RECOGNIZING VOLUMETRIC OBJECTS IN THE PRESENCE OF UNCERTAINTY

Tal Arbel

Department of Electrical Engineering
McGill University

May 1995

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

© Tal Arbel, MCMXCV

Abstract

This thesis describes a new framework for parametric shape recognition. The key result is a method for generating classifiers in the form of conditional probability densities for recognizing an unknown from a set of reference models. Our procedure is automatic. Off-line, it invokes an autonomous process to estimate reference model parameters and their statistics. On-line, during measurement, it combines these with a priori context-dependent information, as well as the parameters and statistics estimated for an unknown object, into a conditional probability density function, which represents the belief in the assertion that the unknown is a particular reference model. Consequently, the method also permits the assessment of the beliefs associated with a set of assertions based on data acquired from a particular viewpoint. The importance of this result is that it provides a basis by which an external agent can assess the quality of the information from a particular viewpoint, and make informed decisions as to what action to take using the data at hand.

The thesis also describes the implementation of this procedure in a system for automatically generating and recognizing 3D part-oriented models. We show that recognition performance is near perfect for cases in which complete surface information is accessible to the algorithm, and that it falls off gracefully when only partial information is available. This leads to a sequential recognition strategy in which evidence is accumulated over successive viewpoints (at the level of the belief distribution) until a definitive assertion can be made. Experimental results are presented showing how the resulting algorithms can be used to distinguish between informative and uninformative viewpoints, rank a sequence of images on the basis of their information (e.g. to generate a set of characteristic views), and sequentially identify an unknown object.

Résumé

Cette thèse décrit une nouvelle approche pour la représentation paramétrique des formes. Le resultat principal est une méthode pour générer des classes sous la forme de fonctions de densité de probabilité pour identifier un inconnu parmi un ensemble de modèles de référence. Notre procédure est automatique. Dans sa phase d'apprentissage, elle fait appel à un processus autonome pour estimer les paramètres des modèles de référence et leurs statistiques. Dans sa phase d'identification, elle combine les paramètres des modèles de référence avec d'autre information contextuelle ainsi qu'avec les paramètres et statistiques d'un objet à identifier pour produire une fonction de densité de probabilité qui représente la confiance en une hypothèse d'identification de l'inconnu parmi les modèles de référence. Conséquemment, la méthode permet aussi l'estimation de la confiance associée à un ensemble d'hypothèses basés sur les données obtenues d'un certain point de vue. L'importance de ce résultat est qu'il procure une base par laquelle un agent externe peut estimer la qualité de l'information provenant d'un point de vue et en conséquence prendre une décision éclairée quant à l'action à réaliser.

Cette thèse décrit aussi une réalisation concrète de cette procédure dans un système pour générer et reconnaître des modèles 3D représentés par leurs parties. Nous montrons que la performance de la procédure de reconnaissance approche la perfection pour les cas où une description compl'ete de la surface est disponible et que les résultats se dégradent d'une manière prévisible et graduelle quand seulement une information partielle est présentée. Ceci débouche sur une stratégie de reconnaissance séquentielle par laquelle les évidences sont accumulées sur plusieurs vues (au niveau des distributions de confiance) jusqu'à ce qu'une hypothèse définitive puisse être établie. Des résultats expérimentaux démontrent comment l'algorithme peut être utilisé pour: distinguer entre les vues informatives et non-informatives, classer une séquence d'images sur la base de leur information (i.e. pour générer un ensemble de vues caractéristiques) et identifier séquentiellement un object inconnu.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor, Frank P. Ferrie, for his support, both on a financial and morale level, his technical advice, as well as his continuous encouragement throughout my degree. His guidance has changed the course of my career.

Next, I would like to express my gratitude to Peter Whaite, whose mathematical and technical expertise were essential to the development of the approach. Thanks for all your support and patience. Perhaps I can return the favor some day...

A special thank you to Gilbert Soucy whose technical support went above and beyond the call of duty. Thanks for acting as a sounding board for my ideas.

Thanks to my family and friends for their support and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Object recognition represents the highest level of processing in a visual system, synthesizing all the information provided by the lower level processes, and using the result to perform reasoning tasks. However, interest in the problem of object recognition has generally been restricted to identifying and locating an object in a visual scene (see survey paper by Arman & Aggarwal 1993*b*). We argue that absolute identifications are limited in that they are biased to a particular system's criteria as to what constitutes a "winning" hypothesis. Furthermore, since no visual system works in complete isolation, external processes must be completely informed about any ambiguities in the results of recognition to be able to make knowledgeable decisions.

In this thesis, we focus our attention on *model-based* recognition. This implies making use of object models that store a priori knowledge about the features essential for object characterization. Recognition consists of matching an unknown object with a model in a predefined database. We broaden the scope of the conventional recognition problem and investigate the notion of the quality of the identification. In our terms, this qualification refers to determining the degree of confidence in the classification. Ideally, representation of this information takes the form of a conditional probability density function, which we will refer to as a *belief distribution*, describing the likelihood of correspondence between an unknown model and each of the reference models. Such a measure is essential to an active recognition process which can use it as feedback in the collection of further data to resolve ambiguity.

Our process works as follows. On-line, we make measurements of an unknown object, the task being to infer the model in the database which best represents it. Problems of this type fall under the category of *inverse problems*, and are underdetermined. Rather than constrain the solution with prior assumptions about the world, we seek a general solution to the inverse problem that makes the sources of knowledge explicit. To this end, we must address the issue of how one systematically incorporates different sources of knowledge into the process of recognition, specifically ambiguities that arise from measurement and representation. We seek a method that represents all relevant contextual information by informative models, and encompasses these descriptions into the solution. Ideally, we would like to represent these sources of knowledge as probability density functions, so as to communicate all the uncertainties to the recognition engine and, in this manner, make well-informed decisions.

1. Overview of the Approach

The application of this work is three-dimensional object recognition in which objects are represented by parametric shape descriptors such as superellipsoids (Barr 1981, Bajcsy & Solina 1987, Raja & Jain 1992, Ferrie, Lagarde & Whaite 1993), deformable solids (Darrell, Sclaroff & Pentland 1990, Pentland & Sclaroff 1991), and algebraic surfaces (Subrahmonia, Cooper & Keren 1992). We introduce a new framework for parametric shape recognition based on a probabilistic inverse theory first introduced by Tarantola (1987). Application of this theory leads to a Bayesian recognition strategy similar to that used in other approaches (Subrahmonia et al. 1992). However, the important distinction of our methodology is that it leads to a mechanism by which the belief distribution used to classify shape models can be automatically generated. In doing so, important sources of contextual knowledge are taken into account that are less obvious in traditional approaches. Such knowledge includes i) a priori knowledge of the objects comprising the database, ii) information obtained from the process of estimating model parameters for an unknown object, and iii) information from the physical theories giving rise to the reference models themselves. We will show how the theory systematically enumerates each of these sources of knowledge, and combines them so as to create the desired belief distribution.

In our context, object models are constructed through a process of *autonomous exploration* (Whaite & Ferrie 1991, Whaite & Ferrie 1993*b*, Whaite & Ferrie 1994) in which a part-oriented, articulated description of an object is inferred through successive probes with a laser range-finding system. Figure 1.1a shows the set-up used to perform experiments — a two-axis laser range-finder mounted on the end-effector of an inverted PUMA-560 manipulator. For any particular viewpoint, such as the one shown in Figure 1.1b, a process of bottom-up shape analysis leads to an articulated model of the object's shape

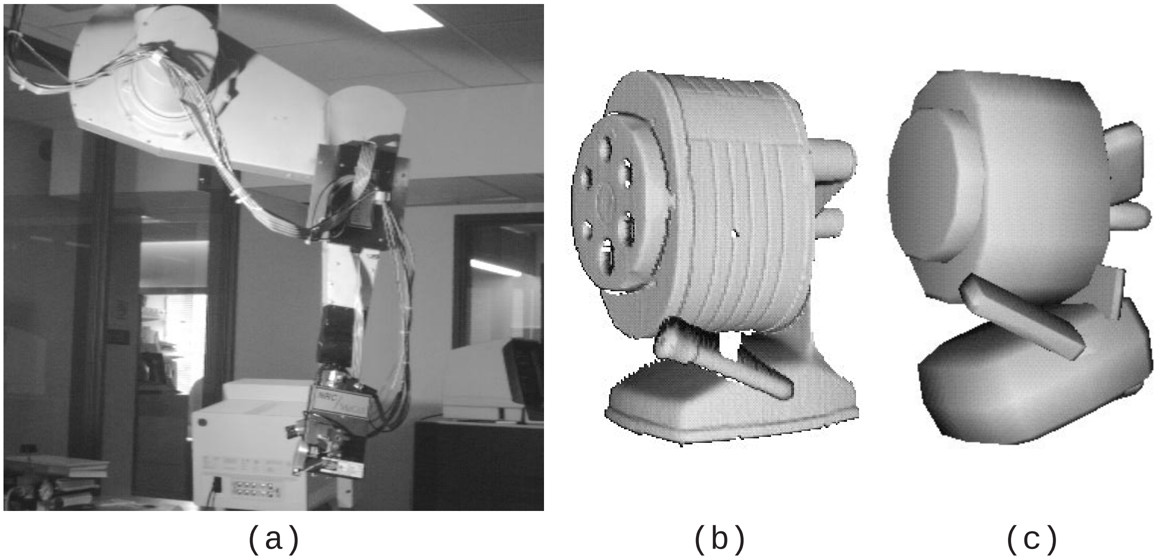


FIGURE 1.1. (a) Mobile laser range-finding system used to construct object models. (b) Laser range-finder image of a pencil sharpener rendered as a shaded image. (c) An articulated, part-oriented model of the sharpener using superellipsoid primitives; 8 superellipsoids are used, one corresponding to each of the parts of the object.

(Figure 1.1c) in which each part is represented by a superellipsoid primitive (Ferrie, Lagarde & Whaite 1993). Associated with each primitive is a covariance matrix **C** which embeds the uncertainty of this representation and which can be used to plan subsequent gaze positions where additional data can be acquired to reduce this uncertainty further (Whaite & Ferrie 1991, Whaite & Ferrie 1993*b*). A system which automatically builds object models based on this principle is reported in (Whaite & Ferrie 1994, Lejeune & Ferrie 1993).

Applying the inverse theory to our context is straightforward. Off-line, a database of object models is generated by presenting each object prototype to the model building system. Each object is in turn represented by several sets of parameters, one corresponding to each part of the object. On-line, the recognition phase proceeds identically to model-building except for one key difference. On each iteration (gaze-point calculation → data acquisition → data merging (fusion) → parameter estimation), the belief (in the form of a conditional probability density function) for each reference object given the current parameter estimate of the unknown object is calculated. If a clear winner stands out in terms of maximum likelihood, the process is terminated. Otherwise the process is allowed to continue and the beliefs in each reference model are updated on the basis of the newly acquired data. In this way, evidence can be incrementally gathered during the process of exploration.

Because the inverse solution to the recognition problem is in the form of a belief distribution, it provides not only descriptions of the results, but of the ambiguities in them as well. This qualification is important in that visual processes rarely work in complete isolation, and external processes using the results of recognition should be fully informed before making decisions. For example, consider an external agent searching for a particular object with limited resources. It must be able to assess what it sees from a particular viewpoint and quickly determine if the extracted information describing the characteristics of the objects in the scene is useful in identifying the target, so as to be able to evaluate alternate strategies. These strategies may include making assessments based on the current information, or using it to decide where to look next. It must do all this while taking into account prior knowledge about the environment. In this thesis, we will show how the resulting belief distributions can be used to (i) assess the quality of a viewpoint based on the assertions it produces, and (ii) sequentially recognize an object by accumulating evidence at a probabilistic level.

