HIERARCHICAL POSE ESTIMATION FROM RANGE DATA FOR SPACE APPLICATIONS

Louis Simard

Department of Electrical and Computer Engineering

McGill University, Montréal

August 2002

A Thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Engineering

Abstract

An emerging application for computer vision systems is satellite servicing, involving pose estimation and tracking. Trackers are available but they often require initialization, considerably reducing autonomy. This thesis presents a new hierarchical pose estimation technique based on view-based analysis. The method can handle very sparse range data, is computationally efficient, is robust to noise, and can handle virtually any type of range data. It partitions the problem into two halves, one dealing with estimation of the translation and the other with the orientation. This greatly reduces the complexity of the overall problem without compromising the accuracy of the solution. The resulting algorithm is able to determine pose to within a prescribed accuracy, and from any vantage point within the sensor field of view, at minimal computational complexity for large variations in image noise. Results showing the performance of the system on a prototype space vision system are presented.

Résumé

L'entretien de satellites est une application des plus prometteuses pour les systèmes de vision artificielle. Plusieurs traceurs sont disponibles mais requièrent toutefois une initialisation, ce qui réduit considérablement l'autonomie. Une nouvelle méthode hiérarchique d'estimation de pose basée sur une analyse visuelle est présentée dans cette thèse. Cette méthode peut traiter des données télémétriques très éparses, est efficace au niveau des calculs, robuste au bruit et peut traiter des données télémétriques de tout genre. Elle partitionne le problème en deux étapes, soit l'estimation de la composante de translation et l'estimation de la composante de rotation. La complexité du problème se voit ainsi grandement réduite tout en conservant la précision sur la solution finale. L'algorithme engendré est capable de déterminer la pose à la précision requise à partir de tout point faisant parti du champ de vision du senseur et ce, dans une complexité minimale des calculs et pour de larges variations en bruit. Des résultats démontrant la performance du système sur un prototype de système de vision spatial sont présentés.

Acknowledgments

First and foremost, I would like to thank my supervisor at the Canadian Space Agency, Denis Laurin, and my supervisor at McGill University, Frank Ferrie, for their support throughout my work. Thanks also to Gilbert Soucy for having facilitated the preliminary contacts with the Canadian Space Agency.

Special thanks to my brother, Philippe, for his great help. As usual, his insight into my work has always surpassed usefulness. Thanks also to the people in the Artificial Perception Laboratory for having provided advices along this project. Thanks to Frédéric Thomas-Dupuis for having drawn my attention on the fundamental aspects of graduate studies.

Heart-felt thanks to my parents, whose encouragement throughout my degree was greatly appreciated. Finally, I would like to thank Stéphanie for all the qualities that make her my “source intarissable d’inspiration”.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

With the advent of fast, reliable 3-D imaging systems capable of acquiring scenes at video rates, the application of range imagery to space robotics has become a viable technology, particularly in satellite servicing. The principal task of a vision system in the latter context is tracking the 6 degree-of-freedom (d.o.f.) pose of an object at rates sufficient for robotic control, generally on the order of 10 Hz or greater. Modern tracking algorithms generally have little difficulty, per se, in achieving such rates as frame-to-frame coherency limits the computational complexity of determining correspondence. However, the situation can become difficult when the coherence assumption is violated, e.g., occlusions by other objects in the scene, acquisition failures by the sensor system, or sudden accelerations beyond the sampling rate of the system. Furthermore, initial correspondence needs to be established in the first place, which often requires the intervention of a human operator (e.g. [13]). All of these are typical of space environments and must be dealt within an operational system.

1. Overview of the Problem

The main issue with most methods used by space tracking systems is that an initial estimate of rotation and translation is required, meaning that they cannot re-establish

tracking upon loss nor initiate automatically. An example is the iterative closest point algorithm (ICP) [3] which consists of aligning a model of an object with the acquired data so as to minimize the distance between the two data sets. While ICP is computationally intensive and usually limits its applications to non-real-time tasks, Jasiobedzki *et al.* [14] have developed an extension which makes it a suitable method for space operations. But while it provides speed and robustness to occlusion, it still requires an initial guess.

As part of its space robotics program, the Canadian Space Agency (CSA) has funded the development of a number of rangefinding technologies, varying from high speed, small volume (12 frames/sec in a 1 m^3 field of view) to relatively large volumes at low sampling rates (1 frame/sec in a 1000 m^3 field of view). For the purposes of the research reported here, the characteristics of this ensemble of devices were approximated in a laboratory setup comprised of a stereo vision system with a field of view of 45° x 30° (1000 m^3) sampled at 720 x 480 pixels. Acquisition speed was limited to 0.9 frame/sec, but compensated for by limiting object velocity accordingly. Conditions were controlled to provide a reasonable facsimile of the expected operating environment. The ICP-based algorithm described above is capable of tracking pose at frame rates using stereo edge features as input [13]. The system is fairly robust, tolerating error in pose of up to 15° in orientation and 150 mm in displacement. Hence, the algorithm used to determine initial pose or to re-establish tracking lock upon loss must do so within this prescribed bound[1] and must be capable of dealing with sparse range data.

[1] With the advantage that a more precise determination of parameters can result in faster convergence of the tracking algorithm.

2. Previous Work

Object recognition generally addresses two problems [20]:

1. Identification: Find the identity of the 3-D object.

2. Localization: Find the pose (position and orientation) of the 3-D object.

While the literature generally groups these two problems together, the focus here is on location rather than identification. Indeed, in space operations, the identification problem is usually avoided as the identity of the object of interest is already known. Furthermore, CAD models of space structures are generally available as all objects are man-made. Only the problem of pose estimation thus remains.

Formally, the problem of pose estimation is defined as determining the set of rotations and translations that map the data acquired in a given view into an object-centered frame of reference defined by a corresponding CAD model,

$$\Omega = (\alpha, \beta, \gamma), \mathrm{T} = (x, y, z), \tag{1}$$

where Ω is the set of 3 Euler angles defining rotations about the **x**, **y**, and **z** axes respectively, and T is the translation vector from the view-centered frame to the object-centered frame (Figure 1.1).

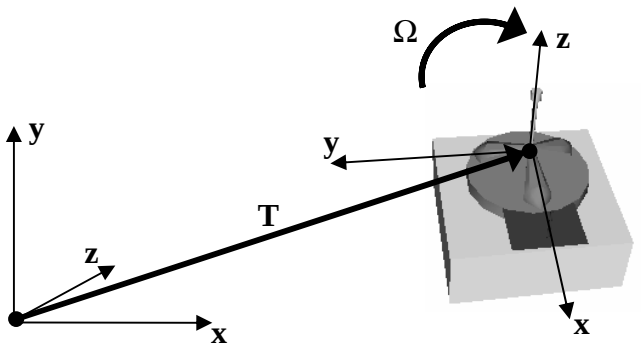


Figure 1.1 - Rotation Ω and translation T from the view-centered frame to the object-centered frame

Although the problem of estimating (Ω, T) is well-studied in the literature, the solution to a particular problem is often operationally constrained. Specifically, the work reported here concerns a satellite tracking system being developed by the CSA. The subject is thus restricted to tracking of range sequences in general, and space applications in particular.

The literature contains many references to pose estimation in space. An example is the work proposed by Cropp *et al.* [9]. They suggest a method of estimating the relative position and orientation of a known target satellite using only passive imagery. Their technique consists of detecting lines in an input image and then matching them to a model. The rotation and the translation that minimize the least-squares line-matching error are then computed. The process is repeated iteratively until convergence. Since this method makes use of intensity images, it is not suitable for the problem considered in this thesis. Indeed, the algorithm should make use of range sequences.

Why use *range* sequences, especially with the success of intensity-based methods

such as Lowe's work using features based on the SIFT [16]? The answer is two-fold. First, rangefinding systems based on active illumination sources can be made largely invariant to the kinds of wide shifts in lighting conditions that typify a space environment (from sunlight to shadow). This variation in lighting has been cited consistently as one of the major difficulties in the design of space vision systems [14,15]. Second, perhaps the most significant advantage of using range data is a calibrated field of view, greatly facilitating figure/ground separation (via depth windowing) and minimizing the complexity of estimating pose parameters (position can be localized from a single viewpoint).

There exist two basic approaches to determine the pose of an object from range data: the feature-based approach and the appearance-based (or view-based) approach. Feature-based methods represent 3-D objects through the type of features and their spatial relations. Several kinds of features can be used [4,12], including corners and edges. Locating an object using this approach then means matching image features with model features, plugging their positions in the projection equations, and solving for the position and orientation of the object. The advantage of these approaches is that they generate compact object descriptors, offer some robustness against occlusion, and some invariance against illumination and pose variations. However, it has long been known that the localization of such features is a non-trivial problem (e.g. [2]), especially when dealing with sparse range data. Furthermore, such techniques assume knowledge of corresponding pairs of model and data features. The applicability is thus limited to objects comprised of geometric features that are easy to both model and extract. These

assumptions are therefore far too restrictive for the kinds of applications considered here.

