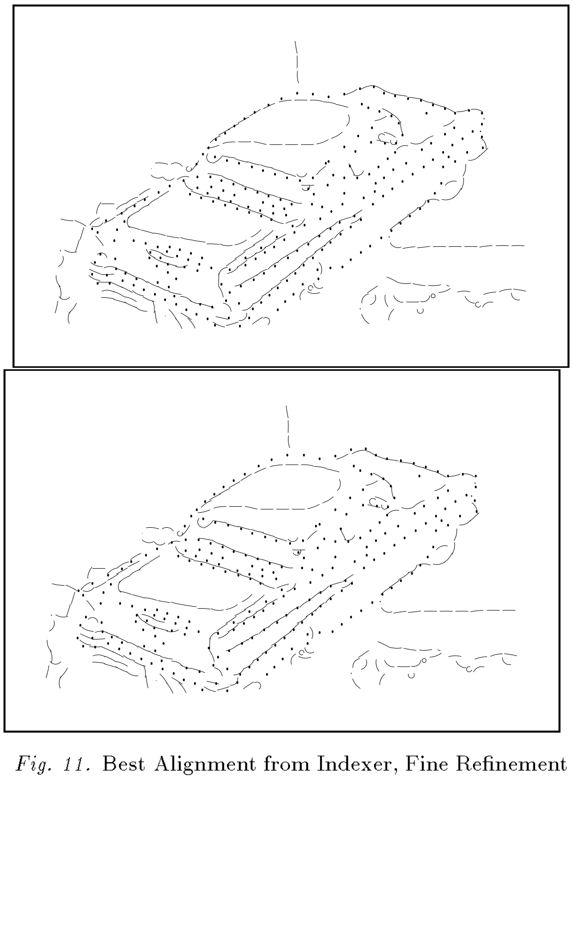

Fig. 11. Best Alignment from Indexer, Fine Refinement

and after the refinement it is

$$[1.00166, 0.0051108, 0.68621, -1.7817]^T \ .$$

The encoding of these poses is described in Section 5.2 (the null pose is $[1,0,0,0]^T$.) The initial pose is in error by about .01 in scale and 5 pixels in position. The final pose errs by about .005 in scale and 1.8 pixels in position. Thus scale accuracy is improved by a factor of about two, and position accuracy is improved by factor of about three.

In these experiments, less that 15 iterations of the EM algorithm were needed for convergence.

Evaluating Random Alignments An experiment was performed to test the utility of PMPE in evaluating randomly generated alignments. Correspondences among the coarse features described in Section 10.2 having assignments from two image features to two model features were randomly generated, and evaluated as in Section 10. 19118 random alignments were generated, of which 1200 had coarse scores better than -30.0 (the negative of the PMPE objective function). Among these 1200, one was essentially correct.

With coarse – fine refinement, the best scoring random alignment converged to essentially the same pose as the best refinement from the indexing experiment, with fine score -355.069. The next best scoring random alignment converged to a grossly wrong pose, with fine score -149.064. This score provides some indication of the "noise" level in the fine PMPE objective function in pose space.

This test, though not exhaustive, produced no false positives, in the sense of a bad alignment with a coarse score better than that of the correct alignment. Additionally, the fine score of the refinement of the most promising incorrect random alignment was significantly lower than the fine score of the (correct) refined best alignment.

Convergence with Occlusion The convergence behavior under occlusion of the EM algorithm with PMPE was evaluated using the coarse features described in Section 10.2. Images features simulating varying amounts of occlusion were prepared by shifting a varying portion of the image. One of these images, along with results of coarse – fine refinement using the EM algorithm are shown in Figure 12.

The starting value for the pose was the correct (null) pose. The refinements converged to good poses in all cases, demonstrating that the method can converge from good alignments with moderate amounts of occlusion.

The final fine score in the most occluded example was lower than the score of one of the random alignments described in Section 10. This indicates that as the amount of occlusion increases, the objective function will not be maximal at the correct pose. In this experiment it happens before the method fails to converge properly locally.