Finally, we note that to be able to solve a large number of problems in vision, we need to be able to model what we know about the world. The inverse theory, which tells us how to represent prior knowledge, and how to combine the knowledge to obtain the solution, is therefore an ideal candidate for the solution of a wide variety of vision problems. Although in this thesis, we concentrate on the problem of object identification, the theory can easily be applied to the problems of object classification or object representation. We will briefly discuss other possible applications of the theory in Chapter 8.

2. Overview and Organization of Thesis

Very few recognition schemes have attempted recognition based on the parameters of volumetric models. One reason for this has been due to the shortage of efficient bottom-up systems capable of building stable representations for multi-part objects. In Chapter 2, we present an overview of the many recognition strategies introduced over the past decade. We will classify the different schemes in terms of the features used to describe the objects, as well as the matching schemes used to compare the unknown object to the models in the database. We will focus our attention on the recognition schemes that do attempt to recognize parametric models (Pentland & Sclaroff 1991, Keren, Cooper & Subrahmonia

1992, Raja & Jain 1992), and illustrate the main differences between those approaches and ours.

The proposed recognition strategy raises a number of fundamental issues. First, how is parametric uncertainty used and communicated between the processes of model building and recognition? Clearly they are not independent. Furthermore, the recognition process must take both the uncertainties in the database, as well as the measurement uncertainties of the unknown object, into account. In Chapter 3, we present an overview of the inverse theory (Tarantola 1987), and in Chapter 4, we show how the appropriate belief distributions used for recognition can be determined from such information by applying the inverse theory to the problem of model recognition. This leads to a method of deriving, for each object model instance, the conditional probability of that model given the current estimated parameters of the unknown and their covariances.

Second, which parametric model would provide the most useful descriptions for recognition? We have chosen to use the parameters of superellipsoid models as features for the purposes of recognition. Representations based on superquadrics, however, pose a number of problems due to degeneracies in shape and orientation. Other parametric forms, e.g. algebraic surfaces (Keren et al. 1992), are sometimes less problematic and can offer a more stable basis for recognition purposes. Nonetheless, it is still desirable to choose forms in which physical attributes can be ascribed to model parameters in an intuitive manner. The finite-element representations introduced by Pentland and his colleagues are a case in point (Darrell et al. 1990, Pentland & Sclaroff 1991). For our purposes, where shape is initially partitioned into part-oriented segments, superellipsoids are attractive both in the range of shapes they can represent as well as their computational simplicity. In Chapter 5, we describe a method of avoiding degeneracies in the case of the superellipsoid, which permits the use of this convenient parametric form without incurring undue computational overhead.

Finally, what is the best manner in which to accumulate information? The model-building process is expensive, the merging of data from different viewpoints in particular (Soucy 1992). While this might be acceptable for database generation, recognition tasks must often be performed rapidly. An alternative is to consider the use of partial information obtained independently from different viewpoints. Because recognition from one view is not always reliable, key to this idea is the ability to assess the quality of the hypotheses from a particular view. In Chapter 6, we illustrate how to use the belief distributions to distinguish between informative and uninformative viewpoints by application of an external threshold. Furthermore, we show how the resulting ambiguities can be resolved without the need for data fusion by accumulating evidence in the form of the belief distributions from sequential viewpoints.

In Chapter 7, we describe and compare the performance of the recognition procedure using beliefs computed from complete and partial surface information respectively. We show that the beliefs generated from partial data retain their selectivity and result in a minimum number of false-positive indications. We illustrate this for single-part objects as well as for parts of complex, articulated models. We show that the majority of the incorrect states are accompanied by very low beliefs, and can be removed by applying a simple threshold. We see that the distributions of the beliefs from different viewpoints are bi-modal, indicating a clear distinction between the informative and uninformative viewpoints. This justifies the use of the threshold to distinguish between them. In addition, we perform a series of incremental recognition experiments that illustrate that the maximum likelihood hypothesis[1] prevails in a largely view-invariant manner. Therefore, we show that, by tabulating the votes for each hypothesis, after a sequence of trials, the correct winner emerges. Finally, we indicate how the system's success at recognizing primitives of articulated models, even with only partial information available, paves the way for recognition of multiple-part objects.

We conclude in Chapter 8 with some general observations on our current work and points for future research.

3. Contributions

In this work, we claim the following contributions:

1. We present a clear and structured recipe for recognition of volumetric models based on a generalized inverse theory.
2. The procedure for both database generation and identification is completely *automatic.*
3. The method explicitly enumerates its sources of contextual knowledge so it can easily be modified to work elsewhere.
4. The result is in the form of a conditional probability density function so ambiguities can be communicated to external processes to evaluate and base decisions upon.
5. The result is a basis by which an external agent can assess the quality of the information from a particular viewpoint by distinguishing between *informative* and *uninformative* viewpoints.
6. An *incremental* recognition scheme is presented.

[1] This refers to the hypothesis that the correct answer is the one with the highest belief.

7. The method is highly discriminant, capable of recognizing models despite wide variations in their size and shape.

8. The system paves way for multiple-part recognition based on graph-matching, by outlining a way to compare the nodes.

9. Strategies for solving other problems in vision such as object classification, and *active* recognition are outlined.

CHAPTER 2

Object Recognition Schemes

1. Introduction

Over the years, much research has been devoted to solving the problem of object recognition. In general, the connotations of the terminology in the field have been fairly widespread. As a result, many classification and model-based representation methods have fallen under the category of object recognition. In this chapter, the first thing we wish to do is clarify the terminology and distinguish model-based object recognition schemes from the others. In doing so, we will restrict ourselves to comparing our work to those methods that extract a series of features from an unknown model, and compare them to a series of models stored a priori in a database. The result we require of the method is a hypothesis, or a group of hypotheses, about the likelihood of the unknown object matching each of the models in the database.

We wish to distinguish recognition schemes from *object classification* schemes, where the goal is to classify the unknown object into one of a series of predetermined categories. Examples of these schemes include work done by Raja & Jain (1992), where objects are represented by superquadric models, and then placed into into one of twelve predetermined categories of 3D shapes (*geons*). In this case, classification is based on low level features derived from the superquadric model, such as bent or straight axis, and straight or curved edges. Other classification schemes include (Hutchinson, Cromwell & Kak 1989).

Within these classification schemes are those methods that attempt to represent an object by a descriptive model, while restricting the possible models to a finite group. These schemes fall under the category of *model-based representation* schemes. Here, measurements of an object are taken and then an attempt is made to recover a higher level representation from them. However, rather than adhere to a strict bottom-up strategy, these methods constrain the search by only permitting the representation to be one of a few possible types of models, stored in a database prior to the experiment. The goal is a model of the object, generated by using top-down information. This differs substantially from the goals of *model-based recognition*, where the descriptive model of the unknown object has already been computed prior to the experiment without the use of top-down information. The goal here is, therefore, not to compute an object model, but rather to hypothesize a match between the computed model and each of a series of predefined models in a database. Examples of model-based representation methods include those that measure the object, and attempt to fit the data to each of the model types stored in the database. The model chosen is the one that fits the data with the smallest overall error (Kriegman & Ponce 1990, Newman, Flynn & Jain 1993, Wu & Levine 1994). Other examples can be found in (Pentland 1987, Dickinson, Pentland & Rosenfeld 1992).

A wide variety of model-based object recognition schemes have been developed over the past thirty years (Chin & Dyer 1986). In this chapter, we wish to review various methods, and distinguish them by the type of features they use to characterize the objects (Section 2) and the way in which they represent the objects in the database (i.e. in what form should the features be combined into object models), as well as the method used to match an object to a model in the database (Section 3). These traits are inherently linked in that the type of representation chosen dictates the features used for recognition, as well as the type of matching strategy chosen, its robustness, and the system's efficiency. The survey will illustrate the problem that in many recognition strategies, implicit assumptions about the nature of the world are applied. These assumptions may include constraints on the kinds of objects that will be recognized (i.e. specialized methods that look for particular features, such as the number of holes in a block), the kinds of features that are interesting (i.e. methods that characterize objects by curvature or boundary features), or the values of the features themselves (i.e. methods that look for sizes within a particular range of values). As a result, the methods may work well in a particular context but, because of the hidden nature of the assumptions, cannot be easily modified to work elsewhere.

2. Features

Most of the previous work in object recognition have used low-level or intermediate level features in order to characterize objects. Low-level schemes look to match edges, corners, curves, lines, silhouettes, contours, boundaries, holes and other predetermined features in their attempt to recognize objects. For example, linear edge fragments, and circular arcs are used in (Grimson 1987, Grimson 1989, Grimson & Lozano-Perez 1987). Line segments, corners, zeros of curvature, other 2D perceptual structures are used in (Lamdan, Schwartz & Wolfson 1988, Thompson & Mundy 1987, Lowe 1985, Huttenlocher & Ullman 1987).

Intermediate schemes extract features of surface patches. For example, Flynn & Jain (1991*a*) use surface area, surface type (cylindrical, spherical or planar), and other surface attributes as features for recognition. Other such schemes use surface normals, centroids, direction of axes of surfaces, centers of sphere (Kim & Kak 1991), or edge adjacency types, i.e. convex, or concave (Fan, Medioni & Nevatia 1987, Fan, Medioni & Nevatia 1989, Fan 1990). Jain & Hoffman (1988) describe models by the area and diameter of the surface patches. They also incorporate the minimum and maximum distances to the adjacent patches.

Some methods incorporate "global" features in their recognition schemes. To date, the majority of the global features have referred to general descriptions such as the number of parts of the object, or the number of local features the objects have (such as the number of edges or corners). Methods that use these kinds of features exclusively are quite inefficient in that these descriptions are generally unstable. For example, the number of object parts depends, quite heavily, on the resolution of the segmenter, and is very sensitive to occlusion. For this reason, the majority of the schemes that use global features, use them in conjunction with other types of features, and use them only to help prune the search space.

Chin & Dyer (1986) state that in order to be able to recognize a wide variety of rigid parts, independent of viewpoint, one needs to be able to extract view-invariant 3D features and match them with features of 3D models. The problem with the majority of low, intermediate, and global features is that they are often unstable, view-variant, and highly susceptible to noise. The shortage of "high-level" features (or stable, global descriptors) features restricts the capabilities of most recognition schemes to a limited class of objects, seen from a few, fixed viewpoints. Examples of such high level features used in recognition schemes are the intrinsic properties of parametric models such as algebraic surfaces (Keren et al. 1992), or superquadrics (Pentland & Sclaroff 1991, Arbel, Whaite & Ferrie 1994*a*). Here, the intrinsic properties used are the parameters of the models themselves. These descriptors will be discussed in more detail in the next section.

3. Matching Strategies

Many methods attempt to find a corresponding match between features of the object models and features extracted from the unidentified object. The matching scheme chosen by a recognition system should be able to achieve this task while accounting for the possibility of missing information due to partial occlusion, measurements from single viewpoint, etc. In many schemes, the dimensionality of the features extracted forces much of the computational burden to be shifted to the matching procedure. As a result, much of the focus of the current literature is to reduce the arduousness of this process. Schemes that represent objects by robust, and stable models, and use rich, global features implicit to their descriptions, reduce the job of the matching process and render it more efficient. In this section, we will discuss the various types of matching strategies that have emerged over the past decade.