Appearance-based methods, on the other hand, are attractive to these applications since they can take advantage of a potentially large ensemble of data samples, yielding better stability of pose estimates relative to landmark features. Indeed, in appearance-based systems, an object is represented as a set of its possible appearances. In other words, intensity images of the object are acquired under different poses and illumination directions. To facilitate comparison during matching, the system encodes individual images as points in a multidimensional space. Techniques such as principle components analysis (PCA) can be used to compress this space into a lower dimensionality based on a statistical analysis of the set of training images. Acquired images are then projected onto the resulting subspace and are matched to the closest point. Basically, the Euclidean distance in this space is equivalent to image correlation [20]. Appearance-based methods can be applied to range images as well. Campbell and Flynn proposed in [8] the use of 2½D data to capture the shape appearance of a view, which makes this method suitable for the problem considered here.

However, the major drawback of appearance-based methods is their lack of ability to handle more than one object in the scene with the possibility of occlusion. Furthermore, they require the object to be segmented from its background [7]. While these problems have been extensively studied for intensity images, less has been done with respect to range images. Although the proposed methods do indeed perform pose estimation (in addition to recognition), the object pose is parameterized by only one or

two d.o.f.. The work reported here addresses the full 6 d.o.f. pose estimation problem.

Consideration of how to solve this problem given the constraints outlined thus far, leads to the principal contribution of this paper, a novel pose estimation technique using a view-based method. As will be demonstrated, the resulting algorithm is reasonably robust and capable of dealing with sparse range data. Since real-time issues are also considered, a novel hierarchical method is proposed to avoid the time-consuming process of matching.

3. Contributions

The contributions of this thesis consist of the following:

- The development of a new pose estimation technique which offers
    - A partitioning of the problem into two halves (translation and orientation)
    - A hierarchical determination of the orientation
    - Robustness to sparse range data
    - Real-time execution
    - Sensor independence

- The implementation and evaluation of the algorithm, including testing on synthetic and real data
- The development of a complete system, part of a deliverable to the Canadian Space Agency
- The integration of the system to an actual space vision system.

4. Outline of Thesis

Chapter 2 is concerned with an analysis of the constraints that can be exploited to partition the problem into two halves, one dealing with estimation of T and the other with Ω. This greatly reduces the complexity of the overall problem without compromising the accuracy of the solution. Section 3 describes a hierarchical view-based method for estimating Ω given T by extending the technique initially proposed by Skočaj and Leonardis [19] using the robust norm of Black and Jepson in [6]. In Chapter 4, it is shown that the resulting algorithm is able to determine pose to within a prescribed accuracy, and from any vantage point within the sensor field of view, at minimal computational complexity for large variations in image noise. Finally, Chapter 5 comprises observations on the current method and intended future work.

CHAPTER 2

Reducing Complexity by Partition

As mentioned in Chapter 1, view-based methods are attractive for space applications since they can take advantage of a potentially large ensemble of data samples, yielding better stability of pose estimates relative to landmark features. However, the difficulty here is the rather large view space that can arise for a 6 d.o.f. parameterization. For example, a brute force approach to representing an object in a volume of 1000 m^3 at a resolution of 150 mm in position and 15° in orientation would imply a view space of over 2 billion images ($1000m^3/0.15^3m^3$ x $360^o/15^o$ x $360^o/15^o$ x $180^o/15^o$). Edwards presents in [11] an appearance-based technique that reduces the problem from six d.o.f. to three. However, the assumptions he makes about the scene are quite limiting. Indeed he assumes a non-occluded object set against a uniform background and ideal illumination conditions.

Fortunately, complexity can be minimized by partitioning the estimation problem in terms of T and Ω respectively. Knowledge of the approximate size and extent of an object is sufficient to achieve rudimentary figure/ground separation and a biased estimate of T, which is denoted by $\hat{T}$. The view-based method used to estimate Ω subsequently can account for this bias in the training process, making it possible to recover an accurate estimate of rotation despite having a biased estimate of T. From here, it is relatively

straightforward to recover the correct value of T.

1. Figure/Ground Separation

Figure/ground separation is the process in which object features are separated from background/noise features. A glance at the range map shown in Figure 2.1 demonstrates the practical difficulty of figure/ground separation (the grapple fixture's location is shown by the rectangular window superimposed on the image). Indeed, while it is generally assumed that an object in space should not be surrounded by other objects, a range map usually contains several background elements. Furthermore, real scenes are often corrupted by noise, and there is no guarantee that they will be uniformly sampled (e.g. data dropouts).

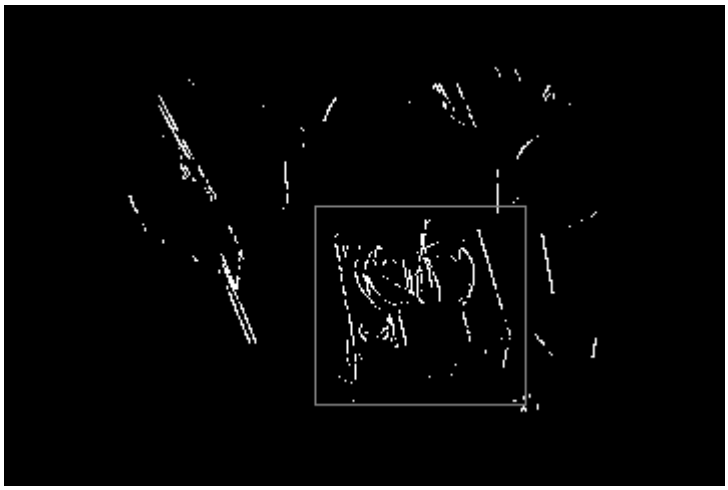


Figure 2.1 - Sparse range data of a grapple fixture with background features

In the literature, the problem of figure/ground separation is generally treated in the context of 2-D images. The goal is then to find which pixels belong to the foreground,

and which pixels belong to the background. Zhang et al. explain in [23] the reasons why this is a difficult problem. First, there are many types of feature that may be present in a scene. Second, in order to determine whether a feature is figure or ground, the spatial/temporal/phasic distribution of other related features must be determined.

Some of these problems can be avoided when dealing with range data. In fact, intensity images are limited in their use as pixel values are related to surface geometry through the illumination conditions, the optical and geometrical properties of the surface, and the viewer position [20]. One of the advantages of a calibrated field of view, on the other hand, is that it is possible to create volumetric templates that are crafted to particular objects. The figure/ground separation then becomes a matter of template matching [20]. Various types of volumetric templates can be used. In the case of the grapple fixture shown in Figure 2.2, the spatial extent of the object is reasonably well approximated by its comprising sphere shown in Figure 2.3.

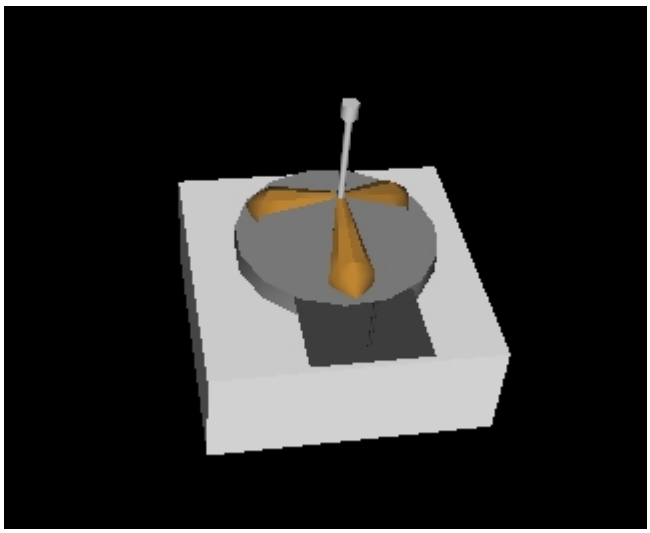


Figure 2.2 - Grapple fixture

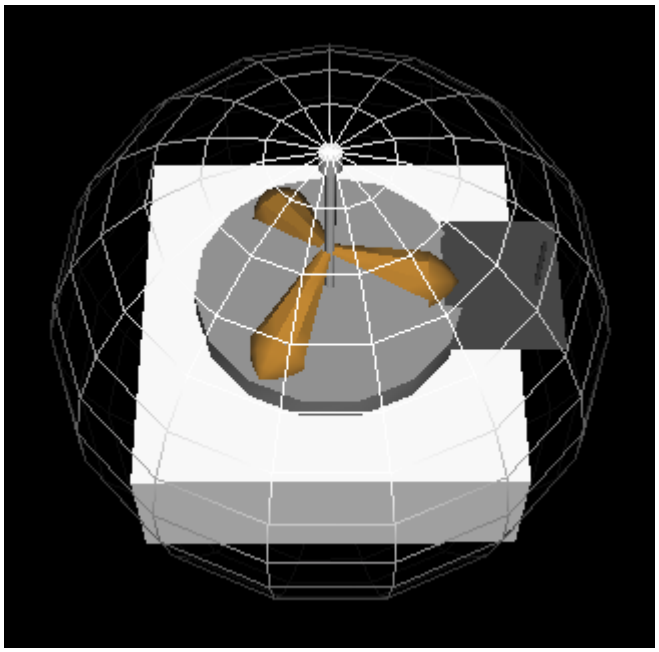


Figure 2.3 - Comprising sphere of the grapple fixture

Figure/ground separation proceeds via template matching that first localizes point clusters with the appropriate shape and/or volume characteristics. One of the advantages of template matching is that it is possible to achieve speedup by judicious sub-sampling without seriously compromising accuracy of localization. Coarse to fine strategies are also possible using a location determined at a coarser scale to initiate a search at a finer one [20]. Since it is also possible that a given template might include data from nearby objects, a clustering algorithm is used to separate patches isolated by gaps, i.e. C^0 discontinuities. By construction, the dominant (largest) patch will always correspond to the surfaces comprising the object.