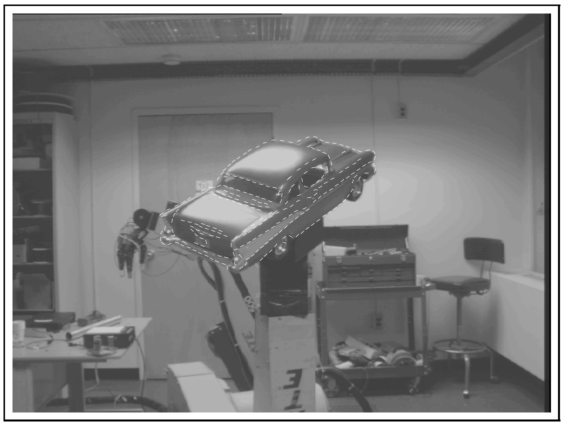

Fig. 13. Result of Fine Refinement

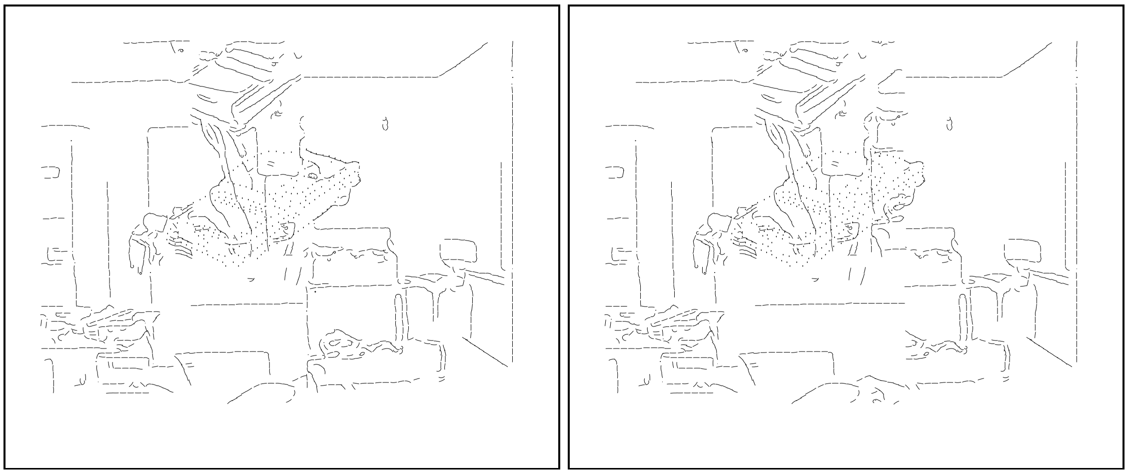

Fig. 12. Fine Convergences with Occlusion

10.3. 3D Recognition Experiment

Refining 3D Alignments using PMPE and the EM Algorithm This section demonstrates use of the EM algorithm with PMPE to refine alignments in 3D. The linear combination of views method is used to accommodate a limited amount of out of plane rotation. A two-view variant of LCV, described in Section 5.4, is used.

A coarse – fine approach was used. In this experiment, coarse features were not used. Instead, coarse PMPE scores were computed by smoothing the PMPE objective function in the following manner:

$$\hat{\psi} \leftarrow \hat{\psi} + \psi_s \quad .$$

The effect of the smoothing matrix ψ_s is to increase the spatial scale of the covariance matrices that appear in the objective function. The smoothing matrix was

$$\mathrm{DIAG}((7.07)^2, (3.0)^2) \quad .$$

These numbers are the amounts of additional (artificial) variance added for parallel and perpendicular deviations, respectively, in the oriented stationary statistics model.

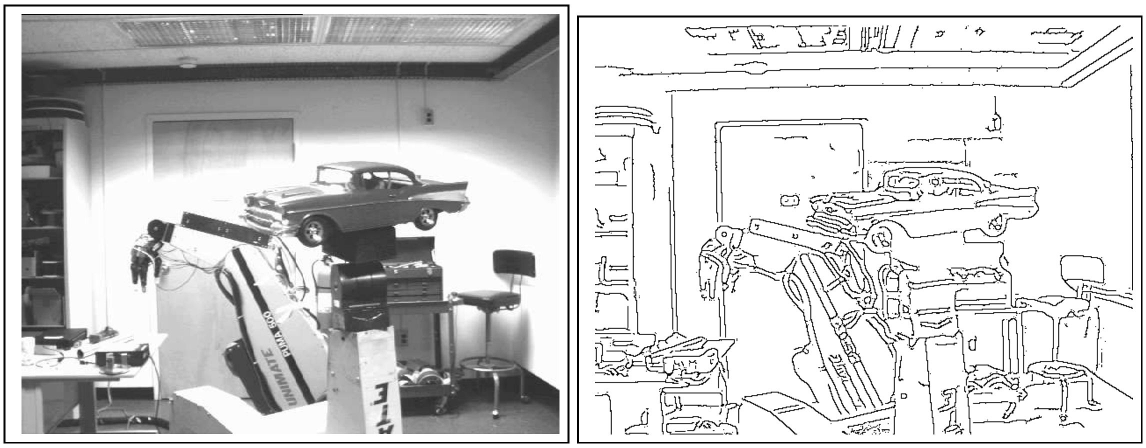

Fig. 14. Grayscale Image and Image Edges

The video test image and edge image are shown in Figure 14. It differs from the model images by a significant 3D translation and out of plane rotation.

The object model was derived from two Mean Edge Images that were constructed as described in (Wells, 1992).

The smoothing used in preparation of the edge maps had 1.93 pixels standard deviation, and the edge curves were broken arbitrarily every 10 pixels. Point-radius features were fitted to the edge curve fragments, as described in Section 5.2, for purposes of display and for computing the oriented stationary statistics, although the features used with PMPE and the EM algorithm were simply the X and Y coordinates of the centroids of the curve fragments. Both views of the model features are shown in Figure 15. The linear combination of views method requires correspondences among the model views. These were established by hand, and are displayed in Figure 16.

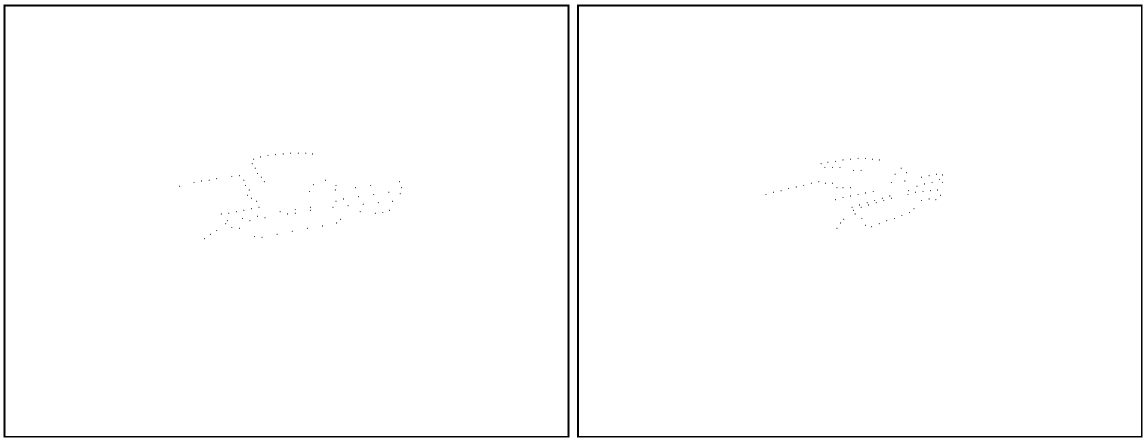

Fig. 15. Model Features (Both Views)

The viewpoints of the model images were 36.6 degrees apart on the view sphere, and the test image viewpoint was roughly in the center of an equilateral triangle of viewpoints containing the model image views.

