Feature Space Trajectory Methods for Active Computer Vision

Michael A. Sipe, *Member*, *IEEE*, and David Casasent, *Fellow*, *IEEE*

Abstract—We advance new active object recognition algorithms that classify rigid objects and estimate their pose from intensity images. Our algorithms automatically detect if the class or pose of an object is ambiguous in a given image, reposition the sensor as needed, and incorporate data from multiple object views in determining the final object class and pose estimate. A probabilistic feature space trajectory (FST) in a global eigenspace is used to represent 3D distorted views of an object and to estimate the class and pose of an input object. Confidence measures for the class and pose estimates, derived using the probabilistic FST object representation, determine when additional observations are required as well as where the sensor should be positioned to provide the most useful information. We demonstrate the ability to use FSTs constructed from images rendered from computer-aided design models to recognize real objects in real images and present test results for a set of metal machined parts.

Index Terms—Active vision, classification, object recognition, pose estimation.

1 Introduction

Object recognition involves processing sensor data in order to assign a *class* label (e.g., a part number) from among a limited number of valid possibilities and estimate the *pose* (i.e., position and orientation) of a three-dimensional object. We consider *active object recognition*, which implies the ability to systematically change sensor parameters to make the object recognition task easier. In our work, we have the capability of changing the viewpoint of the sensor (e.g., the sensor is mounted on a robot). When presented with an ambiguous view, our active object recognition system automatically recognizes the ambiguity and determines the sensor movement needed to obtain a new object view where the ambiguity may be resolved. Our algorithms differ from prior active object recognition work [1], [2], [3] in that they are based on matching global, not local, features.

Although our active object recognition algorithms are applicable to a wide class of problems, we focus on industrial automation. For automated assembly and inspection, it is generally necessary to determine the class of objects with very high reliability and to compute a pose estimate with moderate accuracy (within a few degrees). In an industrial setting, there is some degree of control over the environment in which our active object recognition system operates. Furthermore, it is common in industrial environments to mount sensors on assembly robots so that the sensor viewpoint may be readily changed. We consider images from simple and inexpensive 2D intensity visual sensors (CCD cameras) and assume that the position of the object is known to the extent that the visual sensor may be positioned to obtain an image containing a single, stationary object. We further assume that the scene lighting is reasonably well controlled and that objects of interest lie on a blank planar surface (such as a conveyor belt) designed to simplify the process of segmenting an object from the background.

Prior work in the field of model-based object recognition may be broadly divided into two fundamentally different approaches: local and global feature methods.

Local features are spatially localized (i.e., they are labeled by their coordinates) geometric object properties such as distinguished points, edges, corners, holes, distinctive curves, surface patches, or the axes of simple shapes such as ellipsoids [4]. These features are extracted from a computer-aided design (CAD) model or sensor data for a prototypical object and stored for each known object. The same local features are then extracted from data sensed from an input object and a search is initiated for the transformation (i.e., the pose of the object) which best matches the observed data to the model. The search for the correct interpretation of the sensed data can proceed either in the space of correspondences or in the space of object poses [4]. Either method entails a significant computational burden. Methods such as geometric hashing [5], structural indexing [6], and spin images [7] have been devised to speed up the matching process.

The advantages of local feature matching approaches include tolerance of a cluttered background, multiple objects in the field of view, and occlusion [4] of objects by other objects (we do not encounter these problems in our work). There are also disadvantages to local feature approaches. Construction of geometric models, if a CAD model does not already exist, is much more troublesome than the image-based training process in our method. Furthermore, it is difficult to decide on local features which are appropriate for all objects. For some objects, markings may be more salient than shape [8] (e.g., the lettering on a box of aspirin), however, most models based on local features capture only geometric structure. It is difficult to locate only the proper primitive edge, etc., features in an image. Local techniques tend to have difficulty with objects that have much detail [3] and simple geometric features may not always be reliably obtained from images in many real world applications. Object shape and texture, variations in illumination, occlusion, and noise in the imaging

- *M.A. Sipe is with Cellomics, Inc., Pittsburgh, PA 15219. E-mail: sipe@ieee.org.*
- *D. Casasent is with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: casasent@ece.cmu.edu.*

Manuscript received 18 July 2000; revised 20 July 2001; accepted 6 May 2002. Recommended for acceptance by R.C. Nelson
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 112536.

0162-8828/02/$17.00 © 2002 IEEE

process may limit a geometric feature's visibility. The pose estimates computed using local methods are very sensitive to noisy input data [9]. Finally, local-feature-based recognition methods are generally more computationally demanding than global methods. Since our objects are not occluded by other objects and the background is not cluttered, global features are more robust [10] and more efficient for our application.

Global features may be based on shape, texture, image transforms, moments, or structure [11]. Some of these features (e.g., moments) are designed for invariance to various kinds of distortions.

Our new classification and pose estimation algorithms are based on the *Feature Space Trajectory* [12] (FST) representation for different views (perspective distortion) of a single rigid object. The features used are global. Consider an object viewed at a given range and camera depression angle. As the object rotates about the axis normal to the plane that it rests upon, the *aspect view* of the object changes. In an FST, the different object aspect views are vertices in global feature space. Vertices for adjacent views are connected by line segments so that an FST is created as the viewpoint changes. Each different class of object is represented by a distinct FST. An input test object to be recognized is represented as a point in feature space. In order to classify an unknown object from its features, the FST method computes the Euclidean distance in feature space from the test point to all FSTs known to the system. The class label corresponds to the closest FST. The pose estimate is computed by finding the point on the line segment, on that FST, that is closest to the test point and interpolating the pose from the known poses of the vertices of that line segment.

We can use either a physical prototype or a CAD model to train our system. Our active object recognition system can automatically learn and store an FST representation of each object by moving the camera around a physical object prototype and capturing images from various viewpoints. Alternatively, FSTs may be derived from a CAD model by rendering a set of object images from various viewpoints.

The features employed in the original FST work [12] were wedge and ring samples of the magnitude-squared Fourier transform [13]. Ring features are invariant to in-plane rotations, while wedge features are invariant to changes in scale. Both are invariant to shift. We [14] have compared the performance of the FST using wedge and ring features with that of the FST using features extracted with the Karhunen-Loève (KL) transform [15] and found the performance (for pose estimation) to be superior with KL features. Thus, we use KL features in our present work. Techniques which extract KL features from images are often termed "appearance-based" since the appearance of an object in an image is a function of its shape, reflectance properties, pose, and the illumination conditions [16]. There is a large volume of research using appearance-based methods in eigenspace to detect [17], [18] human faces in images and to search a large databases of faces for those that most closely match an input face [19], [20].

The FST has previously been applied to automatic target recognition (ATR) of military vehicles [14], [21]. In ATR work, the emphasis is on classification rather than pose estimation. In our active object recognition work, pose is also of critical importance. We also introduce new confidence measures that automatically indicate when new aspect views are needed and *actively select* the new views to improve performance.

Murase and Nayar [16] have used techniques quite similar to those of our FST. They represent each object as an appearance manifold in eigenspace (KL features) which is parameterized by object rotation and, in some cases, object scale [22]. They account for variations in illumination by adding degrees of freedom to their appearance manifold for the position of light sources [23]. We differ from this work by minimizing distortions (other than pose) by exercising control over the environment (by fixing the positions of the light sources for example) and modeling nonpose variations by a probability density function (pdf). This prior work did not address actively placing the sensor to obtain the best view of an object or the integration of multiple observations of an object.

Arbel and Ferrie [24] have applied principles similar to ours for selecting desirable viewpoints. They apply the principal of reduction in entropy to select viewpoints while performing active object recognition based on optical flow computation. They also combine multiple observations by applying Bayesian chaining to determine object class however, they do not apply viewpoint selection or combine multiple observations to address the issue of uncertainty in pose estimation as we do.

An extension of the FST to a multidimensional Feature Space Manifold (FSM) to handle multiple degrees of freedom in object pose has previously been discussed [25]. In our work, we handle objects in each of their stable rest positions (this constitutes a second, discrete degree of freedom for object pose) by constructing a separate FST for each rest position of each object. This is more efficient than using an FSM.

This paper is organized as follows: In Section 2, we develop a *probabilistic* FST representation for objects and use it to derive active object recognition capabilities. Section 3 presents experimental results using our approach. We conclude with a discussion in Section 4.