2. Estimation of T

Direct estimation of T is complicated by the fact that only part of the object is visible from a given viewpoint, i.e. some surfaces are always self-occluded. At best a biased estimate, $\hat{T}$, may be obtained, which is offset from T by some constant, Δ. In general, the offset Δ will depend on the particular viewpoint from which the data are acquired. Any metric can be used for $\hat{T}$ provided that it is stable with respect to viewpoint. Two metrics were empirically investigated. The first one was the centroid of the dominant patch resulting from figure/ground separation. However, sparse range data do not necessarily provide an even distribution of the 3-D data points over the object. Important variations were thus observed with respect to the viewpoint. The second was the center of the minimum bounding box enclosing the dominant patch resulting from figure/ground separation. It was determined that it does indeed provide consistent, stable estimates. Formally then, $\hat{T}$ is defined to be the location of the bounding box center in range image coordinates.

To facilitate the matching of appearances, precisely the same metric is used to define the local origin of each of the training views. For each view Ω_i, an associated translation Δ^{Ω_i} to the canonical origin of the object is defined. Hence, given data acquired from an arbitrary viewpoint, it is straightforward to determine the local origin of the corresponding training view as well as to normalize coordinates[2]. What remains is to

[2] And scale, implicitly.

solve the problem of matching appearances which ultimately determines T through Δ^{Ω_i}, and Ω implicitly by the corresponding view which is indexed by rotation.

CHAPTER 3

Orientation Estimation

The success of appearance-based schemes [21,18] is largely dependent on the degree to which the input can be normalized so that it falls within the manifold of similar appearances. Campbell and Flynn [8] were among the first to recognize that range data could be exploited nicely in this context, e.g., using eigensurfaces to capture the rough shape of an object. These eigensurfaces are used in much the same way as eigenpictures to form an eigenspace. A similar approach is taken here using a CAD model to generate a sequence of views in normalized position, indexed by the 3 rotation parameters comprising Ω. This process can include an explicit model of the particular sensor used to acquire the data on-line, so that the appearance manifold more closely resembles the data. An example of simulated data of a stereo vision system is shown in Figure 3.1a.

a) b)

Figure 3.1 - a) Simulated data of a stereo vision system and b) full range data

Unfortunately this approach fails to account for the kinds of sensor variations that can occur in practice, and which often confound standard methods. Examples include so called "missing data" arising from the sparse coverage of passive sensors (e.g. stereo) or return signal dropouts in active sensors (e.g. LIDAR). The effect is analogous to occlusions in intensity-based methods. Another problem typical of range data from active sensors is the presence of strong outliers in the vicinity of occluding contours. This is further exacerbated in space environments due to strong variations in surface reflectance (e.g. reflective blankets used to minimize thermal variation) and ambient illumination.

The "missing data" problem can be handled by training on full range images synthesized from CAD models (an example is shown in Figure 3.1b), and using the approach of Skočaj and Leonardis [19] to find the best fit to the points projected on the eigenspace. Outliers are handled by incorporating a robust norm along the lines of Black and Jepson [6].

1. Training Phase

The set of range images used for training are generated from a CAD model according to a tessellated viewsphere centered at the canonical origin of the object. Tessellations are determined according to the prescribed resolution of the rotation parameters Ω. Further, to account for how the range data are normalized with respect to position (Chapter 2), each image is translated to a new origin determined by the metric for $\hat{T}$. Hence, for a particular training view *i*, the corresponding offset is simply, $\Delta^{\Omega_i} = -\hat{T}_i$.

Standard PCA techniques [18] are used to compute an approximation of the training set to facilitate matching and thus implicitly determine the corresponding Ω for the acquired data. Let each range image correspond to an *n*-dimensional vector in an *m*-dimensional set, $\mathbf{x}^i \in \{x_1^i, ..., x_n^i\}$, where the index $i = 1, \ldots, m$ corresponds to a particular facet of the viewsphere. Each facet in turn corresponds to a unique triple of rotation parameters, Ω_i. It is generally assumed that $m << n$. An approximation to $\mathbf{x}^i$,

$$\tilde{\mathbf{x}}^i = \overline{\mathbf{x}} + \left(\mathbf{a}^i\right)^{\mathrm{T}} \left[\mathbf{e}^1, ..., \mathbf{e}^p\right]^{\mathrm{T}}, \tag{2}$$

is obtained in standard fashion, where $\overline{\mathbf{x}} = \{\overline{\mathbf{x}}_1, ..., \overline{\mathbf{x}}_n\}$, such that

$$\overline{\mathbf{x}}_i = \frac{1}{n} \sum_{i=1}^{n} x_i \ . \tag{3}$$

The basis vectors $\mathbf{e}^j$ correspond to the *p* eigenvectors of the covariance matrix $\mathbf{c} = | c_{ij} |$ with the largest eigenvalues, where

$$\mathbf{c}_{i,j} = \sum_{i=1}^{n} (\mathbf{x}^i - \overline{\mathbf{x}}) \, (\mathbf{x}^j - \overline{\mathbf{x}})^{\mathrm{T}} . \tag{4}$$

Finally, the coefficients for a particular range image, $\mathbf{x}^i$, are given by the inner product

$$a_j = \left\langle \mathbf{e}^j, (\mathbf{x}^i - \overline{\mathbf{x}}) \right\rangle, \ j = 1, ..., p \ . \tag{5}$$

2. Matching Phase

The range images used in training are quite different from the output range images expected from the sensor (compare Figure 2.1 with Figure 3.1b). Thus, sensor range images cannot be simply projected onto the eigenspace using (5) since missing pixels would not be taken into account. Skočaj and Leonardis studied the impact of missing pixels in [19]. To overcome this problem, only the set of non-zero pixels should thus be considered. Let **r** be a vector such that

$$r_j = \begin{Bmatrix} 0 & \text{if } x_j = 0 \\ 1 & \text{otherwise} \end{Bmatrix}. \tag{6}$$

Instead of determining $\mathbf{a}^i$ as in (5), the process used is along the lines of Skočaj and Leonardis, and the computation of the projection coefficients $\mathbf{a}_r^i$ which account for the non-zero pixels is done by minimizing the following norm,

$$E\left(\mathbf{a}_\mathbf{r}^i\right) = \left\| \mathbf{r}\left[\mathbf{x}^i - \bar{\mathbf{x}} + \left(\mathbf{a}_\mathbf{r}^i\right)^T \left[\mathbf{e}^1, \ldots, \mathbf{e}^p\right]^T\right] \right\|. \tag{7}$$

Black and Jepson propose a similar approach in [6] with intensity images, but note that L_2 norms are sensitive to outliers, which is precisely the case here. In a similar fashion, the quadratic norm in (7) is replaced by a robust error norm, yielding the following functional,

$$E(\mathbf{a}_\mathbf{r}^i) = \sum_{j=1}^{n} \rho\left(r_j(x_j^i - \bar{x}_j + \langle \mathbf{a}_\mathbf{r}^i, \mathbf{e}^j \rangle), \sigma\right), \tag{8}$$

where

$$\rho(u,\sigma) = \frac{u^2}{\sigma + u^2}, \tag{9}$$

and σ is a scale parameter that determines where the influence of outliers begins to decrease. Minimizing this non-linear function in the absence of priors can be computationally expensive due to the presence of multiple local minima.

For compact training sets, e.g., for i on the order of several hundred vectors, a reasonable approach is to minimize (8) by exhaustively searching the discrete space of coefficients, $\{\mathbf{a}^i\}$, determined by projecting each training image onto the eigenspace via (5). Doing this removes the need of finding a good seed point and the risk of falling into local minima, making the approach a global process. This is generally sufficient to localize the training view closest to the acquired data, which in turn is sufficient to re-establish tracking lock if the viewsphere corresponding to Ω is tessellated finely enough.

3. Reconstruction Errors

Minimizing (8) ends up to be a minimization between an observed pixel value and a reconstruction of this pixel from the eigenspace. As explained in the previous section, the reconstructed pixel is obtained by using the coefficients $\{\mathbf{a}^i\}$ determined by projecting each training image onto the eigenspace via (5). Doing this allows the reconstruction process to be performed offline, which of course has the advantage of accelerating the online process significantly.

Depending on the number of eigenvectors used, the quality of a reconstructed range image will vary. Indeed, the greater the number of eigenvectors, the better the quality of the reconstruction. But since only the first principal eigenvectors are required to represent the main features of the object, the quality of the reconstruction will be affected for some pixels of the range map. Since the reconstruction process is performed offline, statistics about the quality of the reconstruction can therefore be computed for each pixel over all training images. In other words, during the training phase, the extent to which each reconstructed range image can reliably represent each training range image is verified. An image representing the standard deviation for each reconstructed pixel is created. Then, during the matching process, if an observed pixel falls within the prescribed error bound, the inner difference of (8) is set to zero. This has the advantage of reflecting the importance of each feature. Doing so ensures that the impact of pixels for which the reconstruction error is large is reduced.

