Statistical Object Recognition with the Expectation-Maximization Algorithm in Range-Derived Features

William M. Wells III *†
Massachusetts Institute of Technology
Artificial Intelligence Laboratory
545 Technology Square, Cambridge, Massachusetts 02139
sw@ai.mit.edu

October 13, 1995

Abstract

An extension to the alignment approach is proposed that includes a pose refinement step before verification. In the alignment approach the pose estimates of the initial hypotheses tend to be somewhat inaccurate, since they are based on minimal sets of corresponding features. A method is described that refines the pose estimate while simultaneously identifying and incorporating the constraints of all supporting image features. The strategy also makes practical initial alignments based on low resolution features – which, being less numerous, allow faster running times.

Two statistical formulations of model-based recognition are described: MAP Model Matching, and Posterior Marginal Pose Estimation (PMPE). These formulations use a normal model for feature fluctuations. Empirical evidence is provided from the domain of video edge features indicating that normal probability densities are good models of feature fluctuations – better than uniform densities in that domain. The evidence is provided in the form of observed and fitted cumulative distributions. The results of some statistical tests are reported.

The Expectation – Maximization (EM) algorithm is discussed as a method of carrying out local searches in pose space of the PMPE objective function. A recognition experiment is described where the method is used with features derived from synthetic range images.

*The author is currently affiliated with Harvard Medical School, Brigham and Women's Hospital, Department of Radiology, Surgical Planning Laboratory.

†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.

1 Introduction

1.1 The Statistical Approach and Object Recognition

In this paper, visual object recognition is 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, are 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 are often useful when 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 [1] is one example, while the work of Geman and Geman on image restoration [2] is another. 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 the specification of the problem from the description of the algorithms used to attack 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 [3], Blake and Zisserman [4] in image restoration and Beveridge, Weiss and Riseman [5] in line segment based model matching. With this approach, plausible forms for components of the objective function are often combined using trade-off parameters. Such trade-off parameters are determined empirically. An advantage of deriving objective functions from statistical theories is that the assumptions become explicit in that the forms of the objective function components are clearly related to specific probabilistic models. A second advantage is that the trade-off parameters in the objective function can be derived from measurable statistics of the domain.

1.2 Alignment

The basic strategy of the alignment method of recognition [6] is to use separate mechanisms for generating and testing hypotheses.

Recently, indexing methods have become available for efficiently generating hypotheses in recognition. These methods avoid a significant amount of search by looking up, in pre-computed tables, the object features that might correspond to a group of image features. The geometric hashing method of Lamdan and Wolfson [7] uses invariant properties of small groups of features under affine transformations as the lookup key. Clemens and Jacobs [8] describe an indexing method that gains efficiency by using a feature grouping process to select small sets of image features that are likely to belong to the same object in the scene.

1.3 Align – Refine – Verify

The recognition strategy advocated in this work 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. With the alignment method, 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 verifications) are likely to result from using all supporting image feature data, rather than a small subset. This paper describes a method that refines the pose estimate while simultaneously identifying and incorporating the constraints of all supporting image features.

The strategy also makes practical initial alignments based on lower resolution features. Using low resolution features in indexing is attractive, because there are fewer features at low resolution, allowing faster running times.

1.4 Structure of the Paper

In Sections 2 and 3 probabilistic models of feature correspondences and feature fluctuations are described, while Section 4 discusses linear models of feature projection. Section 5 describes a statistical method of simultaneously evaluating correspondences and object pose. Building on this, Section 6 outlines a statistical method of object pose evaluation: Posterior Marginal Pose Estimation (PMPE). Section 7 describes the use of the expectation – maximization (EM) algorithm for solving the PMPE objective function. An experiment using the EM algorithm with features derived from synthetic range images is presented in Section 8, and related work is discussed in Section 9.

2 Modeling Feature Correspondence

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

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

Image features are discussed in more detail in Sections 3 and 4.

The object to be recognized is also described by a set of features, these are represented by real matrices:

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

Additional details on object features appears in Section 4.

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

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

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 [9]. 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, $\perp$.

2.1 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. This model has been designed to be a reasonable compromise between simplicity and accuracy.

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

$$p(\Gamma) = \prod_i p(\Gamma_i) \quad . \tag{1}$$

There is evidence against using statistical independence here, for example, occlusion is locally correlated. Independence is used as an engineering approximation that simplifies the resulting formulations of recognition. It may be justified by the good performance of the recognition experiments described in Section 8. Few recognition systems have used non-independent models of correspondence. Breuel described one approach [10]. In [11] the independence assumption of Equation 1 has been relaxed in a Markov Random Field (MRF) model of correspondences that is meant to capture the correlated aspects of occlusion.

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

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

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. B may be estimated by taking simple statistics on images from the domain.

The independent correspondence model is used in the experiments described in this paper.

3 Modeling Image Features

3.1 Uniform Model for Background Features

Image features belonging to the background are assumed to be uniformly distributed throughout the v-dimensional coordinate space of the image,

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

(The PDF is zero outside the coordinate space of the image.) The position and orientation, or *pose* of the object is described by β. The W_i are the dimensions of the feature coordinate space.

Equation 3 describes the maximum entropy PDF consistent with the constraint that the coordinates of image features are always expected to lie within the space of the image.

3.2 Normal Model for Matched Features

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

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

$\mathrm{G}_{\psi_{ij}}$ is the v-dimensional Gaussian probability density function with covariance matrix ψ_{ij}: $\mathrm{G}_{\psi_{ij}}(x) = (2\pi)^{-\frac{v}{2}} |\psi_{ij}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} x^T \psi_{ij}^{-1} x\right)$. The covariance matrix ψ_{ij} is discussed more fully in Section 3.3.