3.1. Tree Search Approach. One general category of matching schemes has been the *tree search* approach. Here, after object features are extracted, a tree of possible model-to-object feature matches is built. Each path from root to leaf represents one possible solution to the correspondence problem. The idea is to search for the path that would ensure a consistent matching between object and model. Many people have developed methods to prune the search tree in order to reduce the search time. These include constraining the range of unary feature values (such as the length of an edge), as well as the range of binary feature values describing the interrelationships between unary features (such as the angle between normal vectors) (Grimson 1987, Grimson 1989, Grimson & Lozano-Perez 1987, Flynn & Jain 1991*a*, Flynn & Jain 1991*b*). Swain (1988) developed a *decision tree* approach to object recognition, employing topological, relational and view-dependent information in its decision rules.

3.2. Relational Schemes. Another category of matching schemes is the relational approach. Relational matching schemes attempt to establish correspondence by representing both the sensory data and the model data as graphs, where the nodes represent features, and the arcs represent the geometric relationship among the features. The recognition problem is then a matter of establishing graph isomorphism. Once again, many pruning techniques have been introduced to reduce the search space (Kak, Vayada, Cromwell, Kim & Chen 1987, Faugeras & Hebert 1983, Bhanu 1982). In (Fan et al. 1987, Fan et al. 1989, Fan 1990), objects are represented as attributed graphs, and the approach is to look for the model graph with the largest set of matched nodes. They use low, intermediate, and global level features to prune the search space. Jain & Hoffman (1988) arranged the features described above into groups: shape features, object face features, and boundary information. Range images are represented using these groups as "evidence conditions". The images, along with the weights indicating the uncertainty in the features corresponding to the models, are stored in a database. Matching is performed by computing a similarity measure describing the degree of support for a hypothesis. Bolles, Horaud & Hannah (1984) extended previous work (Bolles & Cain 1982) to develop a 3D *local-feature-focus* technique. The method uses a graph-matching technique to identify the largest cluster of image features that matches a cluster of model features. It works by selecting one feature in the image around which it tries to find a cluster of consistent secondary features. After creating a list of all possible image-feature-to-model-feature assignments, it creates a graph of all possible pairwise assignments. Connections between nodes are established if the two assignments they represent are mutually consistent.

3.3. Pruning the Database by Model-Based Indexing. A major problem facing object recognition schemes has been the enormous complexity involved in searching the database to select the possible candidate models. Many methods have been introduced to reduce the computational complexity. One such method has been the *geometric hashing* scheme (Lamdan & Wolfson 1990, Grimson & Huttenlocher 1990, Flynn & Jain 1992). In these schemes, a hash table, containing surface-surface pairing constraints for all the models in the database, is constructed. Surface pairing measurements are derived from the scene, and the corresponding values are located in the appropriate entry in the table. This results in many possible matches, which are resolved by using predefined sets of rules.

Flynn (1992) investigated the case of large databases. His approach was to reduce the number of prototypes needed to be considered by excluding all redundant feature groups that result from object symmetry. As well, a measure of saliency was assigned to each group in the scene, so that “uninformative” groups are not considered. Other filtering schemes were introduced in (Kim & Kak 1991, Stein & Medioni 1992).

3.4. Automatic Schemes. Many of the schemes described involve a substantial amount of on-line model analysis due, in part, to the additional constraints and conditions computed with the introduction of each new model to the database. In order to reduce the expense of run-time calculations, interest has grown in automatic recognition schemes, with much of the database processing performed off-line. New methods were introduced that performed much of the “precompiling” prior to recognition, improving the efficiency of the task at run-time (Goad 1983). One such scheme uses a representation called an *aspect graph*, first introduced in (Koenderink 1976, Koenderink 1979). These are graphs where each node represents a topologically distinct 2D viewpoint of a 3D object. The arcs, referred to as “visual events”, describe transformations from one viewpoint to another. Essentially, the graph divides the view-sphere into stable regions defining “characteristic views”, where small changes in viewing position do not affect the topological structure of the set of visible features. (Recent work on aspect graphs can be found in (Sripradisvarakul & Jain 1989, Eggert & Bowyer 1989, Eggert, Bowyer, Dyer, Christensen & Goldgof 1992, Kriegman & Ponce 1989, Bowyer & Dyer 1990).) Precomputing an aspect graph for each model in the database can improve the efficiency of the recognition task at run time, by predefining the possible interpretations of the models in the database. The major disadvantages of the representation are the high storage requirements, and large construction times.

Interpretation trees (Ikeuchi 1987*a*, Ikeuchi 1987*b*) are similar to aspect graphs, in that the Gaussian sphere is tessellated into possible viewing positions. This technique includes the additional step of computing a tree containing the possible interpretations of each model in the database. All possible shapes of the model, at the root, are generated, and the similar shapes are grouped into clusters at the leaves of the tree. Different divisions of the aspects form different paths from the root of the tree to the leaves.

Another off-line scheme is the *prediction hierarchy* method. Here, the 2D appearance of some 3D objects is predicted in advance, and merged into a tree-like structure which is traversed during recognition (Burns & Kitchen 1988). Similarly, Dickinson, Pentland & Rosenfeld (1990) introduced hierarchical aspect graphs. The method entails extracting object features, such as the convexity of the contours of the faces, from 3D volumetric primitives. These features, assessed from many viewpoints, are arranged in a hierarchical graph that links facial features to faces to face structures to primitives. In addition, the statistical relations between the features are also stored. On-line matching includes generating hypotheses about the identity at the lowest possible level of the tree. Other automatic schemes have been investigated by (Hansen & Henderson 1988, Hansen & Henderson 1989, Arman & Aggarwal 1993*a*).

3.5. Matching Parametric Models. The final set of matching schemes examined includes those methods that find correspondence by matching the parameters of *parametric models*. A parametric model refers to a representation built by taking measurements of an object, and fitting the data to a model represented by a mathematical equation. These models can be volumetric models, such as superellipsoids and generalized cylinders, or surface descriptors, such as splines, and fourth order polynomials. The parameters of these equations describe implicit, global characteristics of the object, and are therefore stable descriptions for recognition. However, very few schemes find correspondence based on the high-level descriptions themselves. Rather, the majority of the current work in 3D object recognition consists of building the models and extracting externally chosen features from them. An example of this trend is Dickinson et al.'s (1990) choice of convexity of contours of volumetric primitives (see previous section) as features for recognition. In general, extrinsic features are usually much more sensitive to noise, occlusion and viewpoint than the intrinsic ones, such as the parameters themselves (this includes their associated covariances). They usually consist of geometrical (low or intermediate) features, or rather unstable global features (see Section 2). By avoiding using the parameters themselves as features for recognition, limitations on the robustness of the recognition scheme are introduced, especially with complex objects.

There are many reasons for the shortage of recognition schemes based on the parameters of these models. One reason has been the shortage of efficient bottom-up systems capable of building stable representations for multi-part objects. This is due, in part to the shortage of effective segmentation schemes, as well as methods that combine information from different viewpoints. Because of this, it has been thought that recognition of these models is only suitable for single-part objects that are simple in shape, measured from only one viewpoint (see survey paper by Arman & Aggarwal 1993*b*).

In addition, because the uncertainties associated with the parameters are rarely calculated, it is not not generally considered feasible to compare models based on them alone. This is because when fitting a model to data that is noisy, there is an inherent lack of uniqueness in the parameters that describe the model. In these cases, it is impossible to make a definite statement as to which model fits the data best (Whaite & Ferrie 1991). Therefore, matching based on one set of parameters alone would not give accurate results. For this reason, rather than choose external constraints that would force one model over the other, it would be more instructive to embed the uncertainty in the chosen description into the feature set. In Chapter 4, we will show that taking the uncertainties in the measurement parameters into account (as well as the uncertainties of the parameters of the models in the database) in the distance metric permits greater variations in the measured feature, while still maintaining high selectivity in the discrimination between models. We will also show that matching without taking the uncertainties into consideration would cause many false identifications. An example of such a method is that proposed by Pentland & Sclaroff (1991). The authors introduce a method for the recognition of deformable superellipsoid models based on their parameters alone. Using their scheme, proximity is measured by evaluating the normalized dot product of the parameter vectors of the unknown object and of each of the models in turn. The model with the highest dot product value is considered to be the one closest to the unknown, and is the model chosen. We will illustrate the weaknesses of methods such as these later in Chapter 4. Methods that do include uncertainties in the features can be found in (Hutchinson et al. 1989, Kwong & Kim 1993, Subrahmonia et al. 1992).

Often it is the case that problems associated with the parametric model are misunderstood to be insurmountable. For example, using the parameters of superquadrics for the purposes of recognition has been avoided, because the problem of non-uniqueness of parameters has never been addressed. As a result, the power of these representations, namely that they can provide accurate, global descriptions of objects over a wide variety of sizes and shapes with relatively few parameters, has not yet been fully exploited. This has limited their uses to modelling tasks (as in CAD design), and to the recognition of simple objects (see Boult & Gross 1988).

As well, few schemes use a probabilistic approach to the solution. Bayesian recognition of algebraic surfaces has been examined by Subrahmonia et al. (1992). They represent objects by fourth order polynomials (Keren et al. 1992), and measure similarity between the unknown and the models in the database by employing a Mahalanobis distance measure between the coefficient vectors. This distance measure includes the uncertainties in the measured model as well as in the stored models (see (Subrahmonia et al. 1992), Appendix, p.39). Recognition is achieved by choosing the model that results in the smallest Mahalanobis distance. The key difference between their approach and ours (Arbel, Whaite & Ferrie 1994*a*) lies in the techniques used to obtain the solution. They have used strict Bayesian techniques to derive the solution. We have structured the problem within the framework of an inverse problem theory, which offers a clear and structured formula for representing all prior knowledge, as well as a global recipe for combining this knowledge to obtain the posterior information. The result is a *general solution*, which, in our specific case, degenerates to a Bayesian solution similar to theirs. In addition, this framework lends itself to the problem of model-based object recognition, but can be applied to various other problems such as object classification and generic recognition (see Chapter 8).

The other important difference in our schemes is that they, and most others, (see survey papers by Arman & Aggarwal 1993*b*, Chin & Dyer 1986) are interested in the constructing a discriminant that makes an absolute identification of the measured object. In accordance with Marr's (1982) "Principle of Least Commitment", we feel that it is more instructive to retain several possible explanations, rather than choose a single one. This is especially true when the hypotheses are comparable in accuracy. We will demonstrate that making assessments about identity from single measurements can be erroneous, especially when made from viewpoints that provide little information about the characteristics of the object. Rather than make claims about the object's absolute identity, our method communicates the *belief* in the possible hypotheses as feedback to the recognition procedure, in order to further reduce the ambiguity using an active strategy.

In the next chapter, we will introduce the general inverse theory first proposed by Tarantola (1987). We will explain the reasoning behind explicitly enumerating all sources of knowledge available. As well, we will show how, by representing this knowledge as probability density functions, we can easily combine the information to obtain a solution to the inverse solution in the form of a conditional probability density function. Finally, we will illustrate how the general solution reduces to the classical Bayesian solution, providing the desired posterior information. In Chapter 4, we will show how we use this framework within the context of a model-based object recognition system that matches parametric models.