As mentioned in the previous section, the method described here is appropriate for training sets on the order of several hundred vectors. However, for training sets exceeding a certain size (e.g. a few thousand vectors), a more judicious approach has to be taken. The following section presents a novel hierarchical approach that efficiently minimizes Equation 8.

4. Matching by Hierarchy

Equation 8 is linear in the number of training images since the discrete space they form is exhaustively searched. In order to increase the efficiency while recovering the pose, it is suitable to reduce this solution space. A hierarchical approach using several viewsphere tessellations ranging from coarse-to-fine is thus taken.

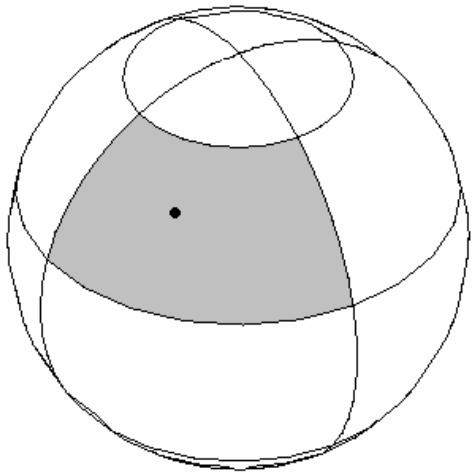
Cyr *et al.* adopt this kind of approach in [10] to determine the pose of a vertebrae spine bone. Given a 3-D model and a 2-D view, they hierarchically search the space of possible poses using a notion of similarity between the projected shape and the 2-D target shape in a fashion reminiscent of the aspect graph approach. Specifically, coarse samples of the viewsphere are matched against the target view, and the process is repeated for samples close to top matches.

While hierarchical approaches are often used in appearance-based techniques to solve the problem of scaling (e.g. [5]), they can also be used to accelerate the time-consuming matching process using coarse-to-fine strategies. The literature contains several examples of hierarchical approaches. Athitsos and Sclaroff exploit this concept in [1] in the context of 3D hand shape classification. Given an input image of a segmented hand, the most similar matches from a large database of synthetic images are retrieved. Database retrieval is done hierarchically, by first rejecting the majority of all database views and then ranking the remaining candidates in order of similarity to the input. However, considering that they make use of intensity images, their method cannot be directly applied. In fact, finding reliable similarity measures is not trivial when dealing

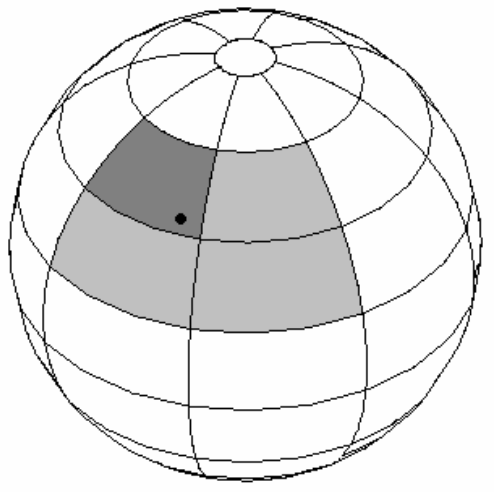
with sparse range data. Black and Jepson proposed in [6] the use of eigen-pyramids. For every image in the training set they construct a pyramid of images by sub-sampling and spatial filtering. Each level of the pyramid then corresponds to an eigenspace. However, their method is used in the context of tracking. An initial estimate is thus required to start the search. Moghaddam and Pentland present in [17] a different method in which the search is performed over scale by constructing multiple input images at various scales and searching over all of them simultaneously. In a similar fashion, Yoshimura and Kanade use hierarchical matching in [22] to determine the angle of rotation of a 2D template. Like the other methods, they build an image pyramid structure with respect to the image size.

The approach described here is different from previous approaches in that the eigenspaces are built corresponding to different samplings of the viewsphere (not image size), from coarse to fine. In other words, blurring is achieved through the averaging provided by eigenspace compression. The process is illustrated in Figure 3.2 where three levels of hierarchy are shown. An eigenspace is associated with each level and the black dot represents the orientation to be found. The process goes as follows. In the first step, the orientation is determined by minimizing (8) over the discrete eigenspace formed by the coarsely sampled viewsphere of the first level. Once the corresponding tile is found (shown in light gray), the search over the second eigenspace is performed but this time, narrowed down to the finer tiles comprised by this tile. The same process is applied up to the third and finer level to determine the final orientation (shown in dark gray). Doing so ensures that an exhaustive search over the third viewsphere is avoided while keeping the same degree of precision. Note that the tessellation factor does not necessarily need to be

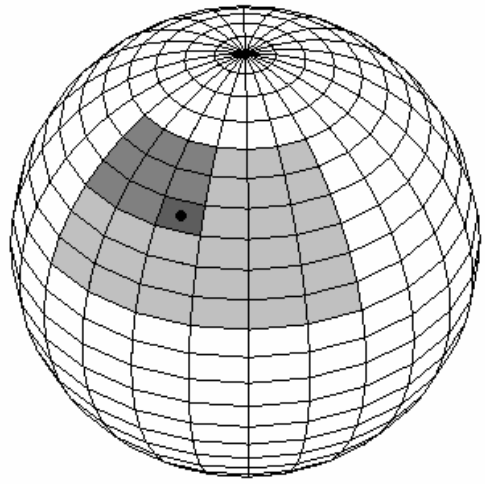
constant between levels, as shown in Figure 3.2. In this case, $L_1 = 2*L_2$, and $L_2 = 3*L_3$, where L_1, L_2, and L_3 are the tessellation factors for each level.



a) First level in the hierarchy



b) Second level in the hierarchy



c) Third level in the hierarchy

Figure 3.2 - Hierarchical approach

Attempting to further refine Ω by using the minimal $\mathbf{a}^i$ as the seed point for subsequent minimization of (8) is unlikely to garner further improvement unless some care is taken to condition the underlying manifold. Specifically, this means sampling Ω so that the views corresponding to adjacent coordinates can largely be linearly

interpolated. As it turns out, for small angles the rotation matrix associated with Ω can be reasonably well approximated by a linear map. It is straightforward to determine analytically the range of Ω for a particular error bound.

CHAPTER 4

Methodology and Results

The performance of the algorithm was evaluated. Tests were conducted using synthetic range data as well as real range images. The following sections describe the experimental setup and the results.

1. Implementation

The technique was implemented in Visual C++ on a Pentium III 750 MHz with 512 Mb of RAM running Windows XP. The graphics engine was an NVidia GeForce 256 with OpenGL support. The Intel OpenCV library was used to implement the standard PCA techniques.

The software, called *Pose Estimator*, was part of a deliverable to the Canadian Space Agency. Figure 4.1 shows a screenshot of the application. Two modes of operation are available: the *training* mode and the *pose estimation* mode. The training mode is used to generate range images of the model and to produce one or many eigenspaces (depending on whether the search method is hierarchical or not). The pose estimation mode is used to determine the pose of the object for which the system was trained for.

The application consists of three views. The *input* view (top window) is used to

display the original data as well as to show the model where the system finds it. The *model* view (bottom-left window) is used to generate the training range data from the model of the object. Finally, the *output* view (bottom-right window) is used to generate range images from the input view or the model view. The location of range data corresponding to this view is normalized following the method described in Chapter 2, i.e., by using the center of the bounding box. Furthermore, since 2½D images are generated (i.e. each pixel corresponds to a depth value), the view is also used to down-sample the data coming from the input view. It should be mentioned that the application can also determine the pose from synthetic data generated from the model.

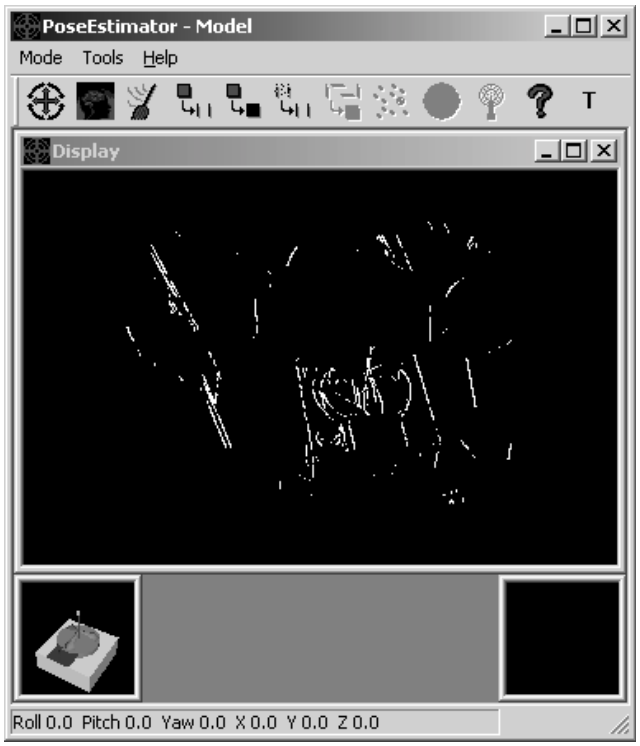


Figure 4.1 - Screenshot of Pose Estimator

Other features are also available. For example, the application can communicate

via sockets with the stereo vision server, allowing remote operations. The user can thus take input data from the server and transmit the pose estimate back to it. Pose Estimator can also operate automatically by checking the confidence level provided by the server. If the level decreases (meaning that the tracking system is not sure anymore about the pose of the object), Pose Estimator will automatically transmit a new pose estimate. The user can also choose whether the search should be done in an exhaustive or hierarchical manner.