When $\Gamma_i = M_j$, 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. Projection and pose are discussed in more detail in Section 4.

3.2.1 Empirical Evidence for Normal Model of Feature Fluctuations

This section describes some empirical evidence from the domain of video image edge features indicating that, in that domain, normal probability densities are good models of feature fluctuations, and that they are better than uniform probability densities. The evidence is provided in the form of observed and fitted cumulative distributions. Additionally, the results of some statistical tests are reported.

The data that is analyzed are the perpendicular deviations of edge features derived from video images. The features are shown in Figure 1.

The model features are from mean edge images (these are averaged with respect to illumination), and the edge operator used in obtaining the image features is ridges in the magnitude of the image gradient. These edge detection methods are described in [11]. The

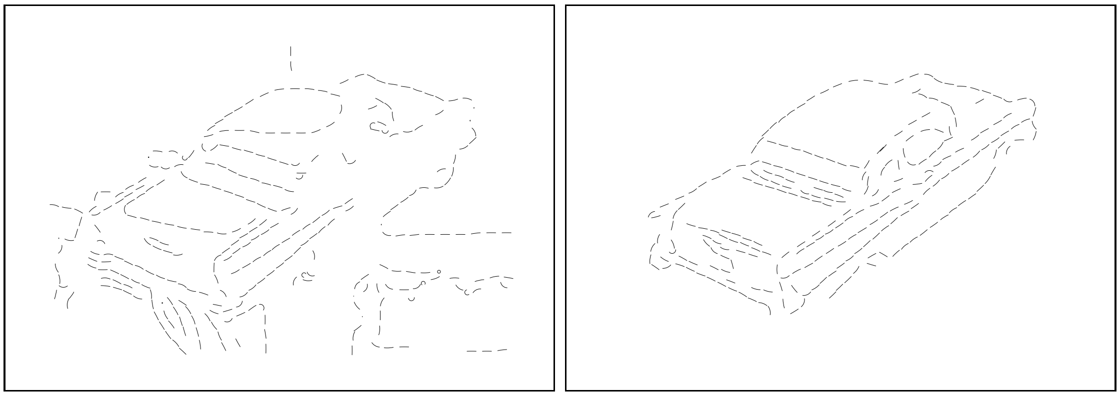

Figure 1: Image and Model Features

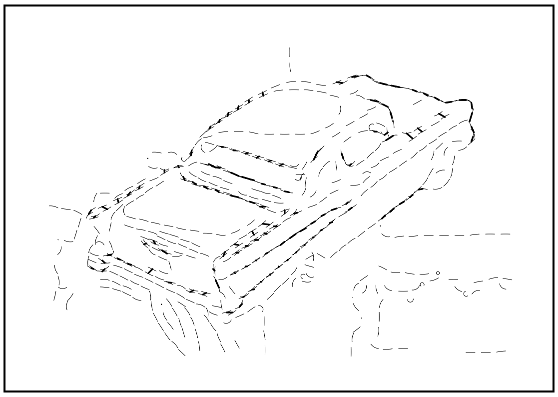

Figure 2: Feature Correspondences

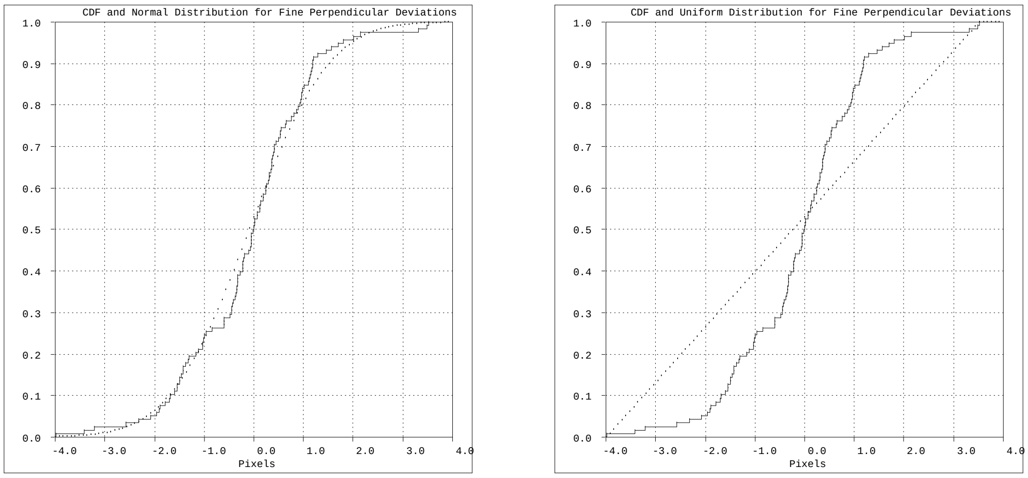

Figure 3: Observed Cumulative Distribution with Fitted Normal and Uniform Cumulative Distributions

smoothing standard deviation used in the edge detection was 1.93 pixels. The object is at the same pose in the model and image scenes, the variation in features is due to changes in lighting, the presence of background features in the "image" scene and the vagaries of edge detection.

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. The features are displayed as circular arc fragments for clarity. The edge curves were broken arbitrarily into 10 pixel fragments.

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. Perpendicular and parallel deviations of the corresponding features were calculated with respect to the normal vectors to edge curves at the image features.

Figure 3 shows the cumulative distribution of the perpendicular deviations. The cumulative distribution of a fitted normal density is plotted in the left figure as heavy dots over the observed distribution. The distribution was fitted to the data using the maximum-likelihood method. These figures show good agreement between the observed distribution, and the fitted normal distribution. The observed cumulative distribution is shown again on the right in Figure 3, this time with the cumulative distribution of a fitted uniform density over-plotted in heavy dots. As before, the uniform density was fitted to the data using the maximum-likelihood method. This figure show relatively poor agreement between the observed and fitted distribution, in comparison to the normal density. Similar results obtained for the parallel deviations, and for a similar set of coarse features [11].