The oriented stationary statistics model of feature fluctuations was used (this is described in Section 3.3). As in Section 10.2, the parameters (statistics) that appear in the PMPE objective function, the background probability and the covariance matrix for the oriented stationary statistics, were derived from matches done by hand.

A set of four correspondences was established manually from the image features to the object features. These correspondences are intended to simulate an alignment generated by an indexing system. Indexing systems that are suitable for 3D recognition are described by Clemens and Jacobs (Clemens and Jacobs, 1990) and Jacobs (Jacobs, 1992). The rough alignment was obtained from the correspondences by one cycle of the EM algorithm, as described above in Section 10. They are displayed in Figure 17, where the four corresponding features appear circled. A coarse alignment was then obtained by running the EM algorithm to convergence with smoothing. This alignment was refined further by running the EM algorithm again, without smoothing.

The resulting alignment is shown in Figure 18. Most of the model features have aligned well. The discrepancy in the forward wheel well may be due to inaccuracies in the LCV modeling process, perhaps in the feature correspondences. This figure demonstrates good results for aligning a smooth 3D object having six degrees of freedom of motion. In these figures, the image features are shown as curve fragments for clarity, although only the point locations were used in the experiment.

10.4. Range Feature Experiment

In this section, an experiment is described where PMPE and the EM algorithm are used in a coarse – fine scheme to estimate the pose of a vehicle appearing in synthetic range images. [3]

Preparation of Features The preparation of the features used in the experiment is summarized in Figure 19. The features were oriented-range features, as described in Section 5.3. Two sets of features were prepared, the "model features", and the "image features".

The object model features were derived from a synthetic range image of an M35 truck that was created using the ray tracing program associated with the BRL CAD Package (Dykstra and Muuss, 1987). The ray tracer was modified to produce range images instead of shaded images.

In order to simulate a laser radar, the synthetic range image described above was corrupted with simulated laser radar sensor noise, using a sensor noise model that is described by Shapiro, Reinhold, and Park (Shapiro, Reinhold and Park, 1986). In this noise model, measured ranges are either valid or anomalous. Valid measurements are normally distributed, and anomalous measurements are uniformly distributed. The corrupted range image appears in Figure 20. To simulate post sensor processing, the corrupted image was "restored" via a statistical restoration method of Menon and Wells (Menon and Wells, 1990). The restored range image appears in 21.

Oriented range features, as described in Section 5.3, were extracted from the synthetic range image, for use as model features – and from the restored range image. The features were extracted from the range images in the following manner. Range discontinuities were located by thresholding neighboring pixels, yielding range discontinuity curves. These curves were then segmented into approximately 20-pixel-long segments via a process of line segment approximation. The line segments (each representing a fragment of a range discontinuity curve) were then converted into oriented range features in the following manner. The X and Y coordinates of the feature were obtained from the mean of the pixel coordinates. The normal vector to the pixels was gotten via least-squares line fitting. The range to the feature was estimated by taking the mean of the pixel ranges on the near side of the discontinuity. This information was packaged into an oriented-range feature, as described in Section 5.3. The model features are shown in Figure 20. Each line segment represents one oriented-range feature, the ticks on the segments indicate the near side of the range discontinuity. There are 113 such object features.

A set of 62 similar features was derived from the restored range image. Some modeled features were lost due to the corruption and restoration of the range image. The set of image features was

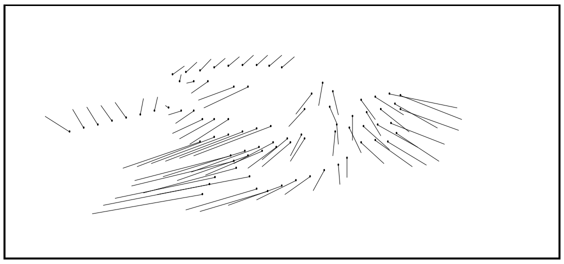

Fig. 16. Model Correspondences

prepared from these features by randomly deleting half of the features, transforming the survivors according to a test pose, and adding sufficient randomly generated features so that $\frac{1}{8}$ of the features are due to the object. There are 248 image features (Figure 21).

Coarse-Fine Method A coarse-fine search method was used for finding maxima of the pose-space objective function. Two levels of smoothing the objective function were used. Peaks, initially located at the coarsest scale, are used as starting values for a search at the next (less smooth) scale. Finally, results of the second level search are used as the initial values for search in the un-smoothed objective function. This coarse-fine method combines the accuracy of the un-smoothed objective function with the larger region of convergence of the smoothed objective function.

The objective function was smoothed by replacing the stationary feature covariance matrix $\hat{\psi}$ in the following manner:

$$\hat{\psi} \leftarrow \hat{\psi} + \psi_s \quad .$$

The effect of the smoothing matrix ψ_s is to increase the spatial scale of the covariance matrices that appear in the objective function. The smoothing matrices used in the experiment were as follows,

$$\mathrm{DIAG}((2.0)^2, (2.0)^2, (.01)^2, (.01)^2) \quad ,$$

and

$$\mathrm{DIAG}((5.0)^2, (5.0)^2, (.025)^2, (.025)^2) \quad .$$

where DIAG$(\cdot)$ constructs diagonal matrices from its arguments. These smoothing matrices were determined empirically.