CHAPTER 3

The Inverse Problem Theory

1. Introduction

The recognition problem requires us to infer from measurements of an unknown object that model which best represents it in a data base of known objects. Like all inverse problems, the recognition problem is ill posed in that, i) several models can give rise to identical measurements and, ii) experimental uncertainty gives rise to uncertain measurements. As a result it is not possible to identify the unknown object uniquely. There are various ways of conditioning ill posed problems, but these all require strong, and often implicit, a priori assumptions about the nature of the world. As a result a method may work well only in specific cases and, because of the hidden implicit nature of the conditioning assumptions, cannot be easily modified to work elsewhere.

For this reason we have adopted the very general inverse problem theory of Tarantola (Tarantola 1987). It makes the sources of knowledge used to obtain inverse solutions explicit, so if conditioning is required, the necessary assumptions about that knowledge are apparent and can be examined to see if they are realistic. Also, and importantly, the question of whether a solution is ill-posed or not is shown correctly to be an operational issue. The theory tells us how the knowledge we have can be combined to obtain a solution, but leaves any decision about the its usefulness up to the tasks that require it. For example, when attempting to recognize objects we would ideally want the unknown model be identified correctly all the time. Because of experimental uncertainties this can never happen, and there is always the possibility that an object will be identified incorrectly. Only the task can know if the likelihood of errors is acceptable.

This raises the interesting question of what we should do if the level of errors is not acceptable. Because the sources of knowledge are explicit they are not only visible to the operational tasks, but are also potentially open to manipulation by them. In principal it should be possible for the task to condition or actively acquire the a priori knowledge required to make the solution acceptable. We have already demonstrated that what we call autonomous exploration functions well at the model building level (Whaite & Ferrie 1993*a*, Whaite & Ferrie 1994) and we now intend, with the aid of this theory, to incorporate feedback from the recognition task as well.

We begin in Section 2 with the introduction of the concept of formal knowledge representation. Section 3 will go on to explicitly enumerate the sources of a priori information used to constrain the inverse problem. Finally in Section 4, we discuss the way the sources are combined to obtain the solution to the inverse problem.

2. States of Information

In a physical system inverse problems are conveniently visualized as a mapping between two different spaces: the *model space* M and the *data space* D. We will assume throughout that M and D are vector spaces with a finite number of real valued parameters. We will define M as an abstract space of points, each representing a conceivable model of the system, and D will refer to the space of all possibly "observable" instrumental responses. A model in M is represented by $\mathbf{m} = (m_1, m_2, \ldots, m_m)$, and a measurement in D by $\mathbf{d} = (d_1, d_2, \ldots, d_n)$.

The view taken by Tarantola is that our knowledge of a physical parameter (model or measurement) is subjective in that it varies from observer to observer depending upon the data in their possession. We can quantify this subjective knowledge by a rule, called the *state of information*, which assigns a positive number reflecting our belief that the true value of the parameter lies within some given range. Mathematically such a rule is a probability[1] (Pfeiffer 1978). For a vector space the rule is represented by a probability density function.

Thus the first postulate of the theory is that our knowledge about a set of parameters is described by a probability density function over the parameter space. This requires us to devise appropriate density functions in order to represent what we know about the world. However, probability theory tells us nothing about the way in which to choose the rule that assigns probabilities. In general the form of these distributions depends on the the interpretation one wishes to place on mathematical probability in the context of a physical system. In some cases, for example a measuring instrument, we can histogram the measurements of a known input and arrive at a rule based on the relative frequencies of measurements occurring within different ranges. In others, for example theoretical knowledge, we must rely

[1] Really a measure – a probability is a normalizable measure.

on our intuition, imagination, and experience to formulate a rule that assigns probabilities, and then verify it through experimental procedure. There are two special and important cases which reflect the fact that our knowledge falls between two extremes: i) the state of perfect knowledge and ii) the state of null information.

The *state of perfect knowledge* is appropriately represented by the Dirac delta function $\delta(\mathbf{x}-\mathbf{x}_0)$, and shows we believe totally that $\mathbf{x} = \mathbf{x}_0$, but not at all that it is any other value. It is the state of information we aspire to but can never attain. In practice we can use it when sources of error are negligible in comparison with others.

The *state of null information* $\mu(\mathbf{x})$ on the other hand is used to represent the fact that we have absolutely no knowledge about the parameters at all. It plays the role of the reference state in the theory, in much the same way that noise is used when measuring information in terms of signal to noise ratios. An obvious choice for $\mu(\mathbf{x})$ is a uniform distribution which, because all parameter values are equally likely, implies no particular belief in any of them.

A uniform $\mu(\mathbf{x})$ is not necessarily correct, especially when dealing with different parametrizations of the same physical system. For example if we are interested in finding the location of some feature in 3D space a uniform distribution over the space of Cartesian coordinates seems a reasonable choice. However a uniform distribution over the space of polar coordinates will result in higher belief values for those features closer to the origin. For our purposes, we will usually assume that $\mu(\mathbf{x})$ is uniform. We claim that this is a reasonable approximation of the true form as we are only dealing with a single class of models, and the same parametrization.

3. Sources of A Priori Information

The second part of Tarantola's theory is a division of the sources of a priori knowledge into two specific categories: the knowledge given by a theory which describes the physical interaction between models and measurements, and knowledge obtained independently of that theory. For our purposes the latter can be broken down into two more independent categories: information we have about the model from measurements, and information from unspecified sources about the kinds of models which exist in the world.

Note that although the theory assumes this information can be represented by probability density functions, it does not tell us their form. Choosing an appropriate form for the a priori distributions can only be done in the context of the problem we are attempting to solve and is largely an intuitive matter. As to whether the form of the distribution is appropriate once chosen, this can only be verified in a scientific manner by experimentally confirming predictions. We are bound by the nature of the scientific method.

3.1. Information Obtained from Physical Theories. A physical theory is a solution to the *forward problem*. It tells us how to predict the error-free values of the observed data $\mathbf{d}$ obtained when observing a given model $\mathbf{m}$,

$$\mathbf{d} = \mathbf{g}(\mathbf{m}). \tag{1}$$

However, no theory is ever exact and there are always "modelization" uncertainties. In the theory these shall be represented by the conditional probability density $\theta(\mathbf{d}|\mathbf{m})$ of observing $\mathbf{d}$ given a model $\mathbf{m}$. When the modelization uncertainties are insignificant we may be able to assume an exact theory, $\theta(\mathbf{d}|\mathbf{m}) = \delta(\mathbf{d} - \mathbf{g}(\mathbf{m}))$. Otherwise $\theta(\mathbf{d}|\mathbf{m})$ effectively places "error bars" on the theoretical relation. Figure 3.1 illustrates these differences in the forward modelization.

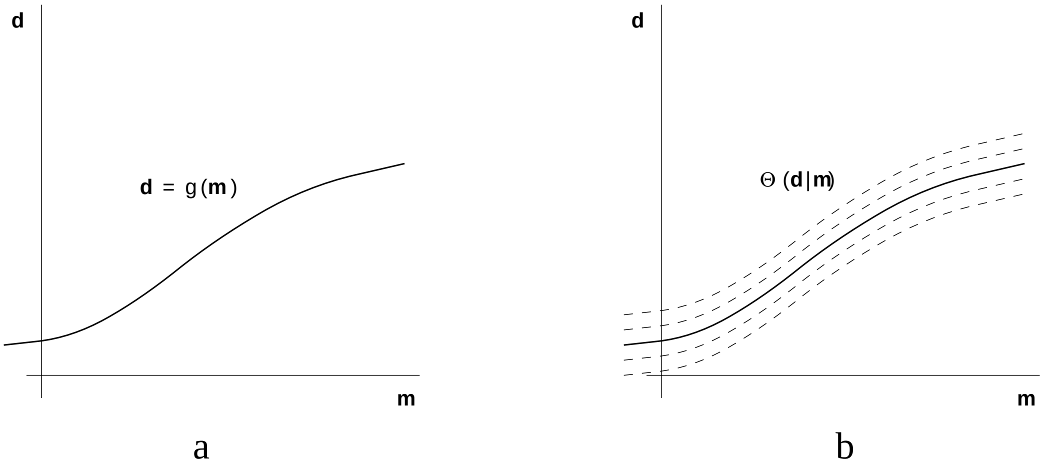


FIGURE 3.1. Forward modelization. (a) If the uncertainties in the forward modelization are neglected, $\mathbf{d} = \mathbf{g}(\mathbf{m})$ gives the predicted data values, $\mathbf{d}$ for each model $\mathbf{m}$. (b) If we cannot neglect the uncertainties in the forward-modelling, they can be described by the conditional probability density function, $\theta(\mathbf{d}|\mathbf{m})$, which gives, for each model $\mathbf{m}$, a probability density for $\mathbf{d}$. This corresponds to placing "error bars" on the theoretical relation $\mathbf{d} = \mathbf{g}(\mathbf{m})$.

Because we are using information in both the data and model spaces we require an expression of the theoretical knowledge in the joint space $M \times D$. Because the non-informative density in the data space $\mu_D(\mathbf{d})$ is independent of the models and by definition contains no information about the data, the joint distribution $\theta(\mathbf{d}, \mathbf{m}) = \theta(\mathbf{d}|\mathbf{m})\ \mu_M(\mathbf{m})$ must contain exactly the same information that $\theta(\mathbf{d}|\mathbf{m})$ does, and can therefore be used to represent the theoretical information over the joint model and data space. Figure 3.2(b) illustrates the joint distribution $\theta(\mathbf{d}, \mathbf{m})$.

3.2. Information Obtained from Measurements and A Priori Information on Model Parameters. Much of the knowledge we have about a problem comes in the form of experimental measurements of observable parameters. All instruments are subject to varying degrees of uncertainty so our knowledge of the observable parameters is imperfect. The probability density function representing the information obtained from measurements will be designated by $\rho_D(\mathbf{d})$. Let $\mathbf{d}_{out}$ denote the value delivered by the instrument at each measurement of a given value of $\mathbf{d}$. The most useful and general way of conveying the results of the statistical analysis of the instrument errors is by defining a probability density function for the value of the output, $\mathbf{d}_{out}$, when the actual input is $\mathbf{d}$. The conditional probability density function conveying this information is denoted $\nu(\mathbf{d}_{out}|\mathbf{d})$. If the actual result of the measurement $\mathbf{d}_{out} = \mathbf{d}_{obs}$ (what we have observed is actually the data outputted by the instrument), then we can use Bayesian reasoning and conclude:

$$\rho_D(\mathbf{d}) = \frac{\nu(\mathbf{d}_{obs}|\mathbf{d})\ \mu_D(\mathbf{d})}{\int_D \nu(\mathbf{d}_{obs}|\mathbf{d})\ \mu_D(\mathbf{d})\ d\mathbf{d}} \tag{2}$$

In specific situations it is often the case that we know something else about the models which can be usefully applied. For example in some industrial applications there may only be a finite number of known objects, and these might always be supported by a conveyer belt. Knowledge like this is a powerful constraint and can be used to eliminate many of the unconstrained solutions. The problem is that this kind of knowledge often appears in the form of ad-hoc selection criteria applied at a late stage of processing, or as conditioning constraints embedded in the formulation of the model. Here it is made explicit as another source of knowledge and represented by the probability distribution $\rho_M(\mathbf{m})$.

For our purposes we will assume that the measurements and the a priori model constraints are obtained independently. In that case the knowledge they represent can be combined to give a probability density function

$$\rho(\mathbf{d}, \mathbf{m}) = \rho_D(\mathbf{d})\ \rho_M(\mathbf{m}) \tag{3}$$

over the joint space $M \times D$.