Kolmogorov-Smirnov tests were carried out [11] to evaluate the compatibility of the data with the fitted normal and uniform distributions. 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.

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 on j allows dependence on the identity of the model feature. This is useful when some model features are know 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. In its strict form this assumption may be too limiting. This section outlines a compromise approach, oriented stationary statistics, that was used in the experiments described in Section 8.

The Oriented Stationary Statistics method involves attaching 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. When an image is presented for recognition, the constant feature covariance is specialized by rotating it to orient it with respect to each image feature. Oriented Stationary Statistics is described in more detail in [11].

4 Linear Projection Models

Pose determination is frequently framed as an optimization problem. The pose determination problem may be significantly simplified if the feature projection model is linear in the parameters of the transformation (the pose vector). The system described in this paper uses a projection model having this property, this enables solving the embedded optimization problem using least squares. Linear projection models may be written in the following form:

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

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

Several linear projection models were described in [12], and in [11].

4.1 2-D Oriented-Range Feature Model

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, is defined by

$$\eta_i = \begin{bmatrix} p'_{ix} \\ p'_{iy} \\ c'_{ix} \\ c'_{iy} \end{bmatrix} \qquad M_i = \begin{bmatrix} p_{ix} & -p_{iy} & 1 & 0 \\ p_{iy} & p_{ix} & 0 & 1 \\ c_{ix} & -c_{iy} & 0 & 0 \\ c_{iy} & c_{ix} & 0 & 0 \end{bmatrix} \qquad \text{and} \qquad \beta = \begin{bmatrix} \mu \\ \nu \\ t_x \\ t_y \end{bmatrix} \quad .$$

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 a vector who's direction is perpendicular to the range discontinuity and pointing 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, where

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

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.

This projection model was inspired by a method used by Faugeras and Ayache in their vision system HYPER [13], and it is used in the experiments described in Section 8.

5 MAP Model Matching

This section outlines MAP Model Matching [1] , a means of evaluating joint hypotheses of match and pose using the MAP criterion.

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*). MAP estimators are especially practical when used with normal probability densities.

The parameters to be estimated in matching are the correspondences between image and object features, and the pose of the object in the image.

[1] An early version of this work appeared in [12] and [14].

The probabilistic models of image features, conditioned on match and pose, that is described in Equations 3 and 4 may be written as follows:

$$p(Y_i \mid \Gamma, \beta) = \begin{cases} \frac{1}{W_1 W_2 \cdots W_v} & \text{if } \Gamma_i = \perp \\ \mathrm{G}_{\psi_{ij}}(Y_i - \mathcal{P}(M_j, \beta)) & \text{if } \Gamma_i = M_j \end{cases} \quad . \tag{6}$$

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 space of the image (the width of the image 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 could be a constant – an assumption that the feature statistics are stationary in the image, or ψ may depend only on i, the image feature index. This 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

$$p(Y \mid \Gamma, \beta) = \prod_i p(Y_i \mid \Gamma, \beta) = \prod_{\Gamma_i = \perp} \frac{1}{W_1 W_2 \cdots W_v} \prod_{ij : \Gamma_i = M_j} \mathrm{G}_{\psi_{ij}}(Y_i - \mathcal{P}(M_j, \beta)) \quad . \tag{7}$$

Next, a joint prior on correspondences and pose is constructed. Prior information on the pose is assumed to be supplied as a normal density,

$$p(\beta) = \mathrm{G}_{\psi_\beta}(\beta - \beta_0) = (2\pi)^{\frac{-z}{2}} |\psi_\beta|^{-\frac{1}{2}} \exp(-\frac{1}{2} x^T \psi_\beta^{-1} x) \quad . \tag{8}$$

Here ψ_β is the covariance matrix of the pose prior and z is the dimensionality of the pose vector, β. With this choice for the form of the pose prior the system is closed in the sense that the resulting pose estimate will also be normal. This is convenient for coarse-fine. If little is known about the pose a-priori, the prior may be made quite broad. This is expected to be often the case. 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), and using the independent correspondence model of Equations 1 and 2, a mixed joint prior probability function may be written as follows,

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

The variable B has been generalized here to have an image feature dependent subscript, i. In the experiments reported in Section 8, $B_i = B$ for all i. This generalization is explored more fully in [11]. 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:

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

where $p(Y)$ is the probability of the image being analyzed – a constant in terms of the parameters being estimated.

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

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

The search for maximizing joint hypotheses of match and pose may be ordered as either a search in correspondence space, or a search in pose space. Recognition experiments in correspondence based search were reported in [12], where heuristic beam search was used. In [11] it is shown that searching a specialization of this MAP criterion in pose space is equivalent to robust chamfer matching.

6 Posterior Marginal Pose Estimation

In the previous section the object recognition problem was posed as an optimization problem resulting from a statistical theory. In that formulation, a hypothesis was the position and orientation, or pose, of the object – as well as the correspondences between object and image features.

The formulation of recognition that is described in this section, Posterior Marginal Pose Estimation (PMPE) [2] , builds on MAP model matching. It provides a smooth objective function for evaluating hypotheses that are simply the pose of the object. The pose is the most important aspect of the problem, in the sense that knowing the pose enables grasping or other interaction with the object.

PMPE 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 in the objective function. Additional motivation was provided by the work of Yuille, Geiger and Bülthoff on stereo [1]. They discussed computing disparities in a statistical theory of stereo where a marginal is computed over matches.