2 Active Object Recognition Theory

In this section, we address the following new FST techniques required for active object recognition:

- a classification confidence measure (Section 2.2.1),
- an uncertainty measure for a pose estimate (Section 2.2.2),
- selecting the best viewpoint for resolving ambiguity in the class of an object (Section 2.3.1),
- selecting the best viewpoint for pose estimation (Section 2.3.2), and
- using multiple observations to produce better estimates and uncertainty measures (Section 2.4).

We proceed by first assuming a specific form for our probabilistic object representation—the pdf for the random feature vector x conditioned on the class ω_i and pose θ of the object (Section 2.1)—and then deriving the new functions listed above using this probabilistic object representation.

2.1 Probabilistic FST Object Representation

The FST object representation explicitly encodes how global features extracted from an image of an object change with aspect view. This is desirable since we need to estimate the pose as well as the class of the object. We must also handle undesirable variations in global features such as: illumination, object texture (dirt, rust, material properties, etc.), and sensor noise. We lump these "nuisance" distortions together and model them collectively as "noise." We consider the observed feature values to be random variables consisting of zero mean random noise added to a deterministic quantity determined by the class and pose of the object. The deterministic quantity is specified by the point on the FST for the object at the given pose. Thus, we view the FST object representation as a basis for deriving conditional pdfs. We apply Bayesian estimation and hypothesis testing theory to these conditional pdfs to derive the required active object recognition system outputs.

Let ω_i, where $\mathrm{i} = 1, 2, 3, \ldots, \mathrm{N_C}$, denote the class hypothesis. Formally, we define an FST $\mathbf{m}_{\omega_i}(\theta)$ as the deterministic vector-valued function that maps the pose parameter (θ) of an object of class ω_i to a point in multidimensional feature space. For a particular object and fixed vision system, the FST describes how the object's features change with θ. If samples include variations within an object *class* or variations in the imaging conditions, a trajectory can be constructed with feature values at each θ that are the *mean* of all samples at a given θ. Thus, we model each observation as a k-dimensional random feature vector

$$\boldsymbol{x} = \mathbf{m}_{\omega_i}(\theta) + \boldsymbol{n}, \tag{1}$$

where random noise (n) is added to a point on the trajectory to account for variations other than perspective distortion. We now derive a specific functional form for the pdf for the observation x conditioned on the class ω_i and pose θ of the object. This is our probabilistic object representation and is denoted by $p_{\boldsymbol{x}|\omega_i,\,\theta}$. First, we assume that $\boldsymbol{n}$ is Gaussian and independent of both object class ω_i and viewpoint. We further assume that $\boldsymbol{n}$ is isotropic—each component of $\boldsymbol{n}$ is independent and identically distributed with variance $\sigma_{\boldsymbol{n}}^2$, which does not vary with the class of the object. We do not claim that this is an accurate model of residual variations, only that it yields useful results. Given sufficient training data, one could easily estimate the parameters of a more complex pdf and adopt a more general form of the results presented here.

Given the assumptions, the pdf for $\boldsymbol{n}$ in (1) becomes

$$p_n(\mathbf{n}) = \frac{\mathbf{1}}{(\mathbf{2}\pi)^{k/\mathbf{2}}\sigma_{\boldsymbol{n}}^k} \exp\,(-\frac{\mathbf{1}}{\mathbf{2}\sigma_{\boldsymbol{n}}^{\mathbf{2}}} \parallel \mathbf{n} - \boldsymbol{\mu_n} \parallel^{\mathbf{2}}), \tag{2}$$

where $\boldsymbol{\mu_n}$ is the mean vector and k is the dimension of vectors $\boldsymbol{x}$ and $\boldsymbol{n}$. Now, we assume that $\boldsymbol{n}$ is zero mean ($\boldsymbol{\mu_n} = \mathbf{0}$). The effect of $\mathbf{m}_{\omega_i}(\theta)$ in (1) is simply to translate the mean of the Gaussian noise model in (2), thus, the pdf for an observation $\boldsymbol{x}$ conditioned on class ω_i and pose θ is

$$p_{\boldsymbol{x}|\omega i,\theta}(\mathrm{x} \mid \omega_i, \Theta) = \frac{\mathbf{1}}{(\mathbf{2}\pi)^{k/\mathbf{2}}\sigma_{\boldsymbol{n}}^k} \exp\left(-\frac{1}{2\sigma_n^2} \parallel \mathbf{x} - \mathbf{m}_{\omega_i}(\Theta) \parallel^2\right). \tag{3}$$

Equation (3) is our probabilistic FST object representation.

We have shown [26] that the FST class and pose estimation methods approximate the maximum a posteriori (MAP) estimators for the class and pose of the object (assuming uniform priors for class and pose). It is not necessary to assume that the noise $\boldsymbol{n}$ is Gaussian to show this; one obtains the same result as long as the noise is purely a function of distance from the FST and decreases with distance.

2.2 Uncertainty and Confidence Measures

In this section, we highlight the uncertainty measures we use to indicate when additional observations are required.

2.2.1 Classification Confidence

We define the confidence $\mathcal{C}_{\omega_\mathrm{i}}$ in class decision ω_i, based on observed feature vector $\mathbf{x}$, to be the approximate a posteriori probability of ω_i, i.e., $\mathcal{C}_{\omega_i}(\mathbf{x}) \approx \mathbf{P}(\omega_i \mid \mathbf{x})$. Using Bayes' theorem [27], we express the a posteriori probability as

$$P(\omega_i \mid \mathbf{x}) = \frac{\mathbf{p}_{\boldsymbol{x}|\omega_i}(\mathbf{x} \mid \omega_i)\mathbf{P}(\omega_i)}{\mathbf{p}_{\boldsymbol{x}}(\mathbf{x})}, \tag{4}$$

where $P(\omega_i) = 1/\mathrm{N_C}$ since we assume equal prior probabilities for each class. We approximate the value of the class conditional density as $p_{\boldsymbol{x}|\omega_i}(\mathbf{x} \mid \omega_i) \approx \mathbf{p}_{\boldsymbol{x}|\omega_i,\theta}(\mathbf{x} \mid \omega_i, \theta_{\omega_i}(\mathbf{x}))$. In effect, we approximate $p_{\boldsymbol{x}|\omega_i}$ by first finding the pose estimate $\widehat{\theta}_{\omega_i}$ for the object, assuming that it is of class ω_i, and then substituting it into $p_{\boldsymbol{x}|\omega_i,\theta}$. Since $\widehat{\theta}_{\omega_i}$ corresponds to the closest point on the FST for class ω_i, the approximation of $p_{\boldsymbol{x}|\omega_i}$ considers only the closest point on the FST. This approximation saves computation time and also has the effect of eliminating undesirable bias. Consider a case, like the cup in Fig. 1, where the appearance of an object is similar over a range of views. If class confidence were computed by integrating over θ, probability would accumulate over the region with similar appearance and bias the result in favor of that object. Our method eliminates such bias and, since our ultimate goal is determining the correct class rather than correctly estimating the posterior class probabilities, our method is justified. The pdf for x in (4) is obtained using the total probability theorem [27] as

$$p_{\boldsymbol{x}}(\mathbf{x}) = \sum_{\mathrm{i}=1}^{\mathrm{N_C}} \mathbf{p}_{x|\omega_i}(\mathbf{x} \mid \omega_i)\mathbf{P}(\omega_i) = \sum_{\mathrm{i}=1}^{\mathrm{N_C}} \mathbf{p}_{\boldsymbol{x}|\omega_i}(\mathbf{x} \mid \omega_i)\frac{\mathbf{1}}{\mathrm{N_C}}. \tag{5}$$

Combining results and simplifying, we obtain the *new classification confidence function* we use:

$$\mathcal{C}_{\omega_i}(\mathbf{x}) = \mathbf{P}(\omega_i \mid \mathbf{x}) \approx \frac{\mathbf{p}_{\boldsymbol{x}|\omega_i,\theta}(\mathbf{x} \mid \omega_i, \widehat{\theta}_{\omega_i}(\mathbf{x}))}{\sum_{\mathrm{p}=1}^{\mathrm{N_C}} \mathbf{p}_{\boldsymbol{x}|\omega_i,\theta}(\mathbf{x} \mid \omega_\mathrm{p}, \widehat{\theta}_{\omega_\mathrm{p}}(\mathbf{x}))}. \tag{6}$$