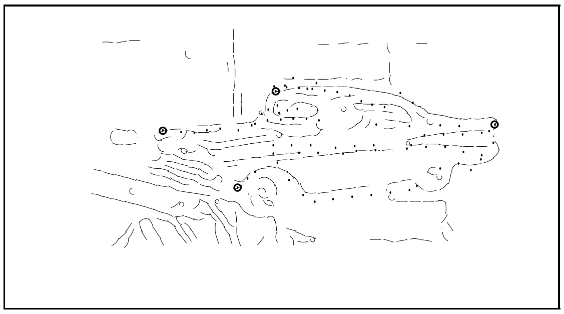

Fig. 17. Initial Alignment

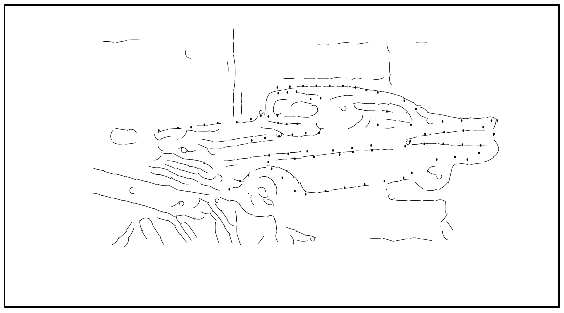

Fig. 18. Fine Refined Alignment

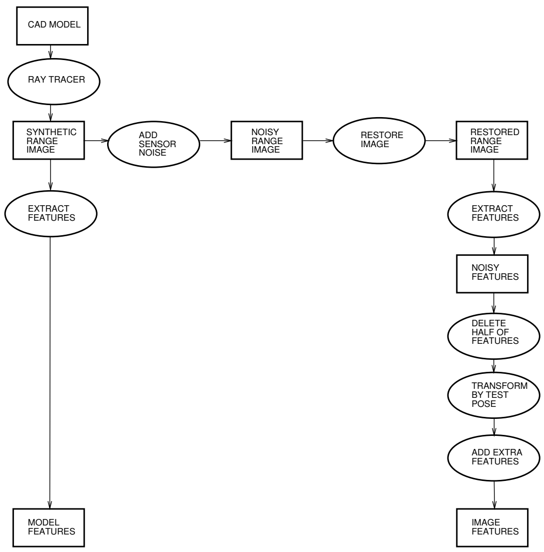

Fig. 19. Preparation of Features

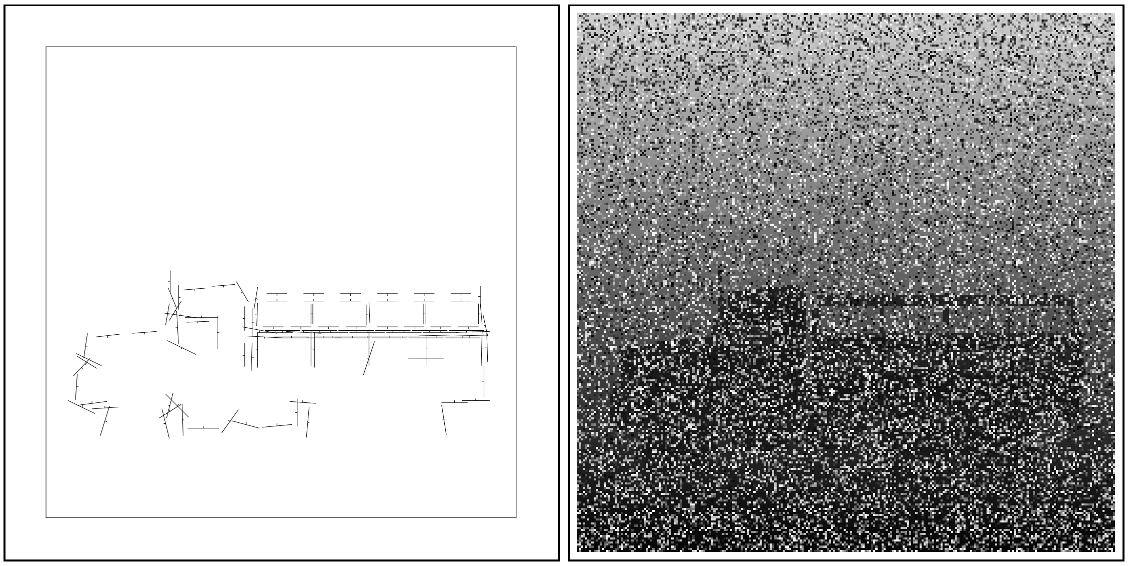

Fig. 20. Model Features and Noisy Range Image

Results An image displaying a sequence of poses from an EM iteration at the coarsest scale appears in Figure 22. The algorithm converged in 21 iterations.