Figure 3.2(a) illustrates the two a priori sources of information represented by their probability density functions: $\rho_D(\mathbf{d})$ and $\rho_M(\mathbf{m})$. Here, they are seen projected onto data and model space. The combination of these sources of information is represented by the probability function $\rho(\mathbf{d}, \mathbf{m})$ lying in joint $M \times D$ space.

4. Solution to the Inverse Problem

The solution to the inverse problem is in principal quite straight forward — it is simply a matter of combining the sources of information, i.e. the theory, the measurements, and the a priori constraints. The complication is the manner in which they are to be combined.

This is the third part of Tarantola's theory. He takes the approach that the classical theory of logic gives rules by which humans handle information. In particular the logical operation of *conjunction* is appropriate, i.e. the solution to the inverse problem is given by the theory AND the measurements AND any a priori information about the models. The notion of logical conjunction is extended to define the conjunction of two states of information (Tarantola 1987, pages 29–31).

DEFINITION 1 (conjunction of states of information). *Let* $f_1(\mathbf{x})$, $f_2(\mathbf{x})$ *be probability density functions representing the states of information* P_1 *and* P_2 *respectively, and* $\mu(\mathbf{x})$ *be the probability density function representing the state of null information. Then*

$$\sigma(\mathbf{x}) = \frac{f_1(\mathbf{x})\, f_2(\mathbf{x})}{\mu(\mathbf{x})} \tag{4}$$

where $\sigma(\mathbf{x})$ *is the* a posteriori *probability density function representing the* conjunction of states of information (P_1 AND P_2).

With this definition we can combine the information from the joint prior probability density function $\rho(\mathbf{d}, \mathbf{m})$ and the theoretical probability density function $\theta(\mathbf{d}, \mathbf{m})$ to get the a posteriori state of information

$$\sigma(\mathbf{d}, \mathbf{m}) = \frac{\rho(\mathbf{d}, \mathbf{m})\, \theta(\mathbf{d}, \mathbf{m})}{\mu(\mathbf{d}, \mathbf{m})} \tag{5}$$

where $\mu(\mathbf{d}, \mathbf{m})$ is the joint non-informative probability density function (the reference state of information). According to Tarantola, this equation is more general that those obtained through traditional approaches, but degenerates to them in specific cases. Under the conditions mentioned, the solution is identical to the Bayesian solution (Tarantola 1987, page 61).

Figure 3.2(c) illustrates the combination of the prior information:$\rho(\mathbf{d}, \mathbf{m})$ and $\theta(\mathbf{d}, \mathbf{m})$ displayed in (a) and (b) respectively. One can see that the conjunction of information, represented by the joint posterior distribution $\sigma(\mathbf{d}, \mathbf{m})$, localizes the knowledge provided by the each of the distributions separately.

What we require however is the a posteriori information about the model parameters, and this is simply given by the marginal probability density function

$$\begin{aligned}\sigma(\mathbf{m}) &= \int_D \sigma(\mathbf{d},\mathbf{m})\; d\mathbf{d} \\ &= \int_D \frac{\rho(\mathbf{d},\mathbf{m})\;\theta(\mathbf{d},\mathbf{m})}{\mu(\mathbf{d},\mathbf{m})}\; d\mathbf{d}\end{aligned} \tag{6}$$

When it is assumed that the model and data non-informative densities are independent, i.e. that $\mu(\mathbf{d},\mathbf{m}) = \mu_D(\mathbf{d})\mu_M(\mathbf{m})$, the equation for the marginal a posteriori density function becomes

$$\sigma(\mathbf{m}) = \int_D \frac{\rho_D(\mathbf{d})\;\rho_M(\mathbf{m})\;\theta(\mathbf{d}|\mathbf{m})\;\mu_M(\mathbf{m})}{\mu_D(\mathbf{d})\;\mu_M(\mathbf{m})}\; d\mathbf{d} \tag{7}$$

This reduces to:

$$\sigma(\mathbf{m}) = \rho_M(\mathbf{m}) \int_D \frac{\rho_D(\mathbf{d})\;\theta(\mathbf{d}|\mathbf{m})}{\mu_D(\mathbf{d})}\; d\mathbf{d}. \tag{8}$$

Equation (8) is *the solution* to the general inverse problem. From $\sigma(\mathbf{m})$ it is possible to obtain any sort of information we wish about the model parameters: mean values, median values, maximum likelihood values, errors, covariances, confidence intervals, etc.

Figure 3.2(d) illustrates the solution to the inverse problem. The resulting distributions representing the posterior model information, $\sigma(\mathbf{m})$, as well as the posterior data information, $\sigma(\mathbf{d})$, are seen projected onto the model and data spaces respectively. By comparing the posterior density function, $\sigma(\mathbf{m})$, to the prior one, $\rho_M(\mathbf{m})$ (displayed in (a)), one can see that some information on the model parameters has been gained. Prior to the conjunction of information, there was only vague information about the kinds of models that exist in the world. After, one can see that a degree of certainty about the model parameters has been gained. This is due to the addition of the prior data information, $\rho_D(\mathbf{d})$, and the theoretical information $\theta(\mathbf{d},\mathbf{m})$.

While the probability density $\sigma(\mathbf{m})$ allows us to estimate the posterior values of the model parameters, the density function $\sigma(\mathbf{d})$ is also useful in that is permits the estimation of the posterior values of data parameters (i.e. "recomputed data"). The posterior data information is computed as follows:

$$\sigma(\mathbf{d}) = \rho_D(\mathbf{d}) \int_M \frac{\rho_M(\mathbf{m})\;\theta(\mathbf{d}|\mathbf{m})}{\mu_D(\mathbf{d})}\; d\mathbf{m}. \tag{9}$$

By comparing $\rho_D(\mathbf{d})$ and $\sigma(\mathbf{d})$ in Figures 3.2(a) and (d) respectively, one can see that knowledge has also been gained about the data parameters.

The *existence* of the solution to the inverse problem simply means that $\sigma(\mathbf{m})$ is not identically null. If it were then this would indicate incompatibility between the theory, the experimental results, and what is assumed a priori about the model parameters.

The *uniqueness* of the solution refers to the fact that there is one and only one solution. This is evident when, by the solution, we mean the probability density $\sigma(\mathbf{m})$ itself. $\sigma(\mathbf{m})$ could be pathological (non-normalizable, multi-model, etc.) but that only indicates the nature of the information possessed on the model parameters. The information itself is uniquely defined as a consequence of the the uniqueness of the conjunction of states of information.

In this chapter, we have introduced the general inverse theory as a framework for solving the recognition problem. We have illustrated how to obtain the solution to the inverse problem in the form of a conditional probability density function, by explicitly naming all sources of knowledge and representing each by a probability density function. We have also shown how the posterior information is obtained under conditions that reduce the general solution to the classical Bayesian solution. In the next chapter, we will show how to apply the theory to the recognition of parts of articulated, parametric models.

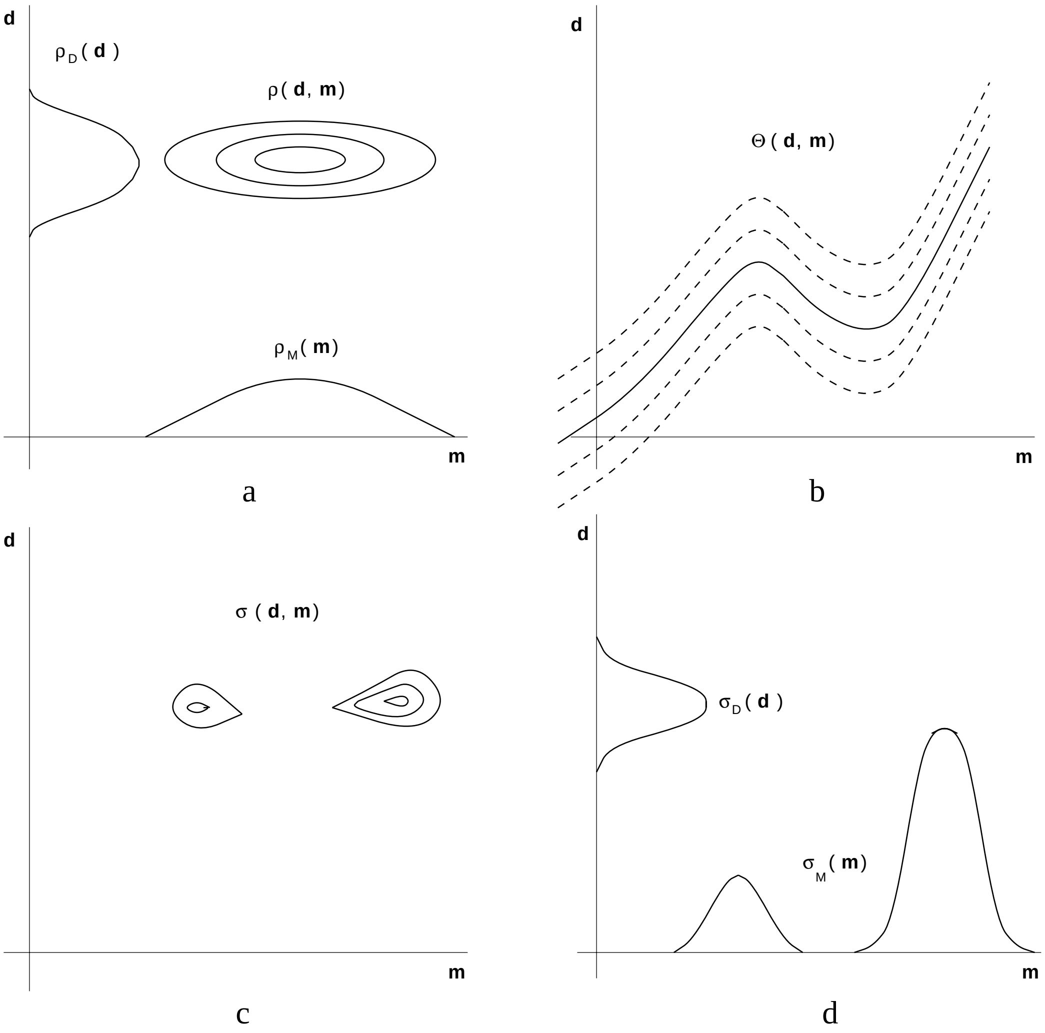


FIGURE 3.2. The probability densities in combined model and data space (Tarantola 1987, page 54). (a) The probabilities $\rho_D(\mathbf{d})$ and $\rho_M(\mathbf{m})$ represent the a priori information on the observable parameters (data) and the a priori information on model parameters respectively. $\rho(\mathbf{d},\mathbf{m})$ represents the joint a priori information in the $D \times M$ space. Since the a priori data information is independent of the a priori model information, we have $\rho(\mathbf{d},\mathbf{m}) = \rho_D(\mathbf{d})\ \rho_M(\mathbf{m})$. (b) $\theta(\mathbf{d},\mathbf{m})$ represents the information on the physical correlations between $\mathbf{d}$ and $\mathbf{m}$, as predicted by a physical theory. (c) $\sigma(\mathbf{d},\mathbf{m})$ represents the joint posterior information, which is the conjunction of the two states of information $\rho(\mathbf{d},\mathbf{m})$ and $\theta(\mathbf{d},\mathbf{m})$, such that: $\sigma(\mathbf{d},\mathbf{m}) = (\rho(\mathbf{d},\mathbf{m})\ \theta(\mathbf{d},\mathbf{m}))/\mu(\mathbf{d},\mathbf{m})$. (d) From $\sigma(\mathbf{d},\mathbf{m})$, we can obtain the marginal probability density functions $\sigma(\mathbf{m}) = \int_D \sigma(\mathbf{d},\mathbf{m})\ d\mathbf{d}$ and $\sigma(\mathbf{d}) = \int \sigma(\mathbf{d},\mathbf{m})\ d\mathbf{m}$. By comparing the the posterior probability density, $\sigma(\mathbf{m})$, to the prior one, $\rho_M(\mathbf{m})$, we can see that some information on the model parameters has been gained. This is due to the addition of the prior data information, $\rho_D(\mathbf{d})$, and the theoretical information, $\theta(\mathbf{d},\mathbf{m})$.