2. Experimental Setup

The stereo vision system described in Chapter 1 is shown in Figure 4.2. It was used to acquire data of a grapple fixture of size 510 x 560 x 530 mm shown in Figure 4.3. The convergence range of the model matching is on average 15° and 30% of the observed object size, corresponding to approximately 150 mm for this particular object. The illumination conditions were similar to what could be encountered in space, i.e. a dark environment with a strong point source illumination.

Figure 4.2 - Stereo vision system

Figure 4.3 - A grapple fixture

3. Training

The system was trained on a CAD-model of the grapple fixture shown in Figure 4.3. Three hierarchy levels were used, such that three sets of training range images were required. The first level consisted of 108 range images (100x100 resolution), corresponding to a 60° increment in roll, pitch and yaw. The second level consisted in 256 images, corresponding to a 45° increment. Finally, 6912 range images were generated for the third level, corresponding to a 15° increment. This particular discretization was chosen to match the maximum initialization error tolerable by the tracking system. A total of 8 eigenvectors were used to represent the first training set, 15 eigenvectors for the second set, and 75 eigenvectors for the third one.

As the spatial extent of the object is reasonably well approximated by its comprising sphere (see Figure 2.3), a sphere of radius 380 mm was used to perform figure/ground separation. Sigma (σ) was set to 2000.

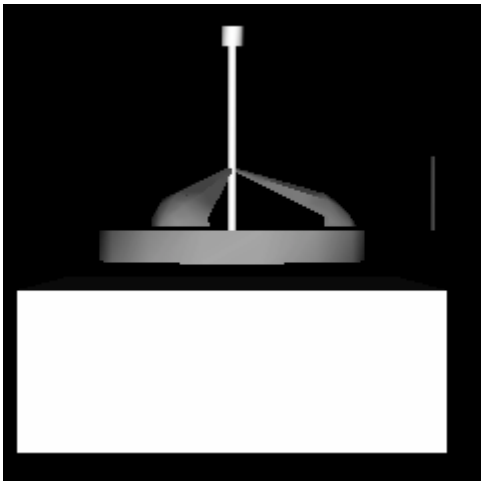


Figure 4.4 - Grapple fixture in an arbitrary pose

Examples of reconstructed range images are shown in Figure 4.5. These correspond to the orientation shown in Figure 4.4. The reconstruction was done for the three hierarchy levels using (2). Observe that the quality of the reconstruction greatly varies through the different levels. Indeed, the generalization provided by the first level is much greater than the third level. This is due to the limited number of eigenvectors used for this level and the lower correlation between each training range image.

The quality of the reconstruction also varies through the pixels of the range map. The maps shown in Figure 4.6 represent the standard reconstruction errors for each pixel for each hierarchy level. Dark regions represent pixels for which the error is high while bright regions represent pixels for which the error is small. Observe that errors are much smaller for higher levels. Indeed, this is due to the fact that the viewsphere is more finely sampled (leading to a higher correlation between range images) and that the number of eigenvectors is greater. As mentioned in the previous chapter, this greatly helps in highlighting the importance of each feature.

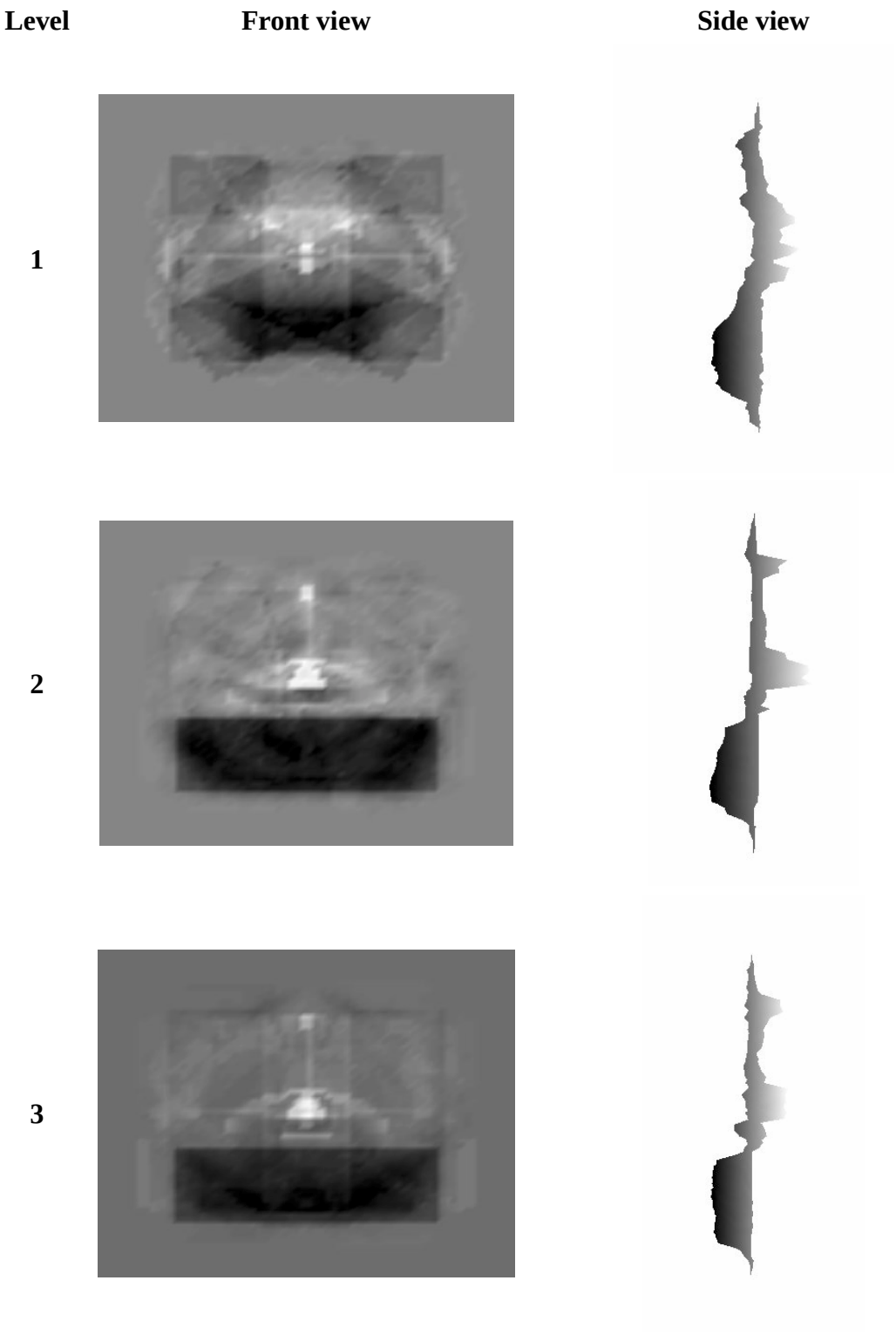
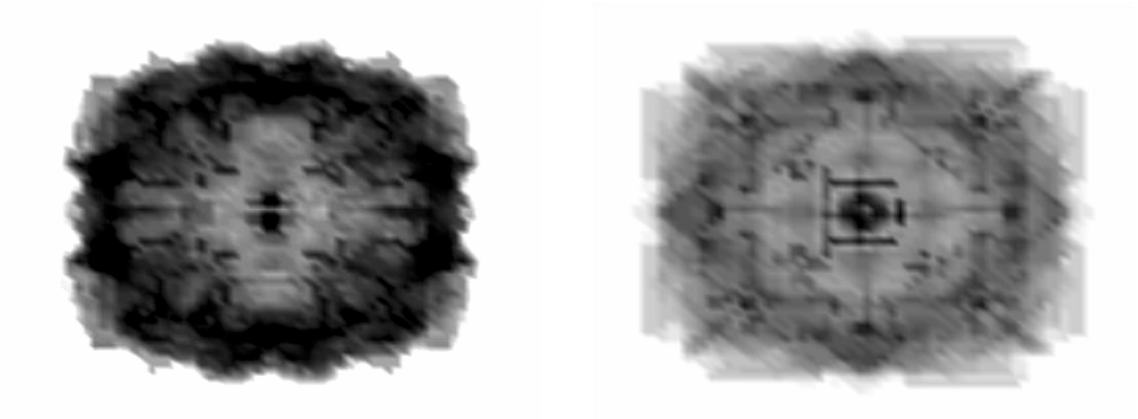
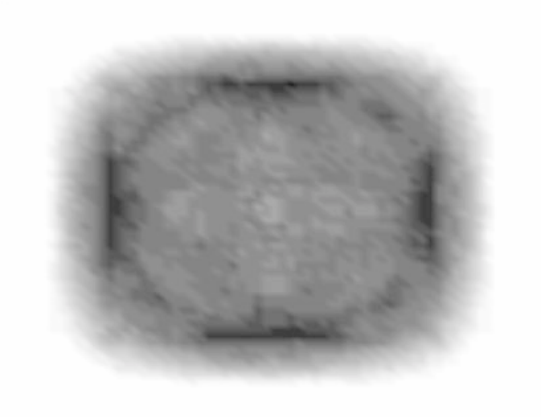


Figure 4.5 – Reconstructed range images for each hierarchy level



a) First hierarchy level

b) Second hierarchy level



c) Third hierarchy level

Figure 4.6 - Reconstruction error maps for each hierarchy level (dark regions representing pixels for which the error is high)

4. Results

Tests were performed in a lab environment where many background features are present. Different poses of the object were used, with distances ranging from 1 to 6 meters. A trial was considered successful when the estimate fit within the tracker tolerance (±15° for roll, pitch and yaw, and ±150 mm for x, y and z of the pose given by the system).