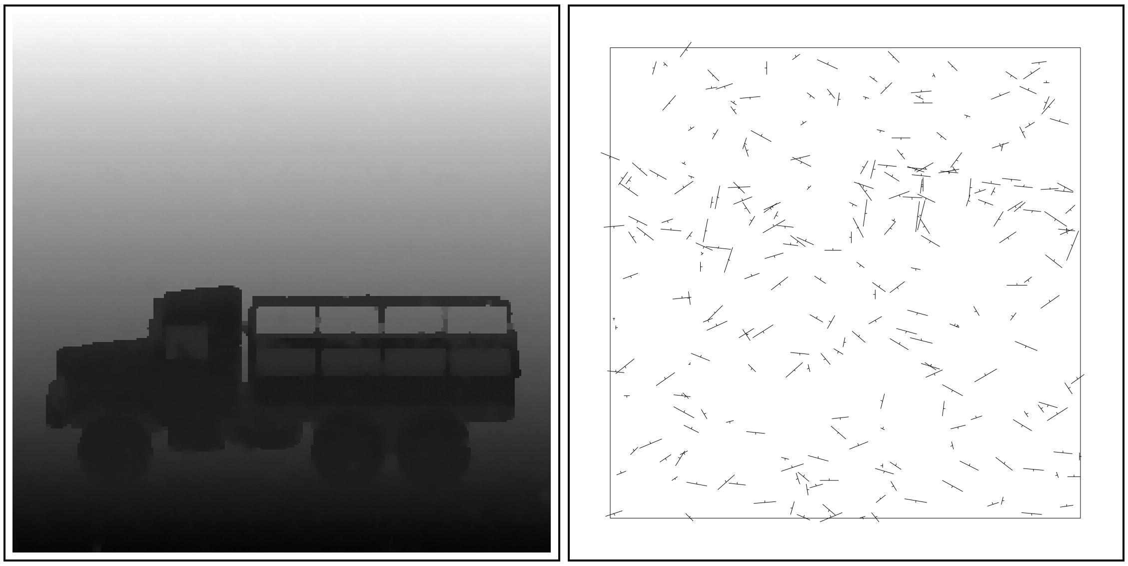

Fig. 21. Restored Range Image and Image Features

11. Related Work

Object recognition has previously been posed as finding the maxima of an objective function. The work of Fischler and Elschlager (Fischler and Elschlager, 1973) is an early example of this approach. Beveridge, Weiss and Riseman (Beveridge, Weiss and Riseman, 1989) use an objective function for line segment based recognition that is similar to that of MMM. In their work, the penalty for deviations is quadratic, while the reward for correspondence is non-linear (exponential) in the amount of missing segment length. (By contrast, the reward in MMM is, for stationary models, linear in the length of aggregate features.) The trade-off parameters in their objective function were determined empirically. Their system performs well in a domain of real images.

The HYPER vision system of Ayache and Faugeras (Ayache and Faugeras, 1986) uses sequential linear-least-squares pose estimation as well as the linear 2D point feature and projection model that inspired the 2D point-radius model. HYPER is described as a search algorithm. Different criteria are used to evaluate candidate matches and to evaluate competing "whole" hypotheses. An ad hoc threshold is used for testing a continuous measure of the metrical consistency of candidate match extensions. Whole match hypotheses are evaluated according to the amount of image feature accounted for – although not according to overall metrical consistency. HYPER works well on real images of industrial parts.

Bayesian approaches to aspects of recognition have been previously reported. Goad outlined a Bayesian strategy of match evaluation based on feature and background statistics in his paper on automatic programming for model-based vision (Goad, 1986). In his system, search was controlled by thresholds on probabilistic measures of the reliability and plausibility of matches.

Lowe describes in general terms the application of Bayesian techniques in his book on Visual Recognition (Lowe, 1985). He treats the minimization of expected running time of recognition. In addition he discusses selection among numerous objects.

Object recognition systems often use a strategy that can be summarized as a search for the maximal matching that is consistent. Consistency is frequently defined to mean that the matching image feature is within finite bounds of its expected position (bounded-error models).

Such an approach may be cast in the framework defined here by assuming uniform probability density functions for the feature deviations. Pose solution with this approach is likely to be more complicated than the sequential linear-least-squares method that can be used when feature deviations have normal models. Cass (Cass, 1992b) (Cass, 1992a) describes an approach to visual object recognition that searches in pose space for maximal alignments under the bounded-error model. The pose-space objective function used there is piecewise constant, and is thus not amenable to local search methods. The search is based on a geometric formulation of the constraints on feasible transformations. It performs well on occluded or fragmented real images. Generalization to other error models is described.

Burns and Riseman (Burns and Riseman, 1992) and Burns (Burns, 1992) describe a classification based recognition system. They focus on the use of description networks for efficiently searching among multiple objects with a recursive indexing scheme.

Fua and Hanson (Fua and Hanson, 1989b) (Fua and Hanson, 1989a) describe a general objective function approach to image understanding. They use a minimum description length (MDL) criterion that is designed to work with generic object models, while the approach presented here is designed for specific object models.

Cohen and Cooper (Cohen and Cooper, 1988) describe a decision-theoretic approach to 3D vision that has been used to estimate the parameters of texture patches and polynomial curves.

Bayesian networks have been used for inference in model based computer vision by Levitt, Binford, and Mann (Levitt, Binford and Mann, 1988). The Bayesian inference used in this paper is simpler, primarily because of the focus on recognizing a single specific object. The systems described here might be utilized as single nodes in Bayesian networks, as they generate the requisite posterior probabilities about the presence of their objects.

There are formal similarities between PMPE and some statistical laser radar processors described by Green (Green, Jr., 1992) and Green and Shapiro (Green, Jr. and Shapiro, 1992). They present a theory of Maximum Likelihood laser radar range profiling, and discuss statistically optimal detectors and recognizers. The single pixel statistics are described by a mixture of uniform and normal components. Range profiling is implemented using the EM algorithm. Under some circumstances, least squares provides an adequate starting value. A continuation-style variant is described, where a range accuracy parameter is varied between EM convergences from a coarse value to its true value. Green (Green, Jr., 1992) computes Cramer-Rao bounds for the complete-data case of Maximum Likelihood range profile estimator, and compares simulated and real-data performance to the limits. Efficiency of the PMPE estimator is discussed in (Wells, 1992).

There are some connections between PMPE and standard methods of robust pose estimation, like those described by Haralick (Haralick et al., 1989), and Kumar and Hanson (Kumar and Hanson, 1989). Both can provide robust estimates of the pose of an object, based on an observed image. The main difference is that the standard methods require specification of the feature correspondences, while PMPE does not – by considering all possible correspondences. PMPE requires a starting value for the pose (as do standard methods of robust pose estimation that use non-convex objective functions).

A limiting-case connection between the Iterative Closest Point (ICP) algorithm (Besl and McKay, 1992) and the use of the EM algorithm with PMPE was described above in Section 8.1.

Thompson has described a method where the initial pose from and indexing system is iteratively improved by making additional correspondences and updating the pose estimate (Thompson, 1989).