Here we use the same strategy as MAP Model Matching for evaluating hypotheses now consisting only of pose – 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, by formally taking the marginal over possible matches:

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

In Section 5, 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

[2] An early version of this work appeared in [15].

posterior marginal may be written as

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

Using Equations 1 (the independent feature model) and 7, we may express the posterior marginal of β in terms of the component densities:

$$p(\beta \mid Y) = \frac{p(\beta)}{p(Y)} \sum_{\Gamma_1} \sum_{\Gamma_2} \cdots \sum_{\Gamma_n} \prod_i [p(Y_i \mid \Gamma_i, \beta) p(\Gamma_i)] \ .$$

Breaking one factor out of the product gives

$$p(\beta \mid Y) = \frac{p(\beta)}{p(Y)} \sum_{\Gamma_1} \sum_{\Gamma_2} \cdots \sum_{\Gamma_{n-1}} \prod_{i=1}^{n-1} [\, p(Y_i \mid \Gamma_i, \beta) p(\Gamma_i)] \sum_{\Gamma_n} p(Y_n \mid \Gamma_n, \beta) p(\Gamma_n) \ .$$

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

$$p(\beta \mid Y) = \frac{p(\beta)}{p(Y)} \prod_i p(Y_i \mid \beta) \ , \tag{10}$$

since

$$p(Y_i \mid \beta) = \sum_{\Gamma_i} p(Y_i \mid \Gamma_i, \beta) p(\Gamma_i) \ . \tag{11}$$

Splitting the Γ_i sum into its cases,

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

Substituting the densities assumed in the model of Section 5 in Equations 6 and 2 yields

$$p(Y_i \mid \beta) = \frac{1}{W_1 \cdots W_v} B_i + \sum_{M_j} \mathrm{G}_{\psi_{ij}}(Y_i - \mathcal{P}(M_j, \beta)) \frac{1 - B_i}{m} \ . \tag{12}$$

Installing this into Equation 10 leads to

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

The objective function for posterior marginal pose estimation may be defined as the scaled logarithm of the posterior marginal probability of the pose, as follows,

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

where K is a constant that has the following definition:

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

K has been chosen to simplify the form of the objective function. This leads to the following expression for the objective function (use of a normal pose prior is assumed, as in Equation 8)

$$L(\beta) = -\frac{1}{2}(\beta - \beta_o)^T \psi_\beta^{-1} (\beta - \beta_0) + \sum_i \ln \left[1 + \sum_{M_j} \frac{W_1 \cdots W_v}{m} \frac{1 - B_i}{B_i} \mathrm{G}_{\psi_{ij}} (Y_i - \mathcal{P}(M_j, \beta)) \right] \quad . \tag{13}$$

This objective function for evaluating pose hypotheses is a smooth function of the pose. Methods of continuous optimization may be used to search for local maxima, although starting values are an issue.

The first term in the PMPE objective function (Equation 13) is due to the pose prior. It is a quadratic penalty for deviations from the nominal pose. The second term essentially 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. The logarithmic term will be near zero if there are no model features close to the image feature in question.

When the MRF correspondence model mentioned in Section 2.1 is used in PMPE, a simple closed form expression for the estimator no longer obtains. However, a tractable dynamic programming method for evaluating the resulting PMPE objective function at a specific pose is described in [11].

7 Expectation – Maximization Algorithm

The Expectation – Maximization (EM) algorithm was introduced in its general form by Dempster, Rubin and Laird in 1978 [16]. 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 8. A linear model is assumed for feature projection.

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:

$$\mathbf{0} = \nabla_\beta \ln p(\hat{\beta} \mid Y) \ . \tag{14}$$

Imposing the condition of Equation 14 on the posterior probability of the pose of an object, given an image, of Equation 10 yields the following,

$$\mathbf{0} = \frac{\nabla_\beta p(\hat{\beta})}{p(\hat{\beta})} + \sum_i \frac{\nabla_\beta p(Y_i \mid \hat{\beta})}{p(Y_i \mid \hat{\beta})} \ . \tag{15}$$

Using the linear projection model, Equation 12 may be re-written as

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

The zero gradient condition of Equation 15 may now be expressed as follows,

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

Using the normal pose prior, and differentiating the normal densities yields

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

.

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

$$\mathbf{0} = \psi_\beta^{-1}(\hat{\beta} - \beta_0) + \sum_{ij} W_{ij} M_j^T \psi_{ij}^{-1}(Y_i - M_j \hat{\beta}) \ , \tag{16}$$

with the following definition:

$$W_{ij} = \frac{\mathrm{G}_{\psi_{ij}}(Y_i - M_j \hat{\beta})}{\frac{B_i}{1-B_i} \frac{m}{W_1 W_2 \cdots W_v} + \sum_j \mathrm{G}_{\psi_{ij}}(Y_i - M_j \hat{\beta})} \ . \tag{17}$$

Equation 16 has the appearance of being a linear equation for the pose estimate $\hat{\beta}$ that satisfies the zero gradient condition for being a maximum. 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 16, the EM algorithm iterates the following two steps:

- Treating the weights, W_{ij} as constants, solve Equation 16 as a linear equation for a new pose estimate $\hat{\beta}$. This is referred to as the M step.

- Using the most recent pose estimate $\hat{\beta}$, re-evaluate the weights, W_{ij}, according to Equation 17. This is referred to as the E step.

The M step is so named because, in the exposition of the algorithm in [16], it corresponds to a maximum likelihood estimate. As discussed there, the algorithm is also amenable to use in MAP formulations (like this one). Here the M step corresponds to a MAP estimate of the pose, given that the current estimate of the weights is correct.