We evaluate $p_{\boldsymbol{x}|\omega_i,\theta}$ in (6) for each class ω_i in by substituting the squared distance from the observation $\mathbf{x}$ to each corresponding FST (this is computed in the classification step) for $\|\mathbf{x} - \mathbf{m}_{\omega_i}(\Theta)\|^2$ in (3). If ω_i is the class with the maximum a posteriori probability (corresponding to the closest FST to the observation), then the numerator is $p_{\boldsymbol{x}|\omega_i,\theta}$ (computed using the minimum squared distance) and the denominator is the sum of the pdfs for all classes. *We use (6) to decide if it is necessary to take additional observations before finalizing the class decision.*

2.2.2 Pose Estimation Uncertainty

We now introduce our new uncertainty measure $\mathcal{U}_{\omega_\mathrm{i}}(\mathbf{x})$ for a pose estimate from an observation $\mathbf{x}$. We use this measure to decide if it is necessary to collect additional observations before finalizing the pose estimate. There are two causes for error in pose estimates. The first is pose ambiguity. Consider the coffee cup object in Fig. 1. The pose of the cup is ambiguous when its distinguishing part (the handle)

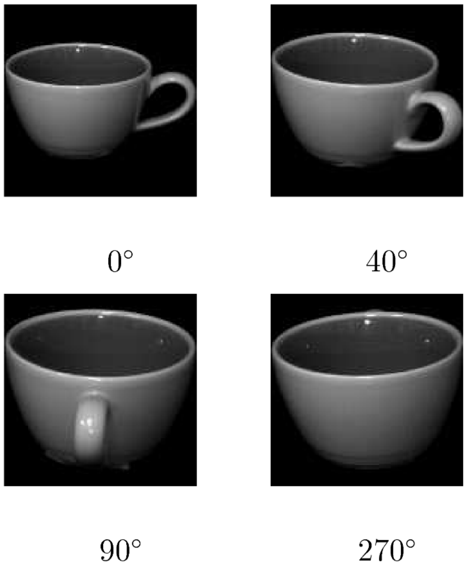

Fig. 1. Selected views of the coffe cup object.

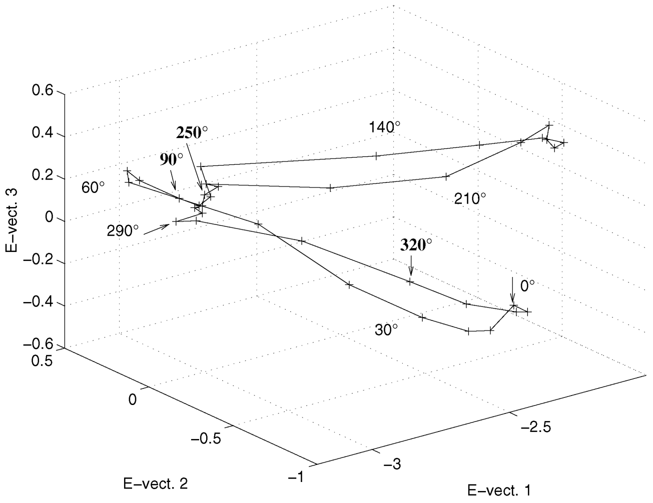

Fig. 2. FST for the coffe cup with $N_{KL} = 3$.

is occluded by the object itself, as is the case in a range of views around 270°. The second cause of error is the collection of variations we refer to as noise. Our uncertainty measure accounts for both causes.

The uncertainty measure we use is the expected value of the magnitude of the difference between the true and estimated poses conditioned on the observation $\mathbf{x}$. The aspect angle θ of an object is periodic (with a typical period $\Theta_{T\omega_i}$ of 360°) therefore, in calculating the magnitude of the difference between the true θ and estimated $\hat{\theta}$ poses, we use $\phi_{\omega_i}(\theta - \widehat{\theta}_{\omega_i}(\mathbf{x}))$ where $\phi_{\omega_i}(\Delta\,\theta) = \min(|\,\Delta\,\theta\,|\,,\,\Theta_{T\omega_i} -\,|\,\Delta\,\theta\,|)$ and $\Delta\theta = \theta - \hat{\theta}$. Applying the definition of the conditional expected value [27 p. 169], our pose estimate uncertainty measure is

$$\begin{aligned}\mathcal{U}_{\omega_i}(\mathbf{x}) &= E[\phi_{\omega_i}(\theta - \widehat{\theta}_{\omega_i}(\mathbf{x})) \mid \omega_i, \mathbf{x}] \\ &= \int_{-\infty}^{\infty} \phi_{\omega_i}(\Theta - \widehat{\theta}_{\omega_i}(\mathbf{x}))\mathbf{p}_{\theta|\omega_i,\boldsymbol{x}}(\Theta \mid \omega_i, \mathbf{x})\mathbf{d}\Theta,\end{aligned} \tag{7}$$

where $p_{\theta|\omega_i,\boldsymbol{x}}$ is the pdf for θ conditioned on the object class and on observation $\mathbf{x}$. It is derived from Bayes' theorem [27] as

$$p_{\theta|\omega_i,\boldsymbol{x}}(\Theta \mid \omega_i, \mathbf{x}) = \frac{\mathbf{p}_{\boldsymbol{x}|\omega_i,\theta}(\mathbf{x} \mid \omega_i, \Theta)p_{\theta|\omega_i}(\Theta \mid \omega_i)}{\mathbf{p}_{\boldsymbol{x}|\omega_i}(\mathbf{x} \mid \omega_i)}. \tag{8}$$

We assume $p_{\theta|\omega_i}$ is uniform over the allowed range of motion and, therefore, a constant (typically, 1/360), and we evaluate $p_{\boldsymbol{x}|\omega_i}$ by multiplying $p_{\boldsymbol{x}|\omega_i,\theta}$ by $p_{\theta|\omega_i}$ and integrating over Θ as

$$p_{\boldsymbol{x}|\omega_i}(\mathbf{x} \mid \omega_i) = \int_{-\infty}^{\infty} \mathbf{p}_{\boldsymbol{x}|\omega_i,\theta}(\mathbf{x} \mid \omega_i, \Theta)p_{\theta|\omega_i}(\Theta \mid \omega_i)d\Theta. \tag{9}$$

Therefore, we have specified everything needed to compute (8). We substitute (8) into (7) and evaluate the integral in (7) for a given pose estimate $\hat{\theta}$. A smaller value of $\mathcal{U}_{\omega_i}(\mathbf{x})$ indicates a more reliable pose estimate.

We gain insight into the usefulness of this metric by looking at the specific example of the coffee cup object in Fig. 1. For a given observation $\mathbf{x}$, there is a closest point on its FST (Fig. 2) that determines $\widehat{\theta}$. There are *different ranges* of θ values that (with noise) have a high probability of having produced the observation $\mathbf{x}$. Fig. 3a shows a hypothetical pdf, $p_{\theta|\omega_i,\boldsymbol{x}}(\Theta \mid \omega_1, \mathbf{x}_1)$ for θ conditioned on observation $\mathbf{x}_1$ and the object class (we expect something like Fig. 3a for an observation of the cup at the 270° viewpoint). Two different θ ranges (80° to 100°, 250° to 290°) are shown to have high probability. This occurs because the 270° observation is very close to two different portions of the FST (the 250°-290° range and the 80°-100° range). The pose estimate $\widehat{\theta}(\mathbf{x}_1)$ is determined by the most likely value of θ (the center of the 250° to 290° range), but there is also a high probability that the actual value of θ lies in the range from 80° to 100°. In this case, the pose estimate derived from $\mathbf{x}_1$ at 270° should be deemed unreliable. Fig. 3b shows a hypothetical pdf for θ conditioned on another observation $\mathbf{x}_2$ (we would expect to see something like Fig. 3b for an observation of the cup around a 40° or 140° viewpoint). In this case, the pdf is unimodal and highly concentrated so we consider the pose estimate $\widehat{\theta}(\mathbf{x}_2)$ derived from $\mathbf{x}_2$ to be more reliable. The value of $\mathcal{U}_{\omega_i}(\mathbf{x}_2)$ will be much smaller than the value of $\mathcal{U}_{\omega_i}(\mathbf{x}_1)$ for the cases in Fig. 3. These observations may be obvious to a human observer; however, our new use of the FST provides an *automated* analysis technique.