As mentioned above, Yuille, Geiger and Bülthoff (Yuille, Geiger and Bülthoff, 1990) discussed computing disparities in a statistical theory of stereo where a marginal is computed over matches. Yuille extends this technique (Yuille, 1990) to other domains of vision and neural networks, among them winner-take-all networks, stereo, long-range motion, the traveling salesman problem, deformable template matching, learning, content addressable memories, and models of brain development. In addition to computing marginals over discrete fields, the Gibbs probability distribution is used. This facilitates continuation-style optimization methods by variation of the temperature parameter. There are some similarities between this approach and using coarse-fine with the PMPE objective function.

Edelman and Poggio (Edelman and Poggio, 1990) describe a method of 3D recognition that uses a trained Generalized Radial Basis Function network. Their method requires correspondences to be known during training and recognition. One similarity between their scheme and PMPE is that both are essentially arrangements of smooth unimodal functions.

There is a similarity between Posterior Marginal Pose Estimation and Hough transform (HT) methods. Roughly speaking, HT methods evaluate parameters by accumulating votes in a discrete parameter space, based on observed features. (See the survey paper by Illingworth and Kittler (Illingworth and Kittler, 1988) for a discussion of Hough methods.)

In a recognition application, as described here, the HT method would evaluate a discrete pose by counting the number of feature pairings that are exactly consistent somewhere within the cell of pose space. As stated, the HT method has difficulties with noisy features. This is usually addressed by counting feature pairings that are exactly consistent somewhere nearby the cell in pose space.

The utility of the HT as a stand-alone method for recognition in the presence of noise is a topic of some controversy. This is discussed by Grimson in (Grimson, 1990), pp. 220. This might be due to an unsuitable noise model implicit in the Hough Transform.

In contrast to the HT, Posterior Marginal Pose Estimation evaluates a pose by accumulating the logarithm of posterior marginal probability of the pose over image features.

The connection between HT methods and parameter evaluation via the logarithm of posterior probability has been described by Stephens (Stephens, 1990). Stephens proposes to call the posterior probability of parameters given image observations "The Probabilistic Hough Transform". He provided an example of estimating line parameters from image point features whose probability densities were described as having uniform and normal components. He also states that the method has been used to track 3D objects.

In Section 2.2 assignments of image features to the various model features were assumed to be equally likely. One route to more accurate modeling of correspondences would exploit bottom-up saliency processes to suggest which image features are most likely to correspond to the object. One such process in described by Shashua and Ullman and (Shashua and Ullman, 1988). If we are provided a per-feature empirical measure of saliency S_i, we could incorporate this information by constructing $p(\Gamma_i = \perp | S_i)$ via the use of Bayes' rule and training data.

Lipson (Lipson, 1992) describes a non-statistical method for refining alignments that iterates solving linear systems. It matches model features to the nearest image feature under the current pose hypothesis, while the method described here entertains matches to all of the image features, weighted by their probability. Lipson's method was shown to be effective and robust in an implementation that refines alignments under Linear Combination of Views. Thompson (Thompson, 1989) describes a somewhat similar iterative method of pose refinement under affine transformation.

12. Conclusions

This paper has explored the application of statistical estimation to the problem of model-based object recognition. The widely-used bounded-error recognition criterion corresponds to correspond to a theory based on uniform models of feature fluctuations.

Evidence was presented indicating, that in some domains, feature fluctuations are better modeled by normal rather than uniform distributions. This motivated the development of new maximum-likelihood and MAP recognition formulations which are based on normal feature models, MAP Model Matching (MMM), which is based on specific correspondences, and Posterior Marginal Pose Estimation (PMPE), which does not commit to particular correspondences, and is amenable to search via the EM algorithm.

MMM was shown to be effective for searching among feature correspondences, and when organized as a search in pose space, it is shown to be equivalent to a robust variant of chamfer matching. PMPE was shown to be an effective criterion for local search in pose space, which is an attractive alternative to combinatorial search for the refinement of hypotheses in recognition.

An extension to the alignment approach that may be summarized as *align refine verify* was proposed.

Acknowledgements

I thank Professors Eric Grimson and Jeffrey Shapiro for stimulating discussions about the material described here. I thank them and an anonymous reviewer for constructive advice on earlier versions of this work. Al Gschwendtner receives thanks for my opportunity to pursue part of this work at MIT Lincoln Laboratory

Notes

1. Early versions of this work appeared in (Wells, 1990) and (Wells, 1991). A more detailed version appears in (Wells, 1992).
2. Early versions of this work appeared in (Wells, 1990) and (Wells, 1991). A more detailed version appears in (Wells, 1992).
3. A more detailed description of this material appeared in (Wells, 1993).

References

Ayache, N. and Faugeras, O. (1986). HYPER: A New Approach for the Recognition and Positioning of Two-Dimensional Objects. *IEEE Transactions PAMI*, PAMI-8(1):44–54.

Barnard, S. (1987). Stereo Matching by Hierarchical, Microcanonical Annealing. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 832–835.

Barrow, H., Tenenbaum, J., Bolles, R., and Wolf, H. (1977). Parametric Correspondence and Chamfer Matching: Two New Techniques for Image Matching. In *Proc. 5th Int. Joint Conf. Artificial Intelligence*, pages 659–663.

Beck, J. and Arnold, K. (1977). *Parameter Estimation in Science and Engineering.* John Wiley & Sons.

Besl, P. and Jain, R. (1985). Three-Dimensional Object Recognition. *Computing Surveys*, 17:75–145.

Besl, P. and McKay, N. (1992). A Method for Registration of 3-D Shapes. *IEEE Transactions PAMI*, 14(2):239–255.

Beveridge, J., Weiss, R., and Riseman, E. (1989). Optimization of 2-Dimensional Model Matching. In *Proceedings: Image Understanding Workshop*, pages 815–830. Morgan Kaufmann.

Blake, A. and Zisserman, A. (1987). *Visual Reconstruction.* MIT Press.