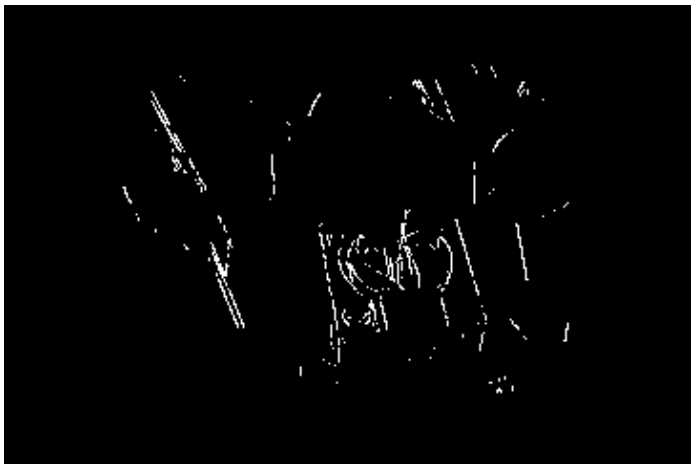
4.1 Exhaustive Matching

The first experiment consisted of retrieving the pose of the grapple fixture using exhaustive matching. In other words, every solution of the discrete space formed by the 6912 training range images was tried. Experimental results are summarized in Table 4.1.

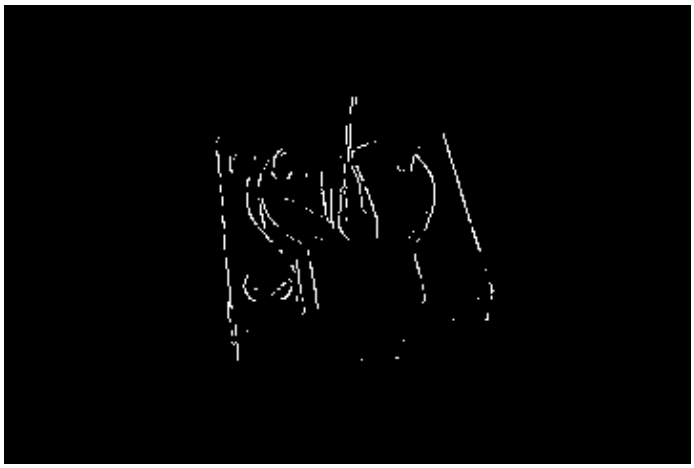
Trials	60
Success	57
Success Rate	95%
Average Time (s.)	T: 0.1, Ω: 1.2

Table 4.1 - Experimental results on a grapple fixture using exhaustive matching

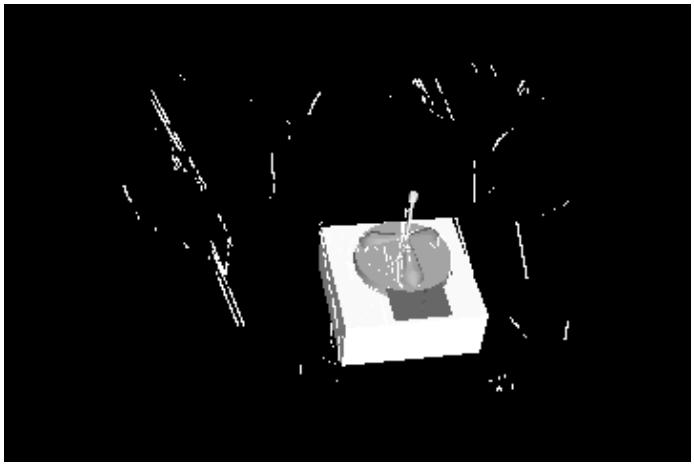
95% of the trials were successful with a total execution time (translation time T + orientation time Ω) of 1.3 seconds. An example of success is shown in Figure 4.7 where the acquired range data are shown in (a). The result of figure/ground separation is shown in (b). The model in its resulting pose is co-rendered with the original range data in (c). On average, 85% of pixels were missing in the range data map (with respect to the full training range image).



a)



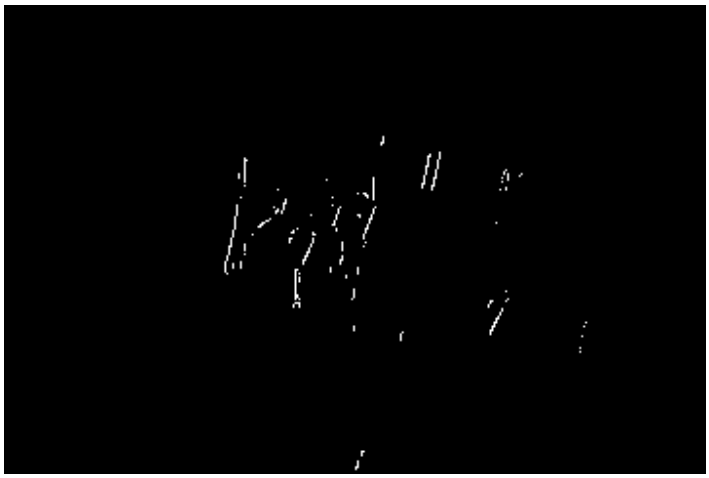
b)



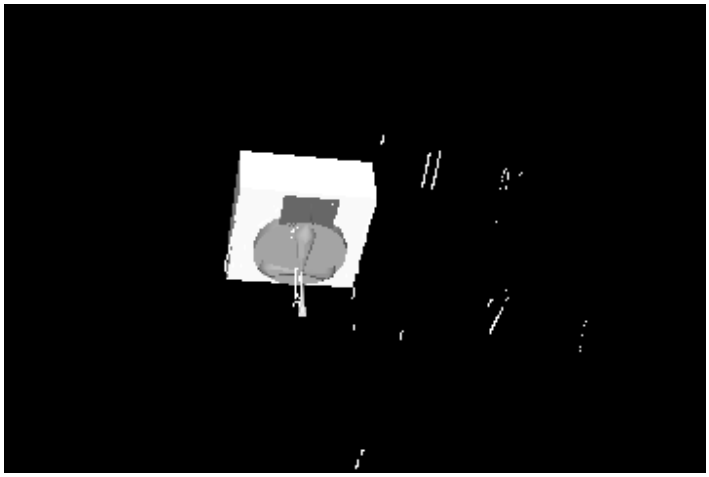
c)

Figure 4.7 - Example of successful pose estimation

Other examples of success are shown in Figure 4.8 and Figure 4.9. The acquired range data are shown in (a). The model in its resulting pose is co-rendered with the original range data in (b).



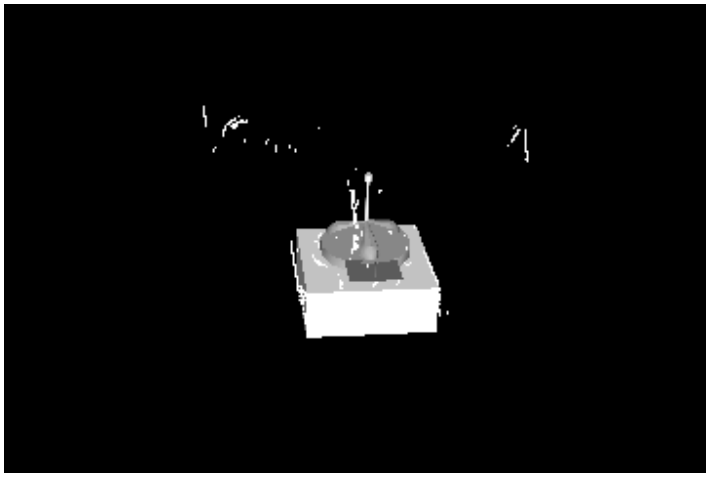
a)



b)

Figure 4.8 - Another example of successful pose estimation

a)