CHAPTER 4

The Part Recognition Problem

1. Introduction

In the previous chapter, we have presented the general inverse theory as a framework for solving the part recognition problem. In this chapter, we will illustrate how to apply the theory to the recognition of parts of articulated models obtained through a classical bottom-up system. We will show how to use the parameters of the models as descriptors for recognition.

In the system we have constructed, articulated object models are created by successive probes of a laser-rangefinder through a process of *autonomous exploration* (Whaite & Ferrie 1991, Whaite & Ferrie 1993*b*, Whaite & Ferrie 1994). For any particular viewpoint, range measurements are taken, surfaces are reconstructed then segmented into parts, and individual models are fit to each part. Each part is represented by a superellipsoid primitive, where points on the surface (x, y, z) satisfy the following implicit equation:

$$f(\mathbf{x}, \mathbf{a}) = \left(\left| \frac{x}{a_x} \right|^{2/\epsilon_2} + \left| \frac{y}{a_y} \right|^{2/\epsilon_2} \right)^{\epsilon_2/\epsilon_1} + \left| \frac{z}{a_z} \right|^{2/\epsilon_1} = 1 \tag{10}$$

where a_x, a_y, a_z indicate extent in the x, y, and z directions respectively, ϵ_1 and ϵ_2 are the shape descriptors, and t_x, t_y, t_z and θ_x, θ_y, and θ_z indicate the translation and rotation in the x, y, and z directions. Associated with each primitive is a covariance matrix $\mathbf{C}$ which embeds the uncertainty of this representation which can be used to plan subsequent gaze positions where additional data can be acquired to reduce this uncertainty further (Whaite & Ferrie 1991, Whaite & Ferrie 1993*b*). Currently, the first five superellipsoid parameters, a_x, a_y, a_z, ϵ_1, ϵ_2, and their associated covariances, are used as part descriptors for object recognition.

Usually, the model fitting process is treated as the solution to an inverse problem where the forward problem is the prediction of the range data that will be gathered from some known volumetric model. However, we will take a larger view and treat *the whole system as a measuring instrument.*

We will let M be the space of volumetric models to be recognized. Given some model $\mathbf{m}$ in the scene, range measurements are taken and from these an *estimate* of the model is obtained, $\mathbf{d}$, which we call a *measurement of the model* in the scene. We denote the space of possible model estimates D.

Given this scenario, we solve the inverse problem (Section 5) by examining the sources of information: the information obtained from physical theories (Section 2), information available through measurement (Section 3), and the a priori information on models (Section 4).

2. Information Obtained from Physical Theories

We first formulate an appropriate distribution to represent what is known about the forward problem. If the entire system were treated as a perfect measuring instrument (free of all uncertainties), the vector function $\mathbf{g}(\mathbf{m})$ introduced in (1) would be the identity function. This would mean that measuring the model would always generate its true parameters: $\mathbf{d} = \mathbf{m}$. However, measuring instruments are never perfect. Formulating a physical theory that enables us to predict estimates of the model parameters given a model in the scene is too difficult given the complications of the system. We therefore collect these estimates empirically through a process called the *training* or learning stage of the recognition process. Here, measures of a known model, $\mathbf{m}$, are collected N times. The measures, $\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_N$, are used to calculate the conditional probability density function $\theta(\mathbf{d}|\mathbf{m})$ for each model by assuming a multivariate normal distribution. These Monte Carlo experiments are exactly like those found in traditional statistical pattern classification methods (Nilsson 1965, Mood & Graybill 1963). A mean, $\bar{\mathbf{m}}$, is computed for each known model:

$$\bar{\mathbf{m}} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{d}_j \tag{11}$$

The covariance matrix, $\mathbf{C}_T$, describing estimated modelling errors for a model $\bar{\mathbf{m}}$, is calculated as follows:

$$\mathbf{C}_T = \frac{1}{N-1}\sum_{j=1}^{N}(\mathbf{d}_j - \bar{\mathbf{m}})(\mathbf{d}_j - \bar{\mathbf{m}})^T \tag{12}$$

Therefore, the final equation for $\theta(\mathbf{d}|\mathbf{m}_i)$ is:

$$\theta(\mathbf{d}|\mathbf{m}) = N(\mathbf{d} - \bar{\mathbf{m}}, \mathbf{C}_T) \tag{13}$$

where N is the multivariate normal distribution:

$$N(\mathbf{d} - \bar{\mathbf{m}}, \mathbf{C}_T) = \frac{1}{\sqrt{(2\pi)^n \ |\mathbf{C}_T|}} \ \exp\left(-\frac{1}{2}(\mathbf{d} - \bar{\mathbf{m}})^T \mathbf{C}_T^{-1}(\mathbf{d} - \bar{\mathbf{m}})\right), \tag{14}$$

n being the dimension of the data space.

Experimental training is not an easy job. A representative sample of models in different poses, and of different scanner positions, must be taken. Otherwise, $\theta(\mathbf{d}|\mathbf{m})$ may either *underestimate* the errors in the estimation process and give high levels of false positive identifications, or conversely *overestimate* them and give low levels of true positive matches.

Later, we will show that, when we have a database of known models in the scene, we need only perform training on these models. The distribution representing the theoretical information, $\theta(\mathbf{d}|\mathbf{m})$, is created by simply summing the individual distributions for each of the known models in the following fashion:

$$\theta(\mathbf{d}|\mathbf{m}) = \sum_i^M \theta(\mathbf{d}|\mathbf{m}_i) \tag{15}$$

where M is the number of models in the scene. This means that it is not necessary to sample all of M, but only the models known to exist a priori. The training process is therefore considerably less complex than it first appears.

The result of training is a database of predefined model classes. Each class can be represented by an ellipsoidal cluster in multi-dimensional parameter space. Figure 4.1(a) illustrates the model classes resulting from training in a scene of four known models. The distributions of each class become elliptical in shape when seen projected onto 2D a_x/a_y parameter space. In (b), one can see how each individual class is created during the training process.

3. Information Obtained from Measurements

The measurement experiment gives a certain amount of information about the true values of the observable parameters. However, often measurement errors are not taken into account, and the estimated model parameters are assumed to be exact. This would imply that the probability density $\rho_D(\mathbf{d})$ would be represented by the Dirac delta function. This is usually an overly optimistic view of the state of information of the measurement, and may end up giving a very positive, but totally unjustifiable, identification of the object.

We do not accept this, however, and have gone to great pains in our system to characterize the ambiguities that exist in the parameters (Whaite & Ferrie 1992). As a result, we obtain not only an estimate of the observed model parameters $\mathbf{d}_{obs}$, but also an estimate

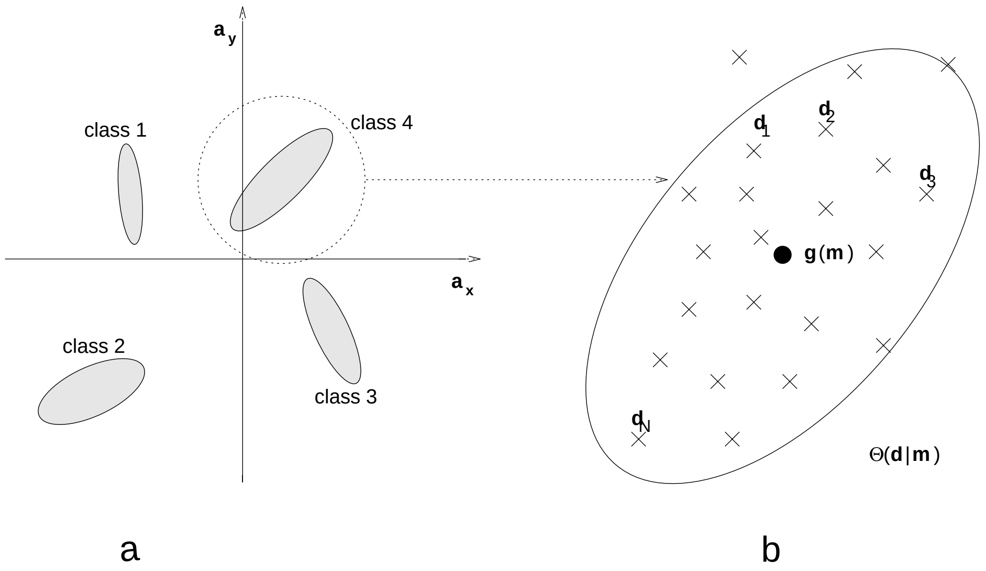


FIGURE 4.1. Results of training. (a) Model classes resulting from training are ellipsoidal clusters in multi-dimensional parameter space. Here, the projection onto the 2D a_x/a_y parameter space is shown. (b) Each model class is created by measuring the known model $\mathbf{m}$ N times. From these measures, $\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_N$, the mean $\bar{\mathbf{m}}$, and associated covariances, $\mathbf{C}_T$, are calculated by assuming a multivariate normal distribution.

of their uncertainty in the covariance operator $\mathbf{C}_d$. The assumption we make is that the multivariate normal distribution $N(\mathbf{d}-\mathbf{d}_{obs}, \mathbf{C}_d)$ represents our belief in the measurements. The probability density function representing this information is the conditional probability density function $\nu(\mathbf{d}_{obs}|\mathbf{d})$, such that:

$$\nu(\mathbf{d}_{obs}|\mathbf{d}) = N(\mathbf{d} - \mathbf{d}_{obs}, \mathbf{C}_d). \tag{16}$$

Therefore, we have:

$$\begin{aligned}\frac{\rho_D(\mathbf{d})}{\mu_D(\mathbf{d})} &= \frac{\nu(\mathbf{d}_{obs}|\mathbf{d})}{\int_D \nu(\mathbf{d}_{obs}|\mathbf{d})\ \mu_D(\mathbf{d})\ d\mathbf{d}} \\ &= \frac{1}{k}\ N(\mathbf{d} - \mathbf{d}_{obs}, \mathbf{C}_d)\end{aligned} \tag{17}$$

where k is the normalization constant:

$$k = \int_D N(\mathbf{d} - \mathbf{d}_{obs}, \mathbf{C}_d)\ \mu_D(\mathbf{d})\ d\mathbf{d}. \tag{18}$$

We have restricted D to the subspace of possible model estimates. We have assumed that $\mu_D(\mathbf{d})$ is a constant uniform distribution, entirely contained within that space, such that:

$$\int_D \mu_D(\mathbf{d})\ d\mathbf{d} = 1 \tag{19}$$

Therefore, the normalization constant reduces to:

$$k = \frac{1}{\int_D d\mathbf{d}} \int_D N(\mathbf{d} - \mathbf{d}_{obs}, \mathbf{C}_d)\; d\mathbf{d}, \tag{20}$$

where $\int_D d\mathbf{d}$ refers to the volume of data space. We assume that the measurement distributions are relatively sharp in that they lie entirely within D. In this case, $\int_D N(\mathbf{d} - \mathbf{d}_{obs}, \mathbf{C}_d)\; d\mathbf{d} \approx 1$, and $k \approx \frac{1}{\int_D d\mathbf{d}}$, a predefined constant.