2.3 Sensor Movement Strategy

In Section 2.2, we discussed the uncertainty and confidence measures which tell us when it is necessary to collect additional data. In this section, we discuss *where* to collect that data, i.e., which new aspect view, or viewpoint, do we choose. For each object known to our active object recognition system, we analyze the probabilistic FST object

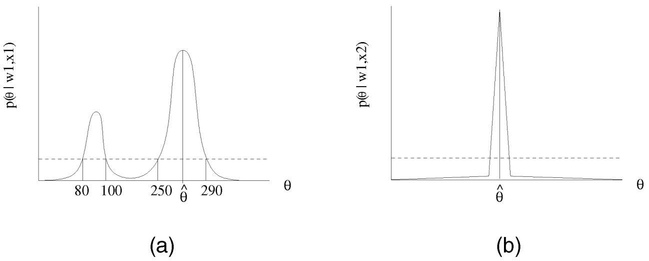

Fig. 3. pdf for θ conditioned on (a) $\mathbf{x}_1$, (b) $\mathbf{x}_2$.

representation *offline* and *automatically* determine and store the best viewpoints for resolving class and pose ambiguities. At runtime, we use the pose estimate from the current observation to compute the sensor motion required to obtain the best viewpoint for classification or pose estimation. This approach assumes that the cost of making an error is much larger than the cost of moving the sensor.

2.3.1 Viewpoint Selection for Best Classification

When the active object recognition system must move the camera to resolve class ambiguity, object viewpoints where the presence of distinguishing characteristics are clearly visible are obviously preferred. We *automatically* select the *best* view to resolve ambiguity, as we now discuss.

Given an initial observation $\mathbf{x}_1$, we seek the sensor rotation $\Delta\Theta$ which maximizes the a posteriori probability of correct classification (which, in effect, reduces the Shannon entropy, as in [24]). We restrict the targeted view to be a training view (FST vertex) of an object. This eliminates any error introduced by the piecewise linear FST representation. The best training viewpoint for discrimination of objects is, approximately, the one with its FST vertex most distant from other FSTs.

We denote the best pose of an object of class ω_i to use in distinguishing it from an object of class ω_j as $\Theta_c(i,j)$. The values of $\Theta_c(i,j)$ (computed offline and stored for each pair of objects) form a matrix that specifies the best camera view for resolving class ambiguity between any pair of objects known to the recognition system. In each iteration of an active object recognition scenario, the system makes an observation. If the classification confidence $\mathcal{C}_{\omega_{\rm i}}$ ((6)) is not sufficiently high after the observation, the system notes the two most likely classes, looks up the best view for distinguishing them, and then drives the camera to that viewpoint using the pose estimate from the current observation. Although we consider only the two most likely classes at each step, our active object recognition system still resolves cases when more than two objects may be confused. Often, there is a single salient view which distinguishes a set of similar parts. Even if this is not the case, moving the sensor to the best viewpoint to discriminate between two objects after each observation tends to discriminate multiple similar objects by a process of elimination.

2.3.2 Viewpoint Selection for Best Pose Estimation

Errors in pose estimation will vary with the sensor viewpoint. In the case of the cup object in Fig. 1, it is not possible to obtain a reliable pose estimate for views in which the handle is not visible (e.g., views around 270°). This is apparent in the FST of the coffee cup object in Fig. 2. Errors in pose estimation are likely to be larger at viewpoints where different parts of the same FST are close in feature space but far apart in aspect angle. The object aspect views around 270° are tightly clustered on the left of the FST (Fig. 2) since the differentiating object feature (the handle) is barely visible in this aspect view range. Thus, we expect large errors in the pose estimate around 270° and smaller errors around 140° and 30° where the handle is visible (the FST in Fig. 2 is more spread out in these angular regions). Thus, the FST representation contains the information required to find the best viewpoint of the object for estimating its pose.

Our new method *automatically finds the best viewpoint of an object for estimating its pose* $\Theta_\theta(\omega_i)$; we define this as the viewpoint with the least pose uncertainty ((7)). Therefore, we find the viewpoint $\Theta_\theta(\omega_i)$ which minimizes the expected value of the magnitude of the pose estimation error (the pose estimation uncertainty in (7)),

$$\Theta_\theta(\omega_i) = \arg\min_{\Theta} E[\phi_{\omega_i}(\theta - \widehat{\theta}_{\omega_i}(\mathbf{x})) \mid \omega_i, \Theta]. \tag{10}$$

We precompute and store this viewpoint $\Theta_\theta(\omega_i)$ for each object ω_i as part of the offline training process. As noted earlier for the classification case, we restrict $\Theta_\theta(\omega_i)$ to be one of the training set views of the object. We estimate $E[\phi_{\omega_i}(\theta\text{-}\widehat{\theta}_{\omega_i}(\mathbf{x})) \mid \omega_i, \Theta]$ at each training view using Monte Carlo (statistical simulation) methods and we select the one that yields the lowest expected error. We use sufficient noise trials (typically less than 100,000) in the simulation to guarantee that the estimate is within $\pm 0.5^\circ$ of the true expected value with probability 0.95. We chose the noise standard deviation σ_n such that the probability is 0.90 that any observation of that object is within a distance $\bar{l}$ of the FST, where $\bar{l}$ is the average distance between adjacent vertices of the FST for that object. This yields a reasonable spread of the pose estimation error for different viewpoints. Our testing has shown that the value of σ_n used is not critical when finding $\Theta_\theta(\omega_i)$.

2.4 Multiobservation Fusion

If the class or pose estimate uncertainties (Section 2.2) are unacceptable, we move the sensor and take another observation. In this section, we highlight the process of fusing the new observation with prior ones to form new estimates and new uncertainty measures. This requires little additional computational overhead beyond the standard FST distance computations.

Consider the classification problem first. We have a set of $\mathrm{N_O}$ observations, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\mathrm{N_O}}$, which we denote by $\wedge_{\rm j}\mathbf{x}_{\rm j}$. In active object recognition, the motion of the sensor is also known. We use all observations *and* the known sensor motions in deriving class and pose estimates. We make the reasonable assumption that the error in sensor motion is minute and treat the rotations $\Delta\Theta_{\rm j}$ between sensor viewpoints as *deterministic* quantities. Using these known sensor viewpoint rotations, we express the object pose associated with each observation with respect to the sensor in its current (last) position. After the transformation into this common coordinate system, there is only one unknown pose in the problem: $\theta_{\mathrm{N_O}}$, the pose of the object for the final viewpoint. We now express the pose $\theta_{\rm j}$ at each step in terms of $\theta_{\mathrm{N_O}}$ and the known sensor rotations. Let $\Delta\Theta_{\rm j}$ denote the rotation from the final sensor viewpoint (at step $\mathrm{N_O}$) to the sensor viewpoint at step j. The pose $\theta_{\rm j}$ of the object at step j may now be expressed in terms of the final object pose $\theta_{\mathrm{N_O}}$ as $\theta_{\rm j} = \theta_{\mathrm{N_O}} + \Delta\Theta_{\rm j}$.

Let $\wedge_{\rm j}\Delta\Theta_{\rm j}$ denote the set of known sensor rotations $\Delta\Theta_1, \Delta\Theta_2, \ldots, \Delta\Theta_{\mathrm{N_O}\text{-}1}$ and let $p_{\wedge_{\rm j}\boldsymbol{x}_{\rm j}|\omega_{\rm i},\theta_{\mathrm{N_O}}}$ denote the joint pdf of the observations $\wedge_{\rm j}\mathbf{x}_{\rm j}$ conditioned on the (unknown) final pose $\theta_{\mathrm{N_O}}$ with the known sensor motions $\wedge_{\rm j}\Delta\Theta_{\rm j}$ as deterministic parameters. We use $p_{\wedge_{\rm j}\boldsymbol{x}_{\rm j}|\omega_{\rm i},\theta_{\mathrm{N_O}}}$ to express *the multiobservation equation* for the MAP pose estimate as

$$\widehat{\theta}_{\omega_{\rm i}}\Big(\bigwedge_{\rm j}\mathbf{x}_{\rm j}, \bigwedge_{\rm j}\Delta\Theta_{\rm j}\Big) = \arg\max_{\Theta_{\mathrm{N_O}}} p_{\wedge_{\rm j}\boldsymbol{x}_{\rm j}|\omega_{\rm i},\theta_{\mathrm{N_O}}}\Big(\bigwedge_{\rm j}\mathbf{x}_{\rm j} \mid \Theta_{\mathrm{N_O}}, \bigwedge_{\rm j}\Delta\Theta_{\rm j}\Big). \tag{11}$$

Similarly, by substituting $p_{\wedge_{\rm j}\boldsymbol{x}_{\rm j}|\omega_{\rm i},\theta_{\mathrm{N_O}}}$ for $p_{\boldsymbol{x}_{\rm j}|\omega_{\rm i},\theta}$ in the corresponding single observation equations, it is easy to derive the multiobservation equations for the pose estimate

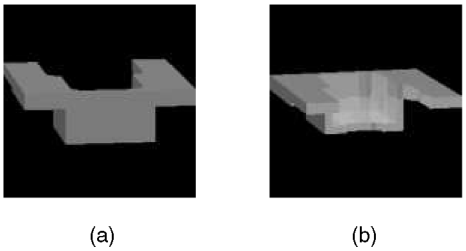

Fig. 4. Rendered views of bracket 1 in rest position 3. (a) 15° aspect. (b) 195° aspect.