The E step is so named because calculating the W_{ij} corresponds to taking the expectation of some random variables, given the image data, and that the most recent pose estimate is correct. These random variables have value 1 if the i'th image feature corresponds to the j'th object feature, and 0 otherwise. Thus, after the iteration converges, the weights provide a continuous-valued estimate of the correspondences, that varies between 0 and 1.

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 1, the rest to 0. Indexing methods are one source of such hypotheses.

8 Range Feature Recognition Experiment

In this section, an experiment is described where the EM algorithm is used in a coarse – fine scheme to estimate the pose of a vehicle appearing in synthetic range images, without the need for feature correspondence information. The region of convergence of the coarse – fine algorithm is explored. The object has four degrees of freedom – translation, rotation, and scaling in the plane. Similar experiments, with full freedom of motion in 3D are described in [11].

8.1 Preparation of Features

The preparation of the features used in the experiment is summarized in Figure 4. The features were oriented-range features, as described in Section 4.1. 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 [17]. The ray tracer was modified to produce range images instead of shaded images. The synthetic range image appears in the first image of Figure 5.

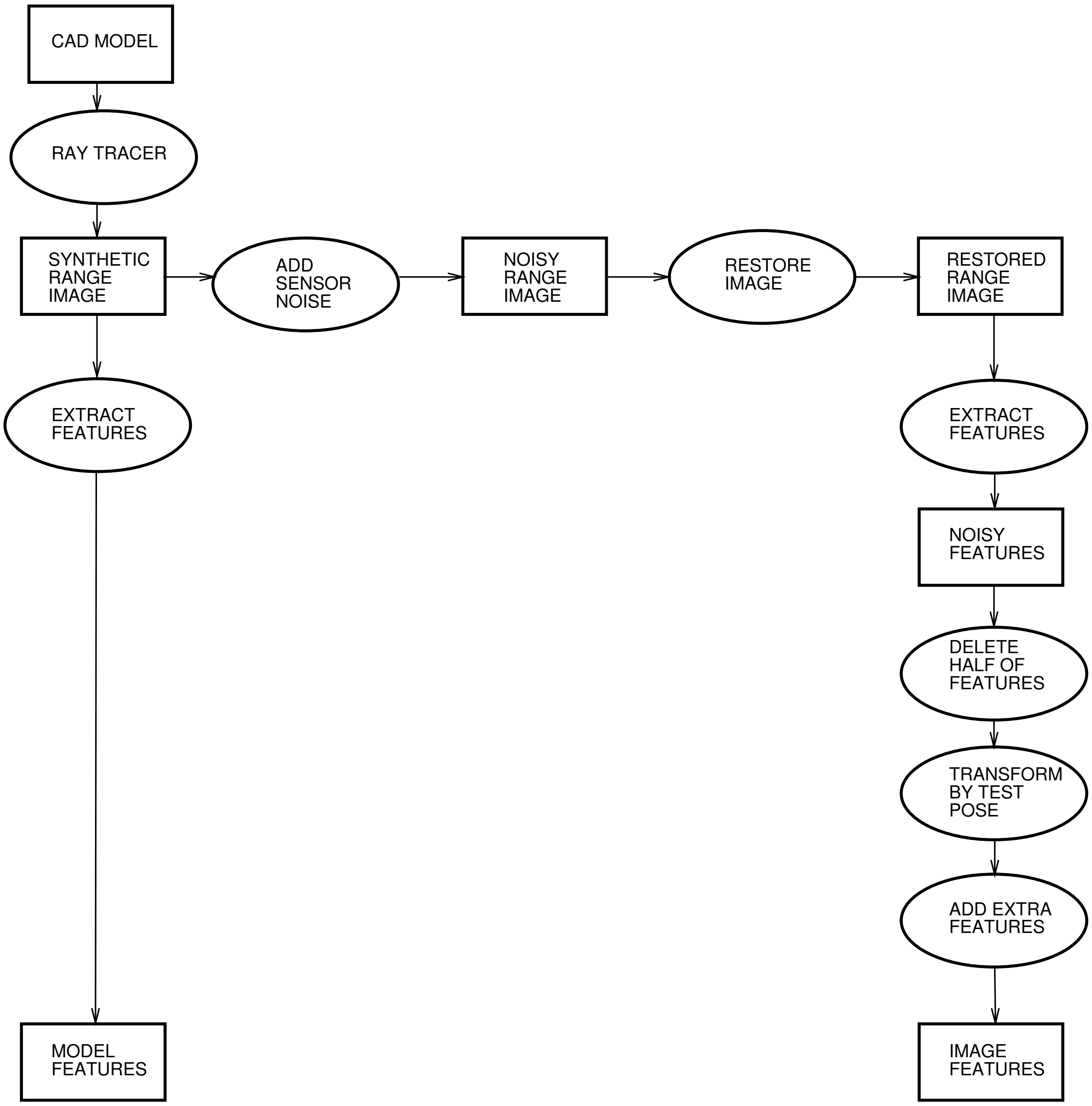

Figure 4: Preparation of Features