The issue of how to define $\int_D d\mathbf{d}$ is a difficult one to address. In order to define such a space, a commitment to a permissible region of observed parameters must be established. As this is very difficult to define prior to measurement, the current framework leaves the measurement knowledge non-normalized. Under the assumption made that the measurement distributions are mostly contained within the data space, we can justify ignoring the normalization constant as it is equal for all measurements. here, different measurements can be compared.

However, for flatter measurement distributions, the assumption that D defines the space of all possible estimates is no longer valid. The normal distribution $\int_D N(\mathbf{d} - \mathbf{d}_{obs}, \mathbf{C}_d)\; d\mathbf{d} \ll 1$, and actually $k \ll \mu_D(\mathbf{d})$. In these cases, the measurement knowledge should actually be much larger than it is, to compensate for the spread out distribution. Because of these cases, independent measurements differ by an uncomputed factor, and can no longer be compared.

4. Information Obtained from A Priori Information on Model Parameters

In the current context, there are a finite number of reference models, $\bar{\mathbf{m}}_i, i = 1 \ldots M$, which are uniformly distributed. The probability density function used to convey this knowledge is

$$\rho_M(\mathbf{m}) = \sum_i^M P(\bar{\mathbf{m}}_i)\; \delta(\mathbf{m} - \bar{\mathbf{m}}_i), \tag{21}$$

where the $P(\bar{\mathbf{m}}_i)$ are the *a priori* model probabilities or weights reflecting the likelihood that the i^{th} model, $\bar{\mathbf{m}}_i$, occurs.

5. Solution to the Inverse Problem

Substituting the probability density functions in (17), (21) into the marginal a posteriori density function in (8) yields

$$\sigma(\mathbf{m}) = \frac{1}{k} \sum_i^M P(\bar{\mathbf{m}}_i)\; \delta(\mathbf{m} - \bar{\mathbf{m}}_i) \;\; \int_D N(\mathbf{d} - \mathbf{d}_{obs}, \mathbf{C}_d)\; \theta(\mathbf{d}|\mathbf{m})\; d\mathbf{d}. \tag{22}$$

Now, because $\delta(\mathbf{m}-\bar{\mathbf{m}}_i) = 0$ for all $\mathbf{m} \neq \bar{\mathbf{m}}_i$, and provided that $\theta(\mathbf{d}|\mathbf{m})$ is finite for $\mathbf{m} \neq \bar{\mathbf{m}}_i$, we may replace it with $\theta(\mathbf{d}|\bar{\mathbf{m}}_i)$. After doing this and regrouping, we get the inverse solution to the part recognition problem to be:

$$\begin{aligned}
\sigma(\mathbf{m}) &= \frac{1}{k} \sum_i^M \left(P(\bar{\mathbf{m}}_i) \int_D N(\mathbf{d} - \mathbf{d}_{obs}, \mathbf{C}_d)\; \theta(\mathbf{d}|\bar{\mathbf{m}}_i)\; d\mathbf{d} \right)\; \delta(\mathbf{m} - \bar{\mathbf{m}}_i) \\
&= \frac{1}{k} \sum_i^M Q_i\;\; \delta(\mathbf{m} - \bar{\mathbf{m}}_i). \qquad (23)
\end{aligned}$$

As we would expect, this tells us that the model *must* be one of the the models given a priori (21), but with a redistribution of the a priori model probabilities $P(\bar{\mathbf{m}}_i)$. For convenience, we will call:

$$Q_i = P(\bar{\mathbf{m}}_i) \int_D N(\mathbf{d} - \mathbf{d}_{obs}, \mathbf{C}_d)\; \theta(\mathbf{d}|\bar{\mathbf{m}}_i)\; d\mathbf{d} \qquad (24)$$

the *a posteriori* model probabilities or weights.

In order that we make a strong positive identification of the part, the Q_i should be concentrated in one model over all the others. If this is not the case, the information we have is inadequate to identify the model, either because the data set is insufficient, or because the empirical distribution, $\theta(\mathbf{d}|\mathbf{m})$, describing the measurement is inadequate.

Now that we have the form of the part recognition solution, we can re-examine in its light the ways in which we might obtain and represent the empirical distribution representing the measurement process. The crucial observation is that:

$$\theta(\mathbf{d}|\mathbf{m}) = \sum_i^M \theta(\mathbf{d}|\bar{\mathbf{m}}_i). \qquad (25)$$

This means, as we would intuitively expect, that the Monte Carlo estimates need not sample all of model space, but only the space of discrete models known to exist a priori, in this case, $\bar{\mathbf{m}}_i$.

Under the normality assumption made in (13) with reference to the conditional probability density function $\theta(\mathbf{d}|\mathbf{m})$, the solution for the a posteriori model probabilities becomes:

$$Q_i = P(\bar{\mathbf{m}}_i) \int_D N(\mathbf{d} - \mathbf{d}_{obs}, \mathbf{C}_d)\;\; N(\mathbf{d} - \bar{\mathbf{m}}_i, \mathbf{C}_T)\; d\mathbf{d}. \qquad (26)$$

The convolution of two normal distributions is a normal distribution (see Appendix A for details), therefore

$$\begin{aligned}
Q_i &= P(\bar{\mathbf{m}}_i) \int_D N(\mathbf{d} - \mathbf{d}_{obs}, \mathbf{C}_d)\; N(\mathbf{d} - \bar{\mathbf{m}}_i, \mathbf{C}_T)\; d\mathbf{d}, \\
&= P(\bar{\mathbf{m}}_i)\; N(\mathbf{d}_{obs} - \bar{\mathbf{m}}_i, \mathbf{C}_D), \qquad (27)
\end{aligned}$$

where $\mathbf{C}_D = \mathbf{C}_T + \mathbf{C}_d$. This result is important because it shows that, under the Gaussian assumption, observational errors and modelization errors simply combine by addition of the respective covariance operators (even when the forward problem is non-linear) (Tarantola 1987, page 58).

Convolving the measurement distribution against each of the reference model distributions has the effect of causing them to be more spread out. Therefore, the contribution of the a priori measurement information is to incorporate its uncertainties into the distributions of the model classes. Figure 4.2(a)–(c) illustrates this concept. In (a), the multivariate normal distributions of the reference models are seen projected onto the 2D a_x/a_y space. The black dot indicates the position of the measured model, $\mathbf{d}_{obs}$ in this space. Here, one can see that the measured model does not fall onto any of the distributions of the reference models. Strict distance metrics such as the one proposed by Pentland & Sclaroff (1991), do not take the uncertainties in the model, defined by the covariances, into account. These methods would find the measured model to be a member of class 3 since it lies closest to it. This identification would be incorrect. To see this, the 2D projection of the measured model distribution, $\rho_D(\mathbf{d})$, is displayed in (b). In (c), one can see the resulting distributions after convolving the measured model with each of the reference models. These distributions are much more spread out than those in (a). The covariances of the measured model define the degree and direction of the spread. Combining the prior information in this manner has lead to the identification of the measured model as being a member of reference class 4. Thus, the combination of the a priori information has improved the solution, in cases where recognition systems that use distance metrics that do not consider the measurement uncertainty would have generated a false identification.

The final equation for the a posteriori probability density function is

$$\sigma(\mathbf{m}) = \frac{1}{k} \sum_{i}^{M} P(\mathbf{m}_i)\; N(\mathbf{d}_{obs} - \bar{\mathbf{m}}_i, \mathbf{C}_D)\; \delta(\mathbf{m} - \bar{\mathbf{m}}_i). \tag{28}$$

This density function is comprised of one delta function for each model in the database. Each delta function is weighted by the *belief* $P(\bar{\mathbf{m}}_i)\; N(\mathbf{d}_{obs} - \bar{\mathbf{m}}_i, \mathbf{C}_D)$ in the model $\mathbf{m}_i$. The final distribution represents the “state of knowledge” of the parameters of $\mathbf{m}_i$. Figure 4.2(d) illustrates this distribution. The beliefs in each of the reference models, $\mathbf{m}_i$, are computed by evaluating each of the convolved distributions at $\mathbf{d}_{obs}$.

Because the normalization constant in (3) is not calculated in the current scheme, the resulting belief distributions are non-normalized. The result is that their values from