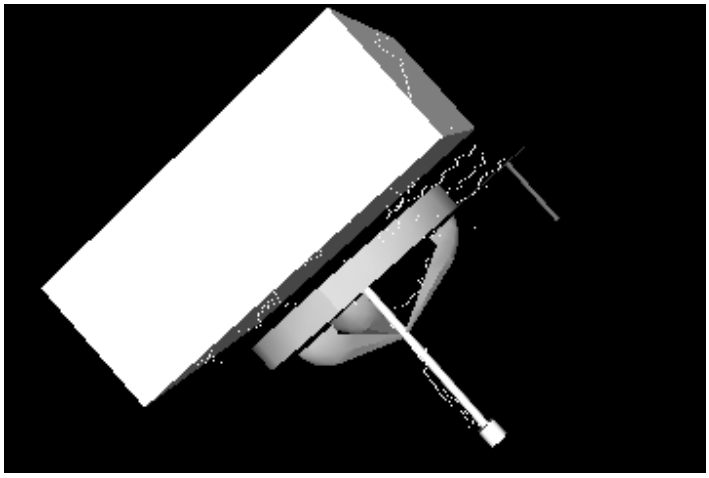
b)

Figure 4.9 - Another example of successful pose estimation

Observe that while many background features are present, the system was able to localize the object within the prescribed accuracy. These examples also demonstrate that the technique is robust to sensor noise. While range data seem to provide sharp and straight edges, a closer look reveals that this is not necessarily the case. Figure 4.10 shows the same data as in Figure 4.8, but rendered from a different viewpoint (i.e. the data set is rotated). Observe that they are quite noisy (in depth). The error is in fact approximately

30 mm at a distance of 3 meters. This can be due to several factors such as stereo mismatches and edge localization errors. In spite of this, the algorithm is able to recover pose successfully.

a)

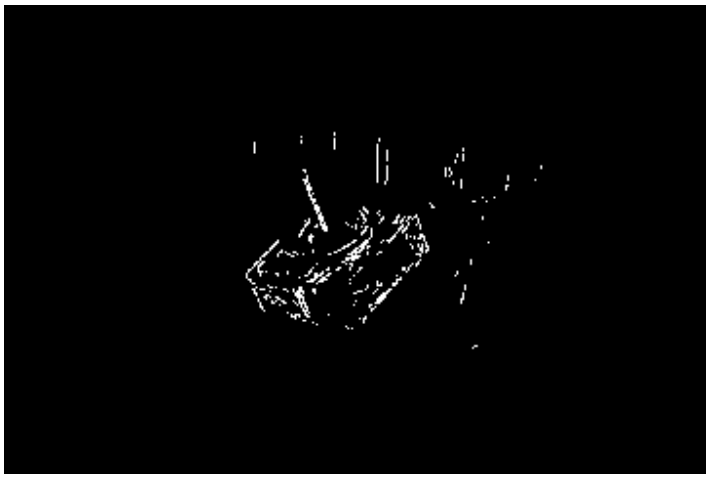


b)

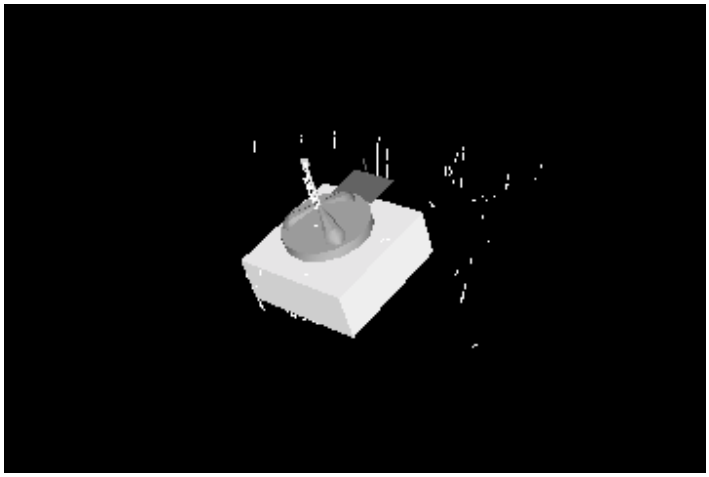
Figure 4.10 - Noise present in range data

An example of failure is shown in Figure 4.11, and corresponds largely to a 180° rotation about the rod protruding from the base. Considering the symmetry of the object, this result is not surprising since the range data do not provide sufficient information

about a distinguishing feature. In this case the top plate, which is clearly visible in Figure 4.7, is poorly sampled in Figure 4.11. In fact, all of the errors recorded corresponded to cases such as this one.



a)



b)

Figure 4.11 - Sample of non-successful pose estimation

4.2 Hierarchical Matching

The second experiment consisted of retrieving the pose of the grapple fixture using hierarchical matching. Experimental results are summarized in Table 4.2. Results

from exhaustive matching are reported here to facilitate comparison.

	Hierarchical	Exhaustive
Trials	**60**	60
Success	**55**	57
Success Rate	**92%**	95%
Average Time (s.)	**T: 0.1, Ω: 0.1**	T: 0.1, Ω: 1.2

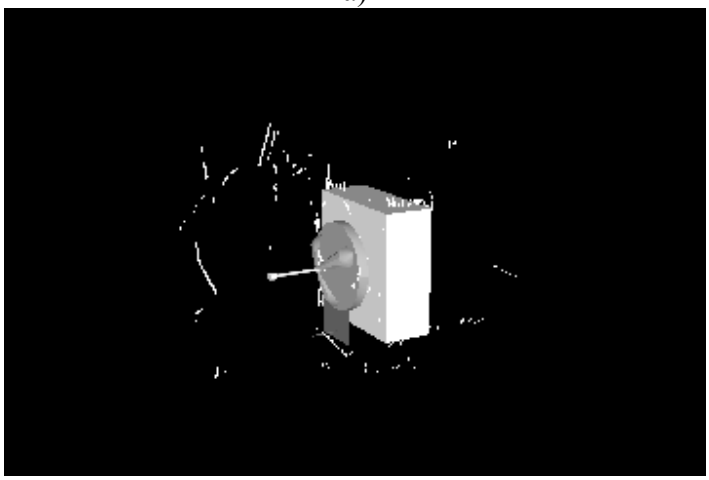
Table 4.2 - Experimental results on a grapple fixture using hierarchical matching

92% of the trials were successful with a total execution time (translation time T + orientation time Ω) of 0.2 second. Using hierarchical matching is almost an order of magnitude faster than using exhaustive matching. In fact, as outlined in the previous chapter, using a hierarchical approach allows reduction of the solution space. Considering that the size of this space is dependent on the number of hierarchy levels that are used, this number can then be chosen in order to meet the requirements of the application. Since the execution time is also dependent on the figure/ground separation, the process can be accelerated by judiciously sub-sampling the data without seriously compromising accuracy of localization.

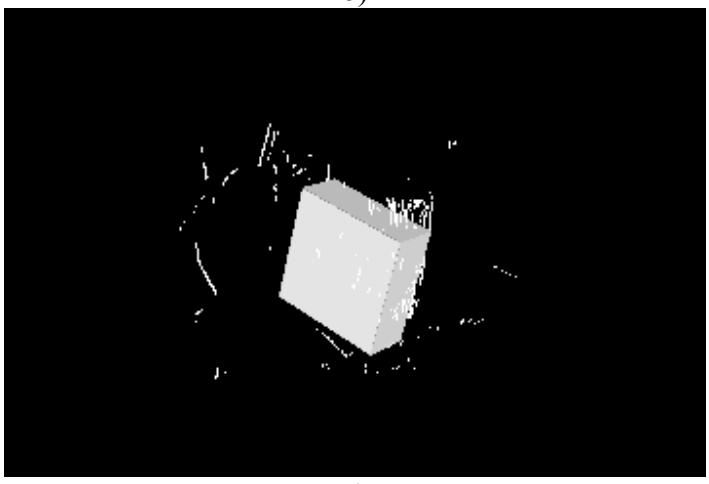
Observe however in Table 4.2 that the success rate is slightly smaller when using hierarchical matching. An example of failure is shown in Figure 4.12. The acquired range data are shown in (a). The model in its correct pose from exhaustive matching is co-rendered with the original range data in (b). The model in its erroneous pose from hierarchical matching is shown in (c). The error corresponds to a 180° rotation about the plate protruding from the base. This clearly indicates that the search was misled right at the first hierarchy level and underscores the primary weakness of the method. In cases

where objects are distinguished by features at finer scales, a hierarchical approach will return an arbitrary choice. The way around this problem is to permit multiple matches in cases where the best match is ambiguous relative to the distance metric. Multiple hypotheses would be carried forward until the ambiguities are resolved at a finer resolution. This approach would be an interesting extension to the algorithm described here. All of the errors recorded corresponded to cases such as this one and the cases explained in the previous section.

a)



b)



c)

Figure 4.12 - Example of failure for hierarchical matching

4.3 Impact of Center Localization Error

The impact of a center localization error on orientation estimation was also studied. Tests were performed on synthetic data (as shown in Figure 4.13) for 30 typical poses of the object. The experiment consisted of adding an error vector to the true view center and determining if the system was still able to correctly determine the pose.