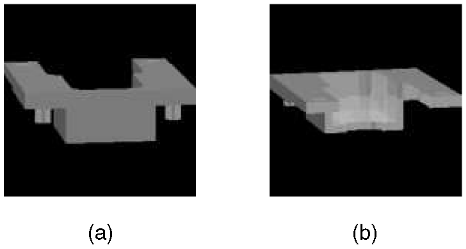

Fig. 5. Rendered views of bracket 2 in rest position 3. (a) 15° aspect. (b) 195° aspect.

uncertainty ((7)) and the a posteriori class probability and classification confidence ((6)). To evaluate these quantities, we derive the joint conditional pdf

$$p_{\wedge_{\mathrm{j}} \boldsymbol{x}_{\mathrm{j}} | \omega_{\mathrm{i}}, \theta_{\mathrm{N_O}}}(\bigwedge_{\mathrm{j}} \mathbf{x}_{\mathrm{j}} \mid \omega_{\mathrm{i}}, \Theta_{\mathrm{N_O}}, \bigwedge_{\mathrm{j}} \Delta\Theta_{\mathrm{j}}) = \prod_{\mathrm{j}=1}^{\mathrm{N_O}} p_{\boldsymbol{x}|\omega_{\mathrm{i}},\theta}(\mathbf{x}_{\mathrm{j}} \mid \omega_{\mathrm{i}}, \Theta_{\mathrm{N_O}} + \Delta\Theta_{\mathrm{j}}), \tag{12}$$

by assuming that the observations x_{j} are statistically independent. This is easily done by substituting (3) into (12) and simplifying to obtain

$$p_{\wedge_{\mathrm{j}} \boldsymbol{x}_{\mathrm{j}} | \omega_{\mathrm{i}}, \theta_{\mathrm{N_O}}}(\wedge_{\mathrm{j}} \mathbf{x}_{\mathrm{j}} \mid \omega_{\mathrm{i}}, \Theta_{\mathrm{N_O}}, \wedge_{\mathrm{j}} \Delta\Theta_{\mathrm{j}}) = \frac{1}{(2\pi)^{\mathrm{N_O k}/2} \sigma_n^{\mathrm{N_O k}}} \exp(-\frac{1}{2_{\sigma_n^2}} \Sigma_{\mathrm{j}=1}^{\mathrm{N_O}} \parallel \mathbf{x}_{\mathrm{j}}\text{-}\mathbf{m}_{\omega_{\mathrm{i}}}(\Theta_{\mathrm{N_O}} + \Delta\Theta_{\mathrm{j}}) \parallel^2). \tag{13}$$

3 Experimental Results

We refer to the learned object information (FSTs and the best object viewpoints for classification and pose estimation) as the *FST knowledge base*. For this case study, the knowledge base contains four parts: the two brackets in Fig. 4 and Fig. 5, the socket in Fig. 6, and a similar socket object in Fig. 7. The images in Fig. 4, Fig. 5, Fig. 6, and Fig. 7, were ray-traced from CAD models. We train our active object recognition system using images rendered from CAD models of each part and then recognize real versions of bracket 1 (Fig. 8) and socket 1 (Fig. 9) since we possessed real metal prototypes for those two parts only. Consider the two different brackets in Fig. 4 and Fig. 5. Bracket 2 is identical to bracket 1 except for the addition of two small circular bosses (cylindrical protrusions visible in the foreground of Fig. 5a) and two holes through the top surface. Due to the small camera depression angle, the holes are barely visible in the rest position shown. Distinguishing bracket 2 from bracket 1 is a good test case for active vision because the task is only possible when the circular bosses are visible. Socket 1 (Fig. 6) is an interesting test case for active pose estimation because the left and right (side) views are indistinguishable and the front and back views are distinguishable only by the larger holes in the back view of the object. Socket 2 differs from socket 1 in the width of the mounting bracket, as seen in Fig. 6b and Fig. 7b. Brackets 1 and 2 may appear in any of the three stable rest positions shown for bracket 1 in Fig. 8. Sockets 1 and 2 may appear in either of the two stable rest positions shown for socket 1 in Fig. 9. Thus, the set of machine parts consists of 2 x 3 + 2 x 2 = 10 object/rest position combinations. We placed the real metal prototypes on a rotating stage to capture images from various aspect angles.

We consider this a 10 class pose and class estimation problem. Our goals for this recognition task are high confidence in the object's class and moderate accuracy in the estimate of its pose. In our tests, we obtained additional object views if the classification confidence $\mathcal{C}_{\omega_{\mathrm{i}}}$ was less than 0.9 or if the pose estimate uncertainty $\mathcal{U}_{\omega_{\mathrm{i}}}$ was more than 3.5°.

All images (rendered and real) were taken from a 9° camera depression angle and were preprocessed such that the object filled at least one dimension (this provides invariance to changes in scale) of the final 128×128 pixel image (the images in Fig. 8 and Fig. 9 are shown prior to preprocessing). Realistic images of each object were rendered by carefully replicating laboratory conditions in the CAD environment [28]. We used 120 images of each object in each rest position (in 3° increments over the full 360° aspect angle range) to construct each FST. Eighty KL features (retaining 95 percent of the variability in the rendered training data) were used to extract features from both rendered and real images. An FST was constructed for each object, in each rest position, from the rendered aspect views. Real images were used as the test set for class and pose estimation.

For each of the 10 initial FSTs, we applied a *new active vertex selection method* that indicates new FST vertex viewpoints

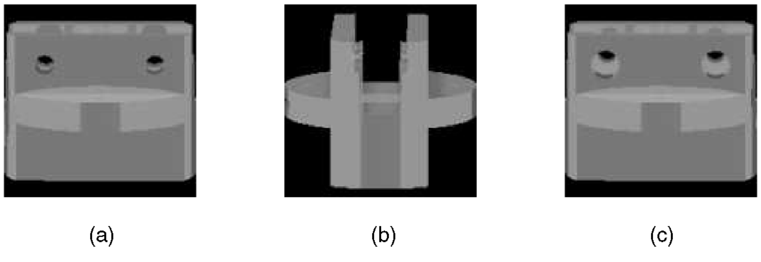

Fig. 6. Rendered views of socket 1. (a) front (0°), (b) side $(90^\circ/270^\circ)$, and (c) back (180°).

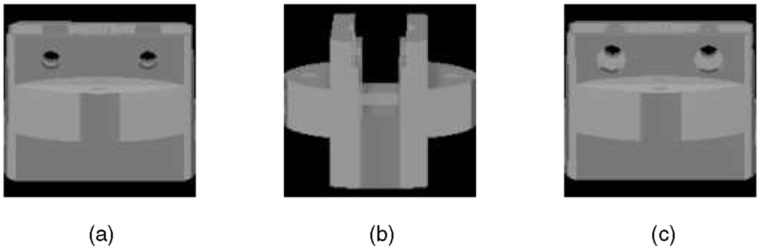

Fig. 7. Rendered views of socket 2. (a) front (0°), (b) side ($90^\circ/270^\circ$), and (c) back (180°).

needed to capture sharp transitions in object appearance due to specular reflections from the shiny metal surfaces. The number of FST vertices was then reduced via a suboptimal pruning technique whereby the vertices least essential for representing the training set are removed in a sequential fashion [14] until a threshold in the representation error is reached. The final number of vertices used for these tests ranged from 124 to 162 for the 10 object/rest position cases. It is possible to produce representations which are much more compact by pruning more aggressively.

3.1 Finding the Best Views for Classification

For each object class $\omega_{\rm i}$, we determined the aspect angle $\Theta_c(\omega_{\rm i}, \omega_{\rm j})$ of the object which best distinguishes that object class from each of the other object classes $\omega_{\rm j}$. As discussed in Section 2.3.1, for the case of two FSTs, this is the training view of the object whose FST vertex is most distant from the other FST. $\Theta_c(\omega_{\rm i}, \omega_{\rm j})$ is stored for each pair of classes in the FST knowledge base.