Borgefors, G. (1988). Hierarchical Chamfer Matching: A Parametric Edge Matching Algorithm. *IEEE Transactions PAMI*, 10(6):849–865.

Breuel, T. (1992). *Geometric Aspects of Visual Object Recognition.* PhD thesis, MIT Department of Brain and Cognitive Sciences.

Brooks, R. (1983). Model-Based Three-Dimensional Interpretations of Two-Dimensional Images. *IEEE Transactions PAMI*, PAMI-5(2):140 – 150.

Burns, J. (1992). *Matching 2D Images to Multiple 3D Objects Using View Description Networks.* PhD thesis, University of Massachusetts at Amherst, Dept. of Computer and Information Science.

Burns, J. and Riseman, E. (1992). Matching Complex Images to Multiple 3D Objects Using View Description Networks. In *Proceedings: Image Understanding Workshop*, pages 675–682. Morgan Kaufmann.

Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Transactions PAMI*, PAMI-8(6):679–698.

Cass, T. (1992a). *Polynomial-Time Geometric Matching for Object Recognition*. PhD thesis, MIT Department Electrical Engineering and Computer Science.

Cass, T. (1992b). Polynomial-Time Object Recognition in the Presence of Clutter, Occlusion, and Uncertainty. In Sandini, G., editor, *Computer Vision – ECCV '92*, pages 834–851. Springer Verlag.

Chin, R. and Dyer, C. (1986). Model-Based Recognition in Robot Vision. *Computing Surveys*, 18:67–108.

Clemens, D. and Jacobs, D. (1990). Model Group Indexing for Recognition. In *Symposium on Advances in Intelligent Systems*. SPIE.

Clemens, D. and Jacobs, D. (1991). Space and Time Bounds on Indexing 3-D Models from 2-D Images. *IEEE Transactions PAMI*, 13(10):1007–1017.

Cohen, F. and Cooper, D. (1988). A Decision Theoretic Approach for 3-D Vision. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition*, pages 964–972. IEEE.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Roy. Statist. Soc.*, 39:1 – 38.

Duda, R. and Hart, P. (1973). *Pattern Classification and Scene Analysis*. John Wiley and Sons.

Dykstra, P. and Muuss, M. (1987). The BRL CAD Package: an Overview. In *Proceedings of the Fourth USENIX Computer Graphics Workshop*, pages 73–80, Cambridge MA.

Edelman, S. and Poggio, T. (1990). Bringing the Grandmother Back Into the Picture: a Memory-Based View of Object Recognition. A.I. Memo 1181, Massachusetts Institute of Technology.

Fischler, M. and Elschlager, R. (1973). The Representation and Matching of Pictorial Structures. *IEEE Transactions on Computers*, C-22(1):67–92.

Fua, P. and Hanson, A. (1989a). Objective Functions for Feature Discrimination: Applications to Semiautomated and Automated Feature Extraction. In *Proceedings: Image Understanding Workshop*, pages 676–694. Morgan Kaufmann.

Fua, P. and Hanson, A. (1989b). Objective Functions for Feature Discrimination: Theory. In *Proceedings: Image Understanding Workshop*, pages 443–459. Morgan Kaufmann.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions PAMI*, PAMI-6(6):721–741.

Goad, C. (1986). Fast 3-D Model Based Vision. In Pentland, A. P., editor, *From Pixels to Predicates*, pages 371–391. Ablex Publishing Co.

Green, Jr., T. (1992). *Three-Dimensional Object Recognition Using Laser Radar*. PhD thesis, MIT Department Electrical Engineering and Computer Science.

Green, Jr., T. and Shapiro, J. (1992). Maximum-Likelihood Laser Radar Range Profiling with the Expectation – Maximization Algorithm. *Opt. Eng.*

Grimson, W. (1985). Computational Experiments with a Feature Based Stereo Algorithm. *IEEE Transactions PAMI*, PAMI-7(1):17–34.

Grimson, W. (1990). *Object Recognition by Computer: The Role of Geometric Constraints*. MIT Press.

Grimson, W. and Huttenlocher, D. (1990). On the Verification of Hypothesized Matches in Model-Based Recognition. In *First Europ. Conf. Comp. Vision*, pages 489–498.

Grimson, W. and Lozano-Pérez, T. (1987). Localizing Overlapping Parts by Searching the Interpretation Tree. *IEEE Transactions PAMI*, PAMI-9(4):469–482.

Haralick, R., H.Joo, C.N.Lee, X.Zhuang, V.G.Vaidya, and M.B.Kim (1989). Pose Estimation from Corresponding Point Data. *IEEE Trans. on Systems Man and Cybernetics*, 19(6):1426 – 1445.

Hough, P. (1962). Methods and Means for Recognizing Complex Patterns. U.S. Patent 3069654.

Huttenlocher, D., Kedem, K., Sharir, K., and Sharir, M. (1991). The Upper Envelope of Voronoi Surfaces and its Applications. In *Proceedings of the Seventh ACM Symposium on Computational Geometry*, pages 194–293.

Huttenlocher, D. and Ullman, S. (1988). Recognizing Solid Objects by Alignment. In *Proceedings: Image Understanding Workshop*, pages 1114–1124. Morgan Kaufmann.

Illingworth, J. and Kittler, J. (1988). A Survey of the Hough Transform. *Computer Vision, Graphics, and Image Processing*, 44:87–116.

Jacobs, D. (1992). *Recognizing 3D Objects Using 2D Images*. PhD thesis, MIT Department Electrical Engineering and Computer Science.

Jaynes, E. T. (1982). Where Do We Go From Here? In Smith, C. R. and Grandy, Jr., W. T., editors, *Maximum-Entropy and Bayesian Methods in Inverse Problems*, pages 21 – 58. MIT Press.

Jiang, H., Robb, R., and Holton, K. (1992). A New Approach to 3-D Registration of Multimodality Medical Images by Surface Matching. In *Visualization in Biomedical Computing*, pages 196–213. SPIE.