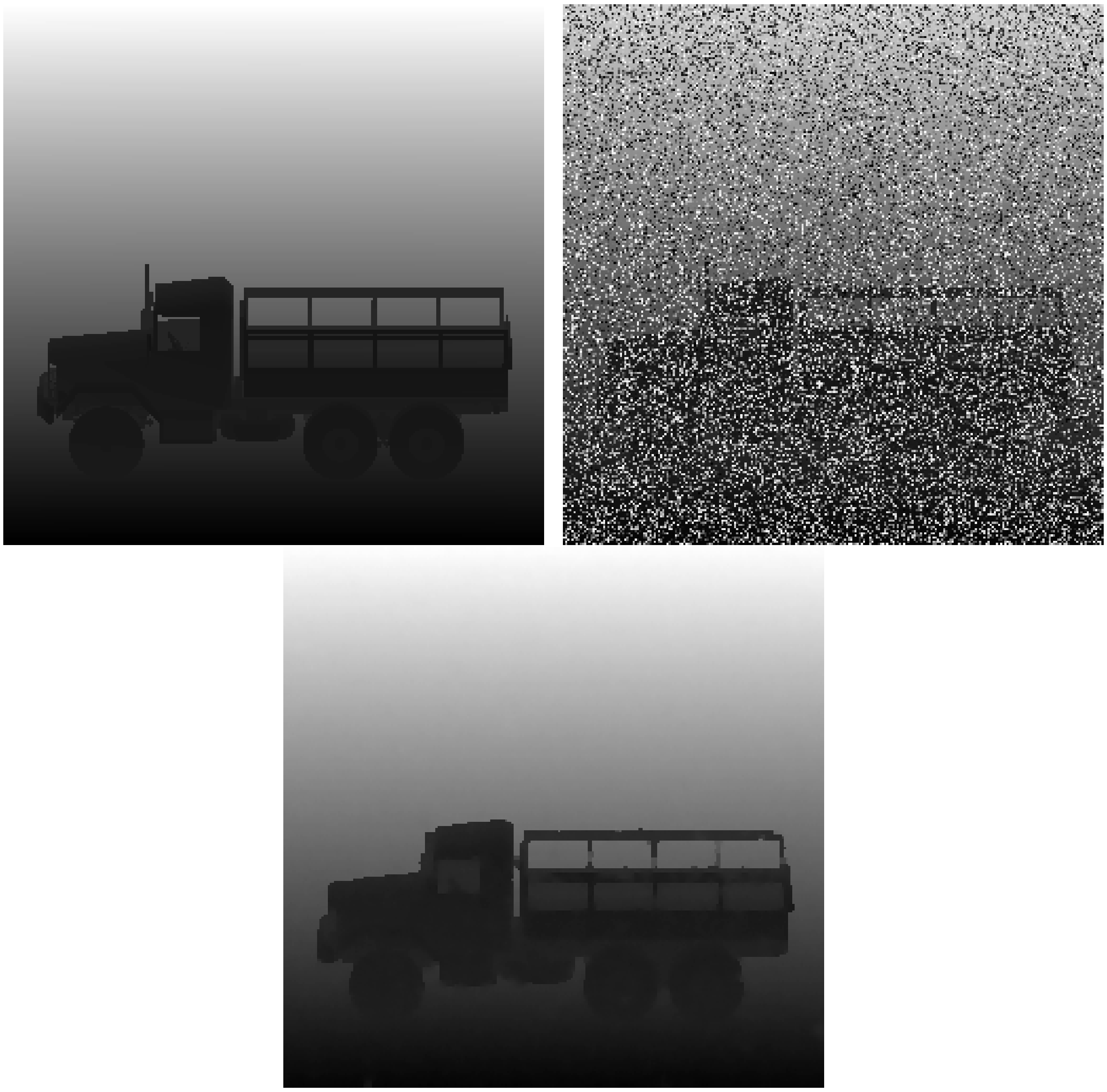

Figure 5: Synthetic Range Image, Noisy Range Image, and Restored Range Image

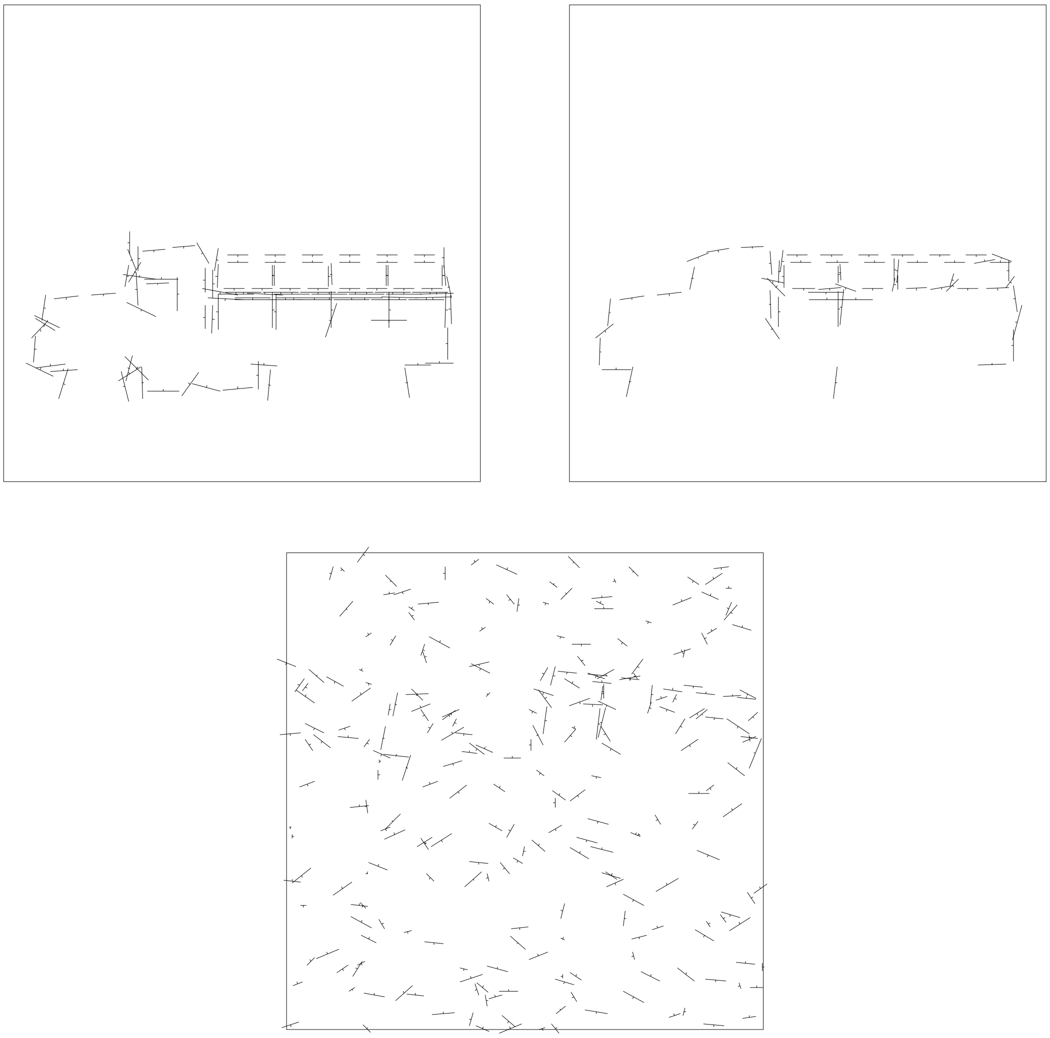

Figure 6: Model Features, Noisy Features, and Image Features

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 [18]. 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 as the second image in Figure 5. To simulate post sensor processing, the corrupted image was "restored" via a statistical restoration method of Menon and Wells [19]. The restored range image appears as the third image of Figure 5.

Oriented range features, as described in Section 4.1, were extracted from the synthetic range image, for use as model features – and from the restored range image, these are called the noisy features. 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 4.1. The model features are shown in the first image of Figure 6. 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.

The noisy features, derived from the restored range image, appear in the second image of Figure 6. There are 62 noisy features. Some features have been lost due to the corruption and restoration of the range image. The set of image features was prepared from the noisy 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. The 248 image features appear in the third image of Figure 6.

8.2 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:

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

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,

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

and

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

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

8.3 Results

Figure 7 illustrates the approximate region of convergence of the coarse-fine search with the EM algorithm described above, when used with the image and model features described in Section 8.1. Eight poses are displayed with the truck shown with light lines – the true pose of the truck is shown for reference with heavy lines. These eight poses are displacements from the true pose in both directions along the four coordinate axes of the pose space. They represent, approximately, the region of convergence of the coarse-fine method – the algorithm converges from these poses, but not from much farther away from the true pose.

An image displaying a sequence of poses from an EM iteration at the coarsest scale appears in Figure 8. The algorithm converged in 21 iterations.

9 Related Work

Green [20] and Shapiro and Green [21] describe a theory of maximum-likelihood laser radar range profiling. The research focuses on 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.

Cass [22] 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 continuous search methods. The search is based on a geometric formulation of the constraints on feasible transformations.

There are some connections between PMPE and standard methods of robust pose estimation, like that described by Haralick [23]. Both can provide robust estimates of the pose of an object, based on an observed image. The main difference is that the standard methods