To predict which objects may be confused (and at what viewpoints), we observe the distance between pairs of FSTs as a function of the aspect angle θ. For example, in Fig. 10, we plot the distances from the vertices of the FST for bracket 1 to the FST for bracket 2, both in rest position 3, as a function of the aspect angle θ of bracket 1. Bracket 1 and bracket 2, *in the same rest position*, are very similar in some aspect ranges. The spike in the distance between FSTs near 260° in Fig. 10 is caused by a rapid change in image appearance due to the effects of specular reflection. Studying Fig. 10, we see that there are ranges of views (60°-111°, 168°-192°, and 252°-300°) where the distance between the FSTs is very small. If bracket 1 in rest position 3 is viewed in these ranges, it will be necessary to rotate the viewpoint to resolve the ambiguity with bracket 2 at the same pose. From Fig. 10, we find that the best viewpoint is 3°, where the inter-FST distance is maximum.

3.1.1 Finding the Best Viewpoints for Pose Estimation

In order to find the best view $\Theta_\theta(\omega_{\rm i})$ of each object class $\omega_{\rm i}$ for pose estimation, we used the Monte Carlo technique described in Section 2.3.2 to find an estimate $\widehat{\phi_{\omega_{\rm i}}}$ of the expected value of the pose estimation error magnitude for each training view of the object. The standard deviation of the added noise σ_n was chosen independently for each object using the average FST link length as described in Section 2.3.2. For each of the 10 object classes $\omega_{\rm i}$, $\Theta_\theta(\omega_{\rm i})$ is stored in the FST knowledge base so that, when the pose estimate uncertainty $\mathcal{U}_{\omega_{\rm i}}$ is too high, the active object recognition system knows to drive the sensor to obtain the best $\Theta_\theta(\omega_{\rm i})$ object aspect view for pose estimation.

We plot $\widehat{\phi_{\omega_{\rm i}}}$ as a function of aspect angle for rest position 1 of socket 1 in Fig. 11 using rendered CAD data. This plot indicates the benefits of active object recognition for this object. As expected, pose estimation is predicted to be very unreliable for views near the sides ($\theta = 90^\circ$ or 270°) of this object. There are viewpoints where the pose estimates are expected to be better, but estimating the pose of this object is difficult from any viewpoint due to its high degree of rotational symmetry. The object features which distinguish the front of the object from the back are subtle, resulting in significant 180° ambiguity in θ even at the best viewpoints. Since the minimum value of $\widehat{\phi_{\omega_{\rm i}}}$ for socket 1 in rest position 1 at any view is 24.6° in Fig. 11, we do not expect that we will be able to estimate θ with acceptable uncertainty $\mathcal{U}_{\omega_{\rm i}}$ for these objects using *any* single observation. However, when

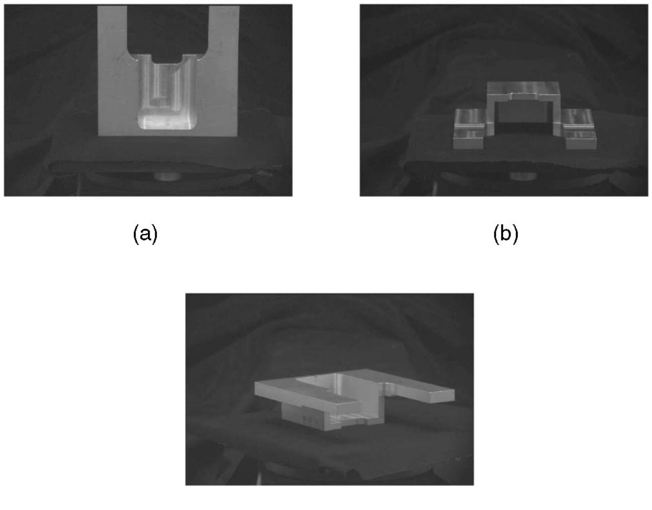

Fig. 8. Real imagery of bracket-1 in different rest positions. (a) Position #1, (b) position #2, and (c) position #3.

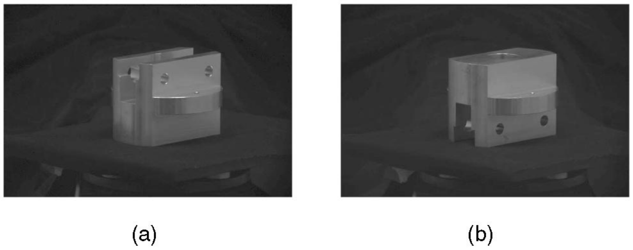

Fig. 9. Real imagery of socket-1 in different rest positions. (a) position #1 and (b) position #2.

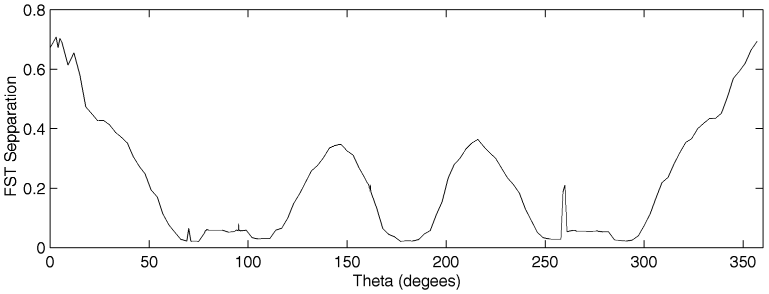

Fig. 10. Separation between the FSTs for brackets 1 and 2 as a function of the pose of bracket 1. Both objects are in rest position 3.

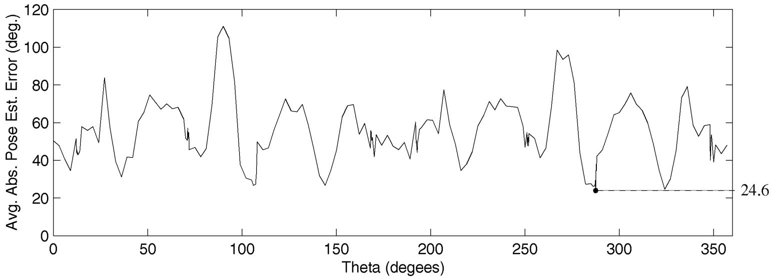

Fig. 11. Estimated expected pose estimation error magnitude as a function of pose for socket 1 in rest position 1.

multiple observations are combined, we show (Section 3.2) that $\mathcal{U}_{\omega_i}$ is often acceptable.

The analyses of socket 1 in rest position 2 and socket 2 in both rest positions are similar to that noted above for socket 1 in rest position 1. The worst case pose estimation error magnitude for the two bracket objects is much better, less than 3°, therefore, we expect that any view of these objects will be sufficient for pose estimation.

These FST *analyses for best classification and pose estimation viewpoints (Sections 3.1 and 3.1.1) are performed offline and automatically. The best poses for classification and pose estimation are stored in our knowledge base and are used in our online active object recognition tasks.* Using a distinct training set of real images, we selected a noise model variance σ_n of 0.1795 for use in (3) and (13).

3.2 Active Object Recognition Tests

Images of both real object prototypes (socket 1 and bracket 1) were captured in the lab in each stable rest position (two for socket 1 and three for bracket 1) at 72 different aspect views (at 5° increments of aspect angle (θ) over the 360° aspect range). For each of the two prototype objects in each starting pose (5 classes/rest positions x 72 aspect angles = 360 tests total), we performed active classification and pose estimation tests. We first describe the results from one of these tests in detail. We then discuss summary statistics for all 360 tests.