Kumar, R. and Hanson, A. (1989). Robust Estimation of Camera Location and Orientation from Noisy Data Having Outliers. In *Proc. of the Workshop on Interpretation of 3D Scenes*, pages 52–60. IEEE Computer Society.

Lamdan, Y. and Wolfson, H. (1988). Geometric Hashing: A General and Efficient Model-Based Recognition Scheme. In *Second Int. Conf. Comp. Vision*.

Levitt, T., Binford, T., and Mann, W. (1988). Bayesian Inference in Model-Based Machine Vision. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE.

Lipson, P. (1992). Model Guided Correspondence. Master's thesis, MIT Department Electrical Engineering and Computer Science.

Lowe, D. (1985). *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers.

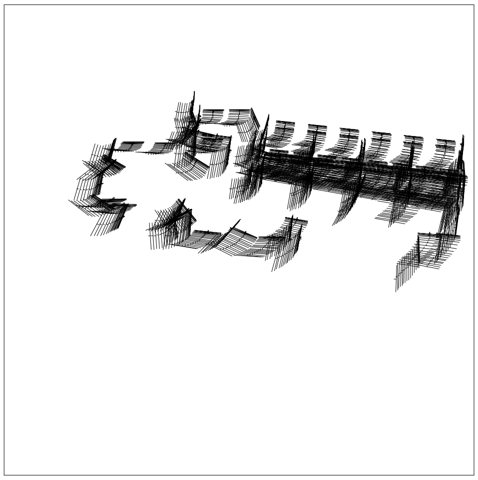

Fig. 22. EM Iteration Poses

Marr, D. (1982). *Vision*. Freeman.

Marroquin, J. (1985). *Probabilistic Solution of Inverse Problems*. PhD thesis, MIT Department Electrical Engineering and Computer Science.

Marroquin, J., Mitter, S., and Poggio, T. (1987). Probabilistic Solution of Ill-posed Problems in Computational Vision. *Journal of the Am. Stat. Assoc.*, 82(397):76–89.

Menon, M. and Wells, W. (1990). Massively Parallel Image Restoration. In *Proceedings of the International Joint Conference on Neural Networks*, San Diego, CA. IEEE.

Perrett, D. et al. (1985). Visual Cells in the Temporal Cortex Sensitive to Face View and Gaze Control. *Proc. Roy. Soc. London B*, 223:293 – 317.

Ponce, J. and Kriegman, D. (1989). On Recognizing and Positioning Curved 3D Objects From Image Contours. In *Image Understanding Workshop (Palo Alto, CA, May 23-26, 1989)*, pages 461–470.

Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. (1986). *Numerical Recipes: the Art of Scientific Computing*. Cambridge University Press.

Rivlin, E. and Basri, R. (1992). Localization and Positioning Using Combinations of Model Views. Technical Report CAR-TR-631, Center for Automation Research, University of Maryland.

Shapiro, J., Reinhold, R., and Park, D. (1986). Performance Analyses for Peak-Detecting Laser Radars. *Proceedings of the SPIE*, 663:38–56.

Shapiro, S., editor (1987). *Encyclopedia of Artificial Intelligence*. John Wiley & Sons.

Shashua, A. and Ullman, S. (1988). Structural Saliency: The Detection of Globally Salient Structures Using a Locally Connected Network. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 321–327.

Stephens, R. (1990). A Probabilistic Approach to the Hough Transform. In *Proceedings of the British Machine Vision Conference*, pages 55–60, Univ. of Oxford.

Thompson, D. (1989). Edge Based Transform Refinement. In *Proceedings: Image Understanding Workshop*, pages 1070–1075. Morgan Kaufmann.

Thompson, D. and Mundy, J. (1987). Three Dimensional Model Matching From an Unconstrained Viewpoint. In *Proceedings IEEE International Conference on Robotics and Automation*, pages 208 – 219. IEEE.

Ullman, S. and Basri, R. (1989). Recognition by Linear Combinations of Models. A.I. Memo 1152, Massachusetts Institute of Technology.

Wells, W. (1990). A Statistical Approach to Model Matching. In *Symposium on Advances in Intelligent Systems*. SPIE.

Wells, W. (1991). MAP Model Matching. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition*, pages 486–492, Lahaina, Maui, Hawaii. IEEE.

Wells, W. (1992). *Statistical Object Recognition*. PhD thesis, MIT Department Electrical Engineering and Computer Science, Cambridge, Mass. MIT AI Laboratory TR 1398.

Wells, W. (1993). Statistical Object Recognition with the Expectation-Maximization Algorithm in Range-Derived Features. In *Proceedings: Image Understanding Workshop*. Morgan Kaufmann.

Wu, C. (1983). On the Convergence Properties of the EM Algorighm. *The Annals of Statistics*, 11(1):95 – 103.

Yuille, A. (1990). Generalized Deformable Models, Statistical Physics, and Matching Problems. *Neural Computation*, 2:1 – 24.

Yuille, A., Geiger, D., and Bülthoff, H. (1990). Stereo Integration, Mean Field Theory and Psychophysics. In *Computer Vision – ECCV 90*, pages 73–82. Springer Verlag.

William M. Wells III is an instructor in radiology in the Department of Radiology, Harvard Medical School and Brigham and Women's Hospital. He is also a research scientist at the MIT Artificial Intelligence Laboratory.

Dr. Wells received the B.S. degree in physics from the University of California at Santa Cruz in 1976, and the M.S. in computer science (computer engineering) from Stanford University in 1984. In 1984 and 1985 he was employed by Silma Incorporated and from 1985 to 1988 he participated in mobile robot research at SRI International.

He completed his PhD at the MIT AI Laboratory in model based computer vision under the supervision of Professor Eric Grimson, spending several summers at MIT Lincoln Laboratory.