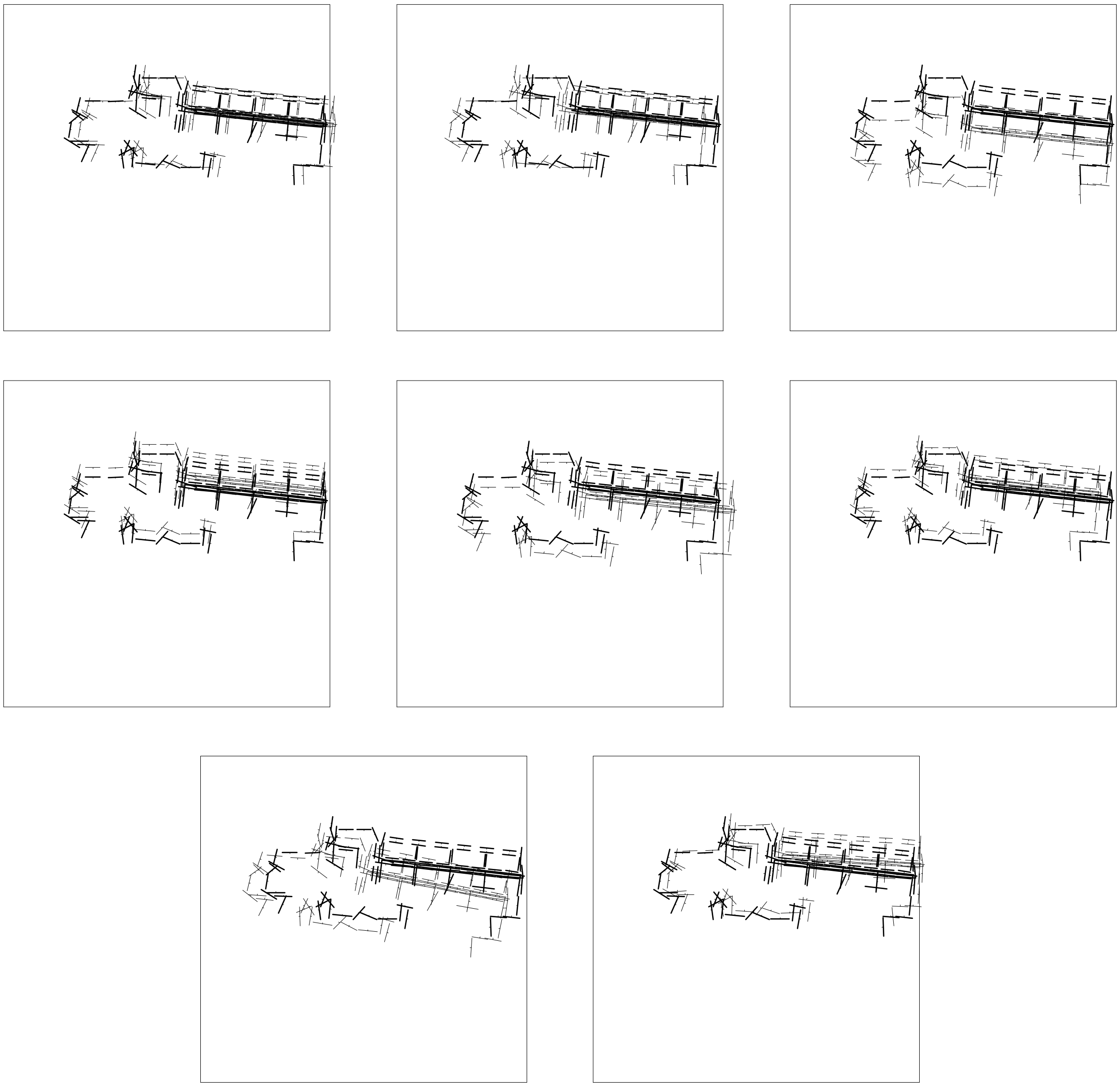

Figure 7: Eight Convergent Starting Poses

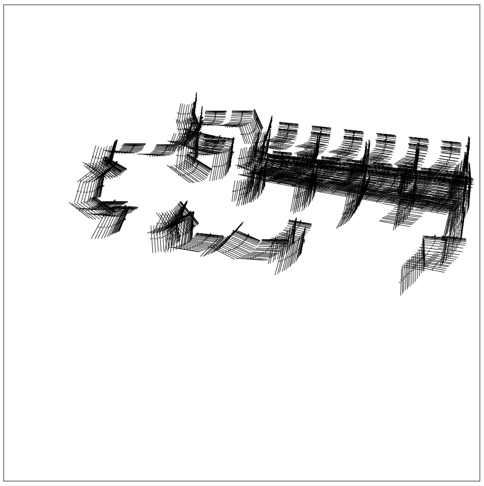

Figure 8: EM Iteration Poses

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).

As mentioned above, Yuille, Geiger and Bülthoff [1] discussed computing disparities in a statistical theory of stereo where a marginal is computed over matches. Yuille extends this technique [24] 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.

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 [25] 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 [26], pp. 220. Perhaps this is due to an unsuitable noise model implicit in the Hough Transform.

PMPE 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 [27]. Stephens proposes to call the posterior probability of parameters given image observations "The Probabilistic Hough Transform." He provided an example of estimating line parameters where image feature point probability densities were described as having uniform and normal components. He also states that the method has been used to track 3D objects, referring to his thesis [28] for definition of the method used.

Lipson [29] describes a system that does refinement of alignments under Linear Combination of Views. The method iterates solving linear systems. It differs by matching 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.

10 Acknowledgments

I thank my thesis advisor, W. E. L. Grimson, for his patience and guidance in the course of this research. I am grateful to Patrick Winston for making MIT's AI Lab the unique place that it is. During several summers this research was supported at MIT Lincoln Laboratory, group 53. Group leader Al Gschwendtner provided a stimulating environment for recognition research in laser radar imagery. There, I appreciate the help Steve Rak and Murali Menon offered me in preparing the images used in the experiments of Section 8.

References

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

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

[3] S.T. Barnard. Stereo Matching by Hierarchical, Microcanonical Annealing. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 832–835, August 1987.

[4] A. Blake and A. Zisserman. *Visual Reconstruction.* MIT Press, 1987.

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

[6] D.P. Huttenlocher and S. Ullman. Recognizing Solid Objects by Alignment. In *Proceedings: Image Understanding Workshop*, pages 1114–1124. Morgan Kaufmann, April 1988.

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

[8] D.T. Clemens and D.W. Jacobs. Model Group Indexing for Recognition. In *Symposium on Advances in Intelligent Systems*. SPIE, 1990.

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

[10] T.M. Breuel. *Geometric Aspects of Visual Object Recognition*. PhD thesis, MIT Department of Brain and Cognitive Sciences, 1992.

[11] W.M. Wells III. *Statistical Object Recognition*. PhD thesis, MIT Department Electrical Engineering and Computer Science, Cambridge, Mass., 1992. MIT AI Laboratory TR 1398.

[12] W.M. Wells III. MAP Model Matching. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition*, pages 486–492, Lahaina, Maui, Hawaii, June 1991. IEEE.

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

[14] W.M. Wells III. A Statistical Approach to Model Matching. In *Symposium on Advances in Intelligent Systems*. SPIE, 1990.

[15] W.M. Wells III. Posterior Marginal Pose Estimation. In *Proceedings: Image Understanding Workshop*, pages 745 – 751. Morgan Kaufmann, January 1992.

[16] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Roy. Statist. Soc.*, 39:1 – 38, 1977.

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

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

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

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

[21] T.J. Green, Jr. and J.H. Shapiro. Maximum-Likelihood Laser Radar Range Profiling with the Expectation – Maximization Algorithm. *Opt. Eng.*, November 1992.

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

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

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

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

[26] W.E.L. Grimson. *Object Recognition by Computer: The Role of Geometric Constraints.* MIT Press, 1990.

[27] R.S. Stephens. A Probabilistic Approach to the Hough Transform. In *Proceedings of the British Machine Vision Conference*, pages 55–60, Univ. of Oxford, September 1990.

[28] R.S. Stephens. *The Hough Transform: A Probabilistic Approach.* PhD thesis, Cambridge University, Department of Engineering, 1990.

[29] P. Lipson. Model Guided Correspondence. Master's thesis, MIT Department Electrical Engineering and Computer Science, 1992.