We consider a test input of the 175° aspect view of the bracket 1 object in rest position 3 (class 33) as the first observation $\mathbf{x}_1$ as shown in Fig. 12a. This is not a training view and, since the presence of the circular bosses cannot be ascertained from this view, it is difficult for the classifier to decide between bracket 1 and 2 in rest position 3 (classes 33 and 43). The object was misclassified as bracket 2 in rest position 3 ($\hat{\omega}_i = \omega_{43}$) after the first observation $\mathbf{x}_1$, but the confidence level $\mathcal{C}_{\omega_i}(\mathbf{x}_1) = \mathbf{0.513}$ was very low and, thus, the class decision was known to be unreliable. The pose estimate was $\hat{\theta} = 173.8^\circ$; from this and the most likely classes (43 and 33), we compute the needed viewpoint rotation. The predetermined best viewpoint for discriminating between classes 43 and 33 $\Theta_c(\omega_{43}, \omega_{33})$ is 3°. Thus, the required rotation is

$$\Delta\Theta = \Theta_c(\omega_{43}, \omega_{33})\text{-}\hat{\theta} = 3^\circ - 173.8^\circ = -170.8^\circ.$$

Since (Fig. 12a) the pose estimation error for $\mathbf{x}_1$ is -1.2° after the first observation, the system misses the best viewpoint (3°) by 1.2°; however, the image from the 4.2° view (Fig. 12b) gives the correct object class ($\hat{\omega}_i = \omega_{33}$) with high classification confidence $\mathcal{C}_{\omega_i}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{0.999}$. Since the pose estimate uncertainty $\mathcal{U}_{\omega_i}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{0.4}^\circ$ is also below the 3.5° threshold, there is no need to move the sensor further to improve the pose estimate. The final pose estimation error is only 0.6°.

We now discuss summary statistics for all of our tests. In total, the object was classified correctly and confidently ($\mathcal{C}_{\omega_i} > 0.90$) after the first observation in only 27.5 percent of the tests. The classification success rate on the first observation for each class ranged from 14 percent for bracket 1 in rest position 3 to 38 percent for socket 1 in rest position 1. In all of the remaining cases (72.5 percent), $\mathcal{C}_{\omega_i}$ was less than 0.90, triggering a viewpoint rotation to Θ_c and another observation. In 30 percent of the low confidence cases, the most likely class identified by the FST was not the correct class. This fact eliminates the possibility that a hidden systematic illumination difference in the training data was responsible for the recognition results obtained. In all cases, the object was classified correctly and with high $\mathcal{C}_{\omega_i}$ after a maximum of three observations. These results constitute clear and convincing evidence that *our classification confidence $\mathcal{C}_{\omega_i}$ is effective in identifying views where the object class is ambiguous and our strategy for relocating the sensor efficiently resolves such ambiguities.*

After the first observation, the pose estimate uncertainty $\mathcal{U}_{\omega_i}$ was acceptable (less than 3.5°) for all of the tests on bracket 1 but was unacceptably high for all of the tests on socket 1. The FST analysis in Section 3.1.1 predicted this result. The best

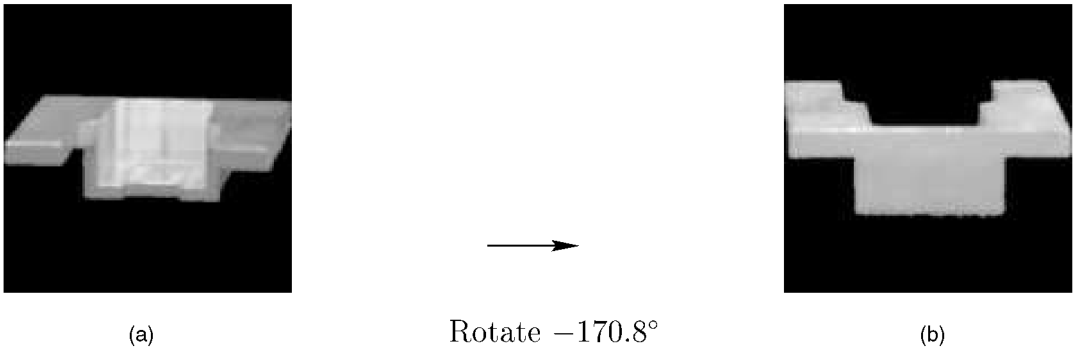

Fig. 12. Active recognition scenario for bracket 1 initially in rest position 3 (class ω_{33}) at $\theta = 175°$ (a) $\hat{\omega}_{\mathrm{i}} = \omega_{43}, \mathcal{C}_{\omega_{\mathrm{i}}}(\mathbf{x_1}) = \mathbf{0.513}$. (b) $\hat{\omega}_{\mathrm{i}} = \omega_{33}, \mathcal{C}_{\omega_{\mathrm{i}}}(\mathbf{x_1}, \mathbf{x_2}) = \mathbf{0.999}$.

single observation of socket 1 was also inadequate for pose estimation; however, when multiple (between two and four) observations were combined, the pose estimate uncertainty was acceptable in all but 12.5 percent (18/144) of the tests on socket 1. *Thus, our fusion of multiple aspect estimates with known camera rotation is essential.* Active pose estimation failed in 18 tests on socket 1. In these tests, $\mathcal{U}_{\omega_{\mathrm{i}}}$ was greater than $3.5°$; the active pose estimation halted—even though the pose estimate uncertainty requirement had not been met—because the estimated pose was very close to the stored best pose. When $\mathcal{U}_{\omega_{\mathrm{i}}}$ was less than $3.5°$, the average pose estimation error was less than one degree, well within our specifications. There were only two cases where the pose estimation error was greater than $3.5°$ when $\mathcal{U}_{\omega_{\mathrm{i}}}$ was less than $3.5°$. We traced both of these errors to a mismatch between lighting conditions in the CAD model and the lab.

3.3 Tests for Robustness

The prior tests have shown that our active object recognition system is tolerant to the differences between real and rendered images and to segmentation error. This section demonstrates the tolerance of our FST algorithms for other types of distortion.

In the first series of tests, we simulated positional shifts of the object on a planar surface by shifting the position of the camera ± 2 inches along the x axis (right to left) and shifting the position of the stage holding the object 0 to -16 inches along the z axis (front to back). Our preprocessing step of cropping the region containing the object from the input image is designed to limit the effects of positional shifts of the object; however, shifts still cause some perspective distortion and segmentation errors which our FST processing must tolerate.

We placed bracket 1 on the rotating stage in rest position 2 (class 32) and $\theta = 9°$. We selected this view because it is the best pose for distinguishing between the two different brackets. In the nominal position (x = z = 0), the object was classified correctly and with high confidence and the pose estimate was $9.8°$ with low uncertainty. We varied both Δz from -16 to 0 and Δx from -2 to +2, in one inch increments and ran active object recognition tests for each Δx, Δz combination (a total of 85 tests). Fig. 13 shows the original (raw) and processed images at the extremes of the shifted positions. In seven of the 85 tests, the initial $\mathcal{C}_{\omega_{\mathrm{i}}}$ was below the threshold of 0.9 (in two of these seven, bracket 1 was misclassified as bracket 2), however, after a second observation, all objects were classified correctly and with high confidence.

We also performed tests of robustness to moderate changes in lighting conditions using rendered images. For these tests we turned off selected light sources and ran active object recognition tests on bracket 1 rendered in rest position 3 at various aspect angles In all cases, the object was classified correctly and the largest pose estimation error was only $1.9°$.

4 Discussion

We have presented a new method for active object recognition using global image features. The method is based on a probabilistic extension to the feature space trajectory (FST) representation for 3D objects. The FST

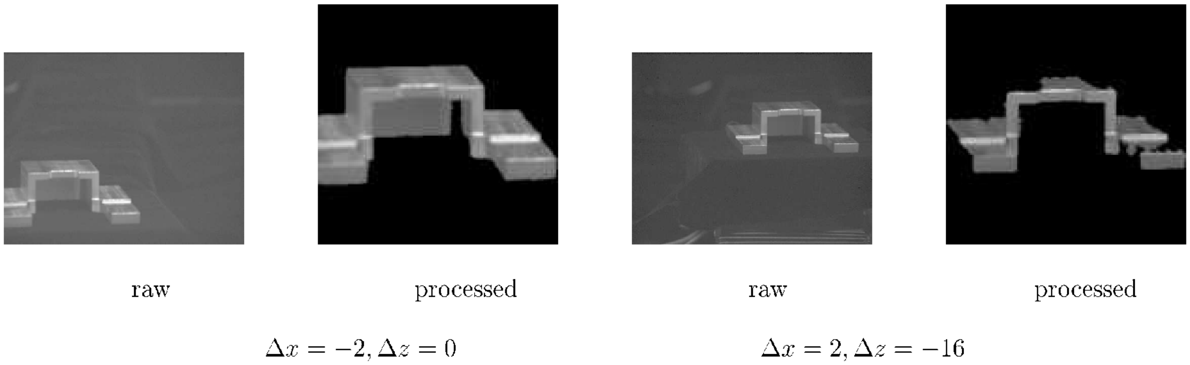

Fig. 13. CIL images of bracket 1 in rest position 1 before (raw) and after (completely processed) segmentation for different object shifts (shown in inches).

representation is compact and easily analyzed to determine the best view of an object to use in estimating its pose or discriminating it from other objects.