Figure 4.13 - Simulated data of a stereo vision system

Results are shown in Figure 4.14, in which the x-axis is the induced translation error in cm and the y-axis the percentage of success. Three different non-coplanar error vectors were used, each having a length ranging from 0.0 to 12.0 cm. As can be seen, determination of Ω is fairly robust, tolerating a translation error of up to 4 cm. In practice, experimental results on real data show that view-based determination of $\hat{T}$, as outlined in Section 2.2, is less than 2.5 cm in error on average.

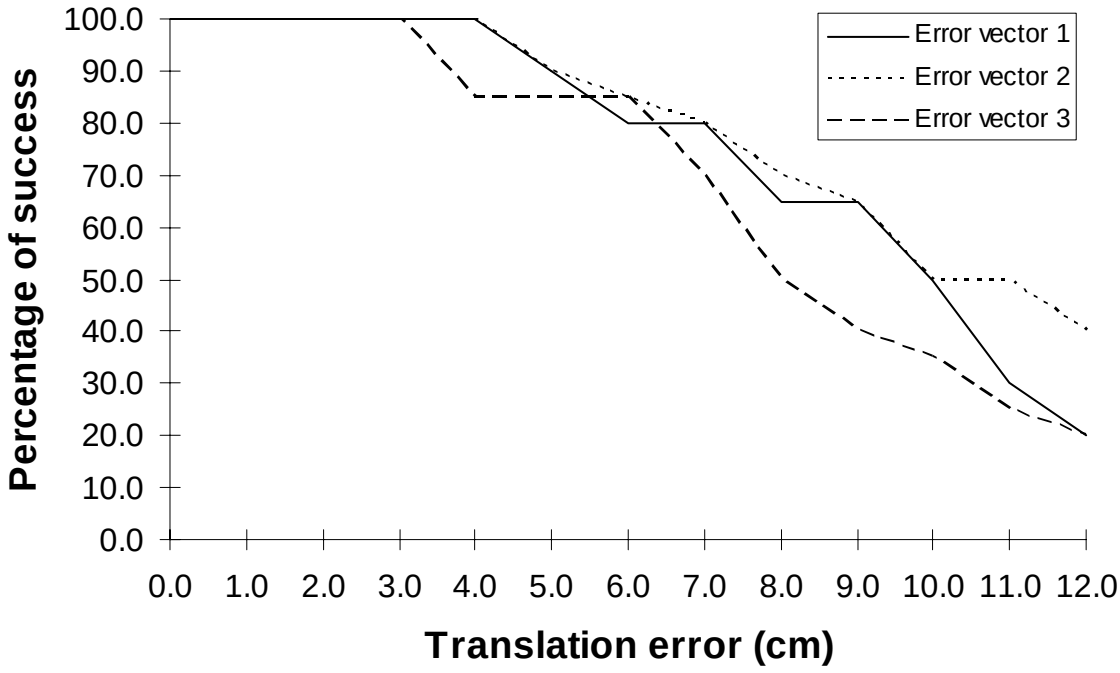


Figure 4.14 - Impact of a center error on recognizing the orientation

The most likely situation where an error in $\hat{T}$ can occur is when only a partial view of the object is available. While the issue of occlusions has been briefly addressed in previous work utilizing view-based analysis, it can hardly be treated in this framework. To some extent, error detection is possible in cases where the measure indicated by Equation 8 exceeds a prescribed bound. This can be made more definitive by applying the requisite statistical analysis to determine an appropriate confidence interval. Unfortunately, this does not get around the problem of ambiguity as demonstrated earlier in Figure 4.11. Such problems can only be resolved using additional observations.

CHAPTER 5

Conclusions

A novel pose estimation technique has been presented. The view-based algorithm partitions the problem into two halves i.e. one dealing with estimation of the translation component and the other with the orientation. As demonstrated through the results, this greatly reduces the complexity of the overall problem without compromising the accuracy of the solution. Figure/ground separation via template matching is used to retrieve the translation component. A novel hierarchical view-based method is used to determine the orientation.

As the preceding chapter demonstrates, the algorithm described in this thesis performs well at recovering pose from sparse, noisy range data. Results presented in this thesis show that the technique is able to determine the pose to within a prescribed accuracy, and from any vantage point within the sensor field of view, at minimal computational complexity for large variations in image noise. However, while the execution time was sufficient for the stereo vision system on which the experiment was conducted, it might not be enough for real-world operations. Aside from the usual speedups obtained by optimizing code and increasing processor speed, the number of hierarchy levels could be increased to reduce the solution space. While the experimental results showed that a hierarchical approach was viable, it was observed however that it

had some weaknesses. Future work could consist of permitting multiple matches at coarser levels such that ambiguities could be resolved at a finer resolution. Furthermore, in practice, even rudimentary knowledge about the object's appearance, such as the object's last known pose before occlusion, can be used to substantially prune the search. Other strategies are also possible. For example, the size of the bounding box could be used as a discriminating factor to reduce the solution space.

From the perspective of machine vision, this thesis nicely demonstrates how appearance-based methods can be used in conjunction with range sensing to solve the pose estimation problem. The key benefit afforded by range sensors with calibrated fields of view is the ability to rapidly normalize views, thus facilitating appearance matching. Here this was used to collapse a 6 d.o.f. problem to 3 d.o.f. by means of view-based position estimates and augmentation of training data. Further, modern range sensors can be made largely invariant to ambient illumination. Although appearance-based techniques are still sensitive to occlusion and outliers, the work reported here shows a way to account for these situations at slight additional cost.

REFERENCES

[1] Athitsos, V. and Sclaroff, S., An appearance-based framework for 3D hand shape classification and camera viewpoint estimation, Proceedings on Automatic Face and Gesture Recognition, pp. 45-50, 2002.

[2] Besl, P.J. and Jain, R.C., Segmentation through variable-order surface fitting, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-10, pp. 167-192, 1988.

[3] Besl, P. and McKay, N., A method for registration of 3-D shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 14, pp. 239-256, 1992.

[4] Bhanu, B., Representation and shape matching of 3-D objects, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-6, pp. 340-351, 1984.

[5] Bischof, H. and Leonardis, A., Robust recognition of scaled eigenimages through a hierarchical approach, Proceedings of Computer Vision and Pattern Recognition, pp. 664-670, 1998.

[6] Black, M.J. and Jepson A.D., EigenTracking: Robust matching and tracking of articulated objects using a view-based representation, Proceedings of ECCV, pp. 329-342, 1996.

[7] Campbell, R.J. and Flynn, P.J., A survey of free-form object representation and recognition techniques, Computer Vision and Image Understanding, Vol. 83, No. 3, 2001.

[8] Campbell, R.J. and Flynn, P.J., Eigenshapes for 3D object recognition in range data, Proceedings of CVPR'99, II : 505-510, 1999.

[9] Cropp, A., Palmer, P., and McLauchlan, P., Estimating pose of known target satellite, Electronics Letters, Vol. 36, Issue 15, pp. 1331-1332, 2000.

[10] Cyr, C. M., Kamal, A. F., Sebastian, T. B., and Kimia, B. B., 2D-3D registration based on shape matching, Mathematical Methods in Biomedical Image Analysis, pp. 198-203, 2000.

[11] Edwards, J. L., An active, appearance-based approach to the pose estimation of complex objects, Proceedings of IROS, Vol. 3, pp. 1458-1465, 1996.

[12] Flynn, P. J., 3-D object recognition with symmetric models: symmetry extraction and encoding, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, pp. 814-818, 1994.

[13] Jasiobedzki, P., Abraham, M., Newhook, P., and Talbot, J., Model Based Pose Estimation Technique for Autonomous Operations in Space, Proceedings of Information Intelligence and Systems, pp. 211-215, 1999.

[14] Jasiobedzki, P., Talbot, J., and Abraham, M., Fast 3D Pose Estimation for On-Orbit Robotics, 31st International Symposium on Robotics, pp. 434-440, 2000.

[15] Jenkin, M., and Jasiobedzki, P., Computation of Stereo Disparity for Space Materials, IROS, pp. 1461-1466, 1998.

[16] Lowe, D., Object Recognition from Local-Scale Invariant Features, International Conference on Computer Vision, pp. 1150-1157, 1999.

[17] Moghaddam, B. and Pentland, A., Probabilistic visual learning for object detection, Proceedings on Computer Vision, pp. 786-793, 1995.

[18] Nayar, S.K., Murase, H. and Nene, S.A., Parametric appearance representation, Oxford University Press, Chapter 6.

[19] Skočaj, D. and Leonardis, A., Robust recognition and pose determination of 3-D objects using range images in eigenspace approach, Proceedings of 3rd International Conference on Pattern Recognition, pp. 171-178, 2001.

[20] Trucco, E. and Verri, A., Introductory Techniques for 3-D Computer Vision, Prentice-Hall Inc., 1998.

[21] Turk, M. and Pentland, A., Eigenfaces for recognition, Journal of Neuroscience, 3(1), pp. 71-86, 1991.

[22] Yoshimura, S. and Kanade, T., Fast template matching based on the normalized correlation by using multiresolution eigenimages, Proceedings of IROS, pp. 2086-2093, 1994.

[23] Zhang, J., Gao, J., and Liu, J., Figure-ground separation by a dynamical system, IEEE Transactions on Image Processing, Vol. 8, No. 1, pp. 115-122, 1999.