FST runtime computations are dominated by inner products of the observed feature vector with each FST vertex. The computational load is thus $O\ (kN_oN_v)$, where k is the number of features, N_o is the number of observations, and N_v is the total number of FST vertices. In practice, the computational burden is amenable to real-time operation.

We have demonstrated the utility of our active object recognition system and its tolerance for the differences between real and rendered imagery.

In prior work [29], we detailed a case study involving a simple assembly in which we recognized the assembly and each of its constituent parts individually. We also introduced a mechanism for rejecting untrained objects, and demonstrated the process of updating the feature space, FSTs, and best viewpoint information for two subassemblies.

Acknowledgments

The support of this work by grant IRI9530657 from the US National Science Foundation is gratefully acknowledged.

References

[1] S.A. Hutchinson and A.C. Kak, “Multisensor Strategies Using Dempster-Shafer Belief Accumulation,” *Data Fusion in Robotics and Machine Intelligence,* M.A. Abidi and R.C. Gonzalez, eds., pp. 165-209, Academic Press, 1992.

[2] F.G. Callari and F.P. Ferrie, “Active Recognition: Using Uncertainty to Reduce Ambiguity,” *Proc. 13th Int'l Conf. Pattern Recognition,* vol. 1, pp. 925-929, Aug. 1996.

[3] S.J. Dickinson, H.I. Christensen, J.K. Tsotsos, and G. Olofsson, “Active Object Recognition Integrating Attention and Viewpoint Control,” *Computer Vision and Image Understanding,* vol. 67, no. 3, pp. 239-260, 1997.

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

[5] Y. Lamdan and H. Wolfson, “Geometric Hashing: A General and Efficient Model-Based Recognition Scheme,” *Proc. Int'l Conf. Computer Vision,* pp. 238-249, Dec. 1988.

[6] F. Stein and G. Medioni, “Structural Indexing: Efficient 3D Object Recognition,” *IEEE Trans. Pattern and Machine Intelligence,* vol. 14, no. 2, pp. 125-144, Feb. 1992.

[7] A.E. Johnson and M. Hebert, “Efficient Multiple Model Recognition in Cluttered 3D Scenes,” *Proc. Computer Vision and Pattern Recognition (CVPR '98),* pp. 23-25, June 1998.

[8] M.J. Black and A.D. Jepson, “Eigentracking: Robust Matching and Tracking of Articulated Objects using a View-Based Representation,” *Int'l J. Computer Vision,* vol. 26, no. 1, pp. 63-84, 1998.

[9] R. Haralick, H. Joo, C.-N. Lee, X. Zhuang, V. Vaida, and M.B. Kim, “Pose Estimation from Corresponding Point Data,” *IEEE Trans. Systems, Man, and Cybernetics,* vol. 19, no. 6, pp. 1426-1446, Nov. 1989.

[10] G. Wells, C. Venaille, and C. Torras, “Vision-Based Robot Positioning Using Neural Networks,” *Image and Vision Computing,* vol. 14, pp. 715-732, 1996.

[11] A.K. Jain, *Fundamentals of Digital Image Processing,* chapter 9, pp. 342-343. Prentice Hall, 1989.

[12] D. Casasent and L. Neiberg, “Classifier and Shift-Invariant ATR Neural Networks,” *Neural Networks,* vol. 8, nos. 7-8, pp. 1117-1130, 1995.

[13] G.G. Lendaris and G.L. Stanley, “Diffraction-Pattern Sampling for Automatic Target Recognition,” *IEEE Proc.,* vol. 58, pp. 198-205, Feb. 1970.

[14] D. Casasent, L. Neiberg, and M.A. Sipe, “FST Distorted Object Representation for Classification and Pose Estimation,” *Optical Eng.,* vol. 37, no. 3, pp. 914-923, Mar. 1998.

[15] C.W. Therrien, *Decision, Estimation, and Classification: An Introduction to Pattern Recognition and Related Topics.* New York: Wiley, 1989.

[16] H. Murase and S.K. Nayar, “Visual Learning and Recognition of 3D Objects from Appearance,” *Int'l J. Computer Vision,* vol. 14, pp. 5-24, 1995.

[17] B. Moghaddam and A. Pentland, “Probabilistic Visual Learning for Object Representation,” *IEEE Trans. Pattern and Machine Intelligence,* vol. 19, no. 7, pp. 696-710, July 1997.

[18] B. Darrell, T. Moghaddam, and A.P. Pentland, “Active Face Tracking and Pose Estimation in an Interactive Room,” *Proc. 1996 IEEE Conf. Computer Vision and Pattern Recognition,* June 1996.

[19] M. Turk and A. Pentland, “Eigenfaces for Recognition,” *J. Cognitive Neuroscience,* vol. 3, no. 1, pp. 71-86, 1991.

[20] A. Pentland, B. Moghaddam, and T. Starner, “View-Based and Modular Eigenspaces for Face Recognition,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* June 1994.

[21] D. Casasent and R. Shenoy, “Feature Space Trajectory for Distorted-Object Classification and Pose Estimation in Synthetic Aperture Radar,” *Optical Eng.,* vol. 36, no. 10, pp. 2719-2728, Oct. 1997.

[22] H. Murase and S.K. Nayar, “Detection of 3D Objects in Cluttered Scenes Using Hierarchical Eigenspace,” *Pattern Recognition Letters,* vol. 18, no. 4, pp. 375-384, 1997.

[23] H. Murase and S.K. Nayar, “Illumination Planning for Object Recognition Using Parametric Eigenspaces,” *IEEE Trans. Pattern and Machine Intelligence,* vol. 16, no. 12, pp. 1219-1227, Dec. 1994.

[24] T. Arbel and F.P. Ferrie, “Viewpoint Selection by Navigation through Entropy Maps,” *Proc. Seventh Int'l Conf. Computer Vision,* pp. 248-254, 1999.

[25] L. Neilberg, “Feature Space Trajectory Pattern Classifier,” PhD thesis, Dept. of Electrical and Computer Eng., Carnegie Mellon Univ., 1996.

[26] M.A. Sipe and D. Casasent, “Global Feature Space Neural Network for Active Computer Vision,” *Neural Computation and Applications,* vol. 7, no. 3, pp. 195-215, 1998.

[27] A. Papoulis, *Probability, Random Variables, and Stochastic Processes.* New York: McGraw-Hill, 1984.

[28] M.A. Sipe and D. Casasent, “Active Object Recognition Using Appearance-Based Representations Derived from Solid Geometric Models,” *Proc. Symp. Intelligent Systems and Advanced Manufacturing,* vol. 3522, pp. 139-150, Nov. 1998.

[29] M.A. Sipe and D. Casasent, “FST-Based Active Object Recognition for Automated Assembly,” *Proc. SPIE Intelligent Robots and Computer Vision XVIII, Algorithms, Techniques, and Active Vision,* vol. 3837, pp. 2-13, Nov. 1999.

Michael A. Sipe received the BS degree and MS degree in electrical engineering from Virginia Polytechnic Institute and State University in 1987 and 1992, respectively, and the PhD degree in electrical and computer engineering from Carnegie Mellon University in 1999. From 1987 until 1994, he was employed with International Business Machines Inc., where he designed and developed computer vision and imaging systems that have been deployed worldwide. He is currently the Systems Integration manager at Cellomics Inc. His work there entails automating the collection and analysis of cellular biology data. His research interests include computer vision, pattern recognition, and machine learning. He is a member of the IEEE and IEEE Computer Society.

David Casasent received the PhD degree in electrical engineering from the University of Illinois in 1969. He is a full professor at Carnegie Mellon University, Pittsburgh, Pennsylvania, in the Department of Electrical and Computer Engineering, where he is the George Westinghouse Professor and director of the Laboratory for Optical Data Processing. He is a fellow of the IEEE, OSA, and SPIE and has received various best paper awards and other honors. He is the author of two books, editor of one text, editor of 60 journal and conference volumes, and contributor to chapters in 20 books and more than 700 technical publications on various aspects of optical data processing, image pattern recognition, and real-time signal processing. His research interests include distortion-invariant pattern recognition, neural networks, Gabor and wavelet transforms, robotics, morphological image processing, and product inspection. He is a past president of INNS (the International Neural Network Society) and of SPIE (the International Optical Engineering Society).