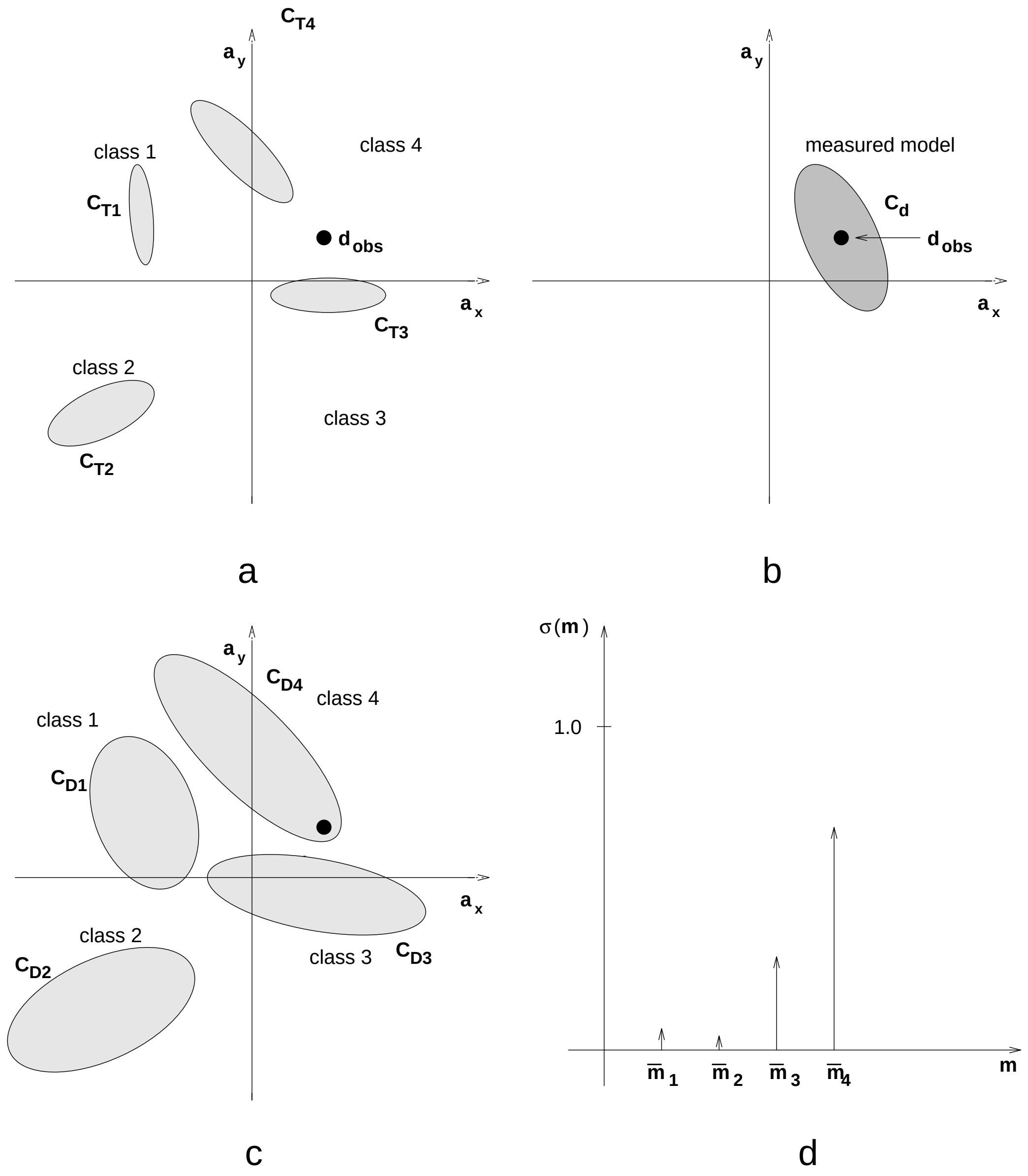


FIGURE 4.2. Creating the belief distribution. (a) Here, the reference model distributions, $\theta(\mathbf{d}|\mathbf{m}_i)$, are seen projected onto 2D, a_x/a_y parameter space. The black dot represents the position of the measured model, $\mathbf{d}_{obs}$ in 2D parameter space. We can see that it doesn't fall on any of the reference model distributions, and lies "closest", by a strict distance metric, to class 3. (b) The measured model distribution, $\rho_D(\mathbf{d})$, projected onto 2D parameter space. (c) The result of convolving the distribution in (b) with each of those in (a) is a version of (a) spread out in parameter space. We can see that now the measured model actually falls within the distribution of the fourth reference model class. (d) The resulting belief distribution. Notice that the system has the highest belief in model class 4, and a small belief in model class 3.

independent measurements cannot be compared. However, our current interest lies in the examining the relative beliefs resulting from each measurement.

The advantage of the method is that rather than establish a final decision as to the exact identity of the unidentified object, it communicates the degree of confidence in assigning the object to each of the model classes. It is then up to the interpreter to decide what may be inferred from the resulting distribution.

Figure 4.3 illustrates the kinds of results we get by applying the theory to a typical recognition problem. Here, the reference models were produced by training on models created with data acquired by scanning the objects all around their surfaces (i.e. complete 3D data). The reference models, consisting of a smaller sphere, a large sphere, and a lemon, can be seen in Figure 4.3a. The larger sphere was then measured from a single viewpoint, and the resulting model is shown in Figure 4.3b. The system's ability to distinguish the larger sphere from both the smaller sphere and the lemon was then tested. The result is the belief distribution found in Figure 4.3c. One can see that the system has a significantly higher degree of confidence in the hypothesis that the measured model was a large sphere.

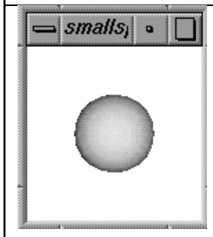
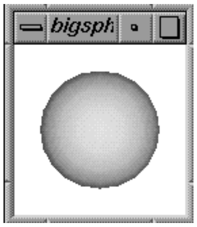
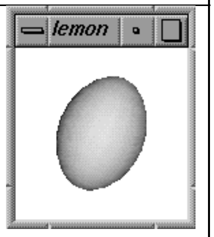
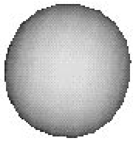
<table>
<tr><td></td><td colspan="3">a) Reference Models</td></tr>
<tr><td></td><td> </td><td> </td><td> </td></tr>
<tr><td>b) Measured Model</td><td colspan="3">c) Beliefs in Reference Models</td></tr>
<tr><td></td><td>6.12×10^{-43}</td><td>0.00273</td><td>0</td></tr>
</table>

FIGURE 4.3. Recognizing a sphere. (a) The reference models are: a smaller sphere, a larger sphere, and a lemon. (b) The measured unknown model. (c) The belief distribution.

In this chapter, we have presented a method for the recognition of volumetric models based on the general inverse theory (presented in Chapter 3). We have specified the probability density functions representing each sources of knowledge involved in the solution. We have also shown how to combine the information to obtain a solution in the form of a conditional probability density function, which we refer to as a belief distribution. In Chapter 7, we will illustrate a system that successfully recognizes real objects based on the methodology presented. We choose to represent objects by superellipsoid models, due to their computational simplicity. In order recognize based on the parameters of these volumetric models, the next chapter will specify how to avoid the degeneracies in shape and orientation associated with them.

CHAPTER 5

Degeneracies in the Superellipsoid Representation

1. Introduction

In the previous chapter, we have shown how to apply the inverse solution to a bottom-up system that produces volumetric models used for recognition. Although the recognition strategy described can be applied to any parametric model of an object, we have decided on the superellipsoid model as an object descriptor, due to the wide range of shapes in can represent as well as its computational simplicity. This type of model is also attractive in that the parameters describe physical attributes of the objects in an intuitive manner (see Chapter 1).

However, representations based on superquadrics pose a number of difficulties due to degeneracies in shape and orientation. By fitting data to superellipsoid models, the resulting covariance matrix defines a local region of parameter space (the ellipsoid of confidence) in which models are non-unique or ambiguous (Whaite & Ferrie 1991). The problem is that the ellipsoid of confidence represents the non-uniqueness at a single minima in parameter space. There might be other parameters at several disjoint minima that fit the data equally well. The problem of detecting all the possible local minima is a difficult one to address. For one thing, many of the minima may be geometrically equivalent. Rotating a model by 90° about an axis of symmetry will result in different rotational parameters, and re-ordered size parameters, without changes in appearance. In addition to these problems, other less obvious equivalence classes occur for superellipsoids. For example, in the $x-y$ plane, squares have shape parameters $\epsilon_2 = 0.1$, and diamonds 1.9. However, a diamond with equal size parameters is simply a square rotated by 45° . Detecting all possible equivalence classes, compounded with the uncertainty of the parameter set, is a difficult problem that must be addressed if one is to compare parameters for the purposes of recognition.

We will begin this chapter by enumerating the possible equivalence classes for the superellipsoid (Section 2). We will then show how to avoid these degeneracies without undue computational overhead, by representing a model by all of its possible equivalent forms. Consequently, models are described by multi-modal distributions (Section 3). Finally, we will show how to encompass multi-modal superellipsoid models into the recognition strategy described earlier (Section 4).

2. Equivalence Classes for the Superellipsoid Parameters

It was determined empirically that there are, in fact, only a finite number of possible equivalence classes for superellipsoids. Here, we will enumerate the most common degeneracies that occur in practice when using superellipsoids to model objects.

Using the superellipsoid description, two identical models may be described differently because of different labelling of the axes of symmetry. This is the most common type of equivalence class for superellipsoids, which we will refer to as *rotational equivalences*. Within this class, the highest number of equivalent states occurs when objects have the same shape in all planes. In superellipsoid terms, this means that the shape parameters, ϵ_1 and ϵ_2, are equal. Here, one can describe the same surface in any one of six different ways, by different assignments of the x, y, z axes. Therefore the size of the model can be appropriately described by any one of six permutations of the extent parameters, while the description of shape of the object remains unaltered. Figure 5.1 shows the six possible rotational equivalences of a model with equal shape parameters: $\epsilon_1 = \epsilon_2$.

It is important to note that this type of rotational equivalence class is only *strictly true* when the shape of the model is identical along all three axes of symmetry. We define strict equivalence to mean that the surfaces are identical in size and shape. It is in this situation only that the model can be described by any of the six permutations of the extent parameters. This is due to the limited descriptive powers of the superellipsoid model where shape is described by only two parameters: ϵ_2 and ϵ_1. ϵ_2 controls the shape in the x–y cross-sectional plane along the $z-axis$, while ϵ_1 describes the shape in two planes simultaneously, x–z, and y–z. When the model has different shape parameters, or $\epsilon_1 \neq \epsilon_2$, the superellipsoid description forces the unique shape to be along the $z-axis$ in all cases. Hence, the number of strict rotational equivalences in this case are limited to two, generated by the permuting the labelling of the x and y axes. In fact, regardless of their shape parameters, two models are rotationally equivalent if they only differ by having opposite labelling of their x and y axes.

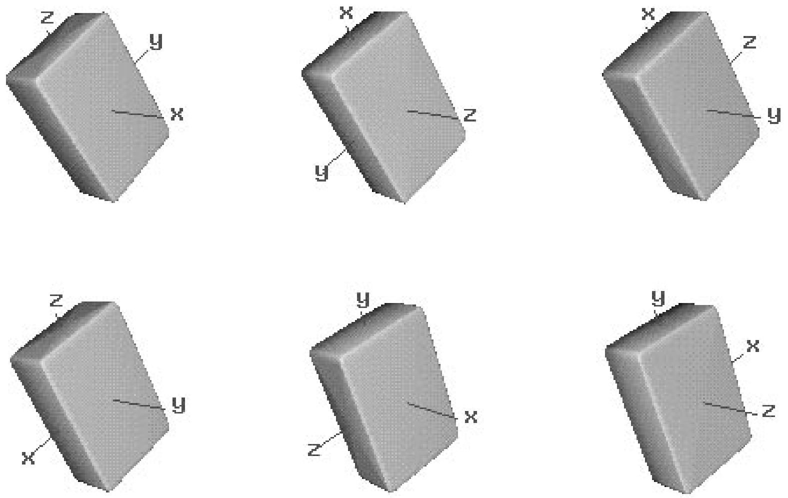


FIGURE 5.1. Rotational equivalences when the shape parameters are equal. Here, $\epsilon_1 = 0.1$ and $\epsilon_2 = 0.1$.

Figure 5.2 illustrates the case of a superellipsoid model of a cylinder with shape parameters: $\epsilon_1 = 0.1$, and $\epsilon_2 = 1.0$. In this case, the model is round in one cross-sectional plane, and rectangular in the other two. The superellipsoid description of the model forces the axis with the unique cross-sectional shape to be the $z - axis$. Figure 5.2(a) shows the original cylinder, and (b)–(f) shows the result of permuting the size parameters of the cylinder in (a). The fact that only (b) is identical to (a) illustrates that, for models with different shape parameters, the only strict rotational equivalence occurs in reversing the x and y axes.

Another type of equivalence class occurs when a superellipsoid model has a cross-sectional shape of a square in the x–y plane: $\epsilon_2 \approxeq 0.1$ and $a_x = a_y$. In this case, the model can also be described as a diamond: $\epsilon_2' \approxeq 1.9$, with the extent parameters: $a_x' = a_y'$, scaled such that $a_x' = \sqrt{2} \times a_x$. The size parameters must be scaled because, with a square, the extent parameters are measured from one face to the opposite one. However, for a diamond, they are measured from corner to corner (see Figure 5.3). This equivalence is only *strictly true* in the limit when the shapes are purely square ($\epsilon_2 = 0.1$) or diamond-like ($\epsilon_2' = 1.9$). In the range in between, $0.1 < \epsilon_2 < 1.9$, the shape of the model becomes more rounded. In this case, one can say that the equivalence between square-like models ($\epsilon_2 < 1$) and diamond-like models ($\epsilon_2' > 1$) is only *approximately true*, especially with the added effect of uncertainty. There is an approximate match between models such that $\epsilon_2' = 2.0 - \epsilon_2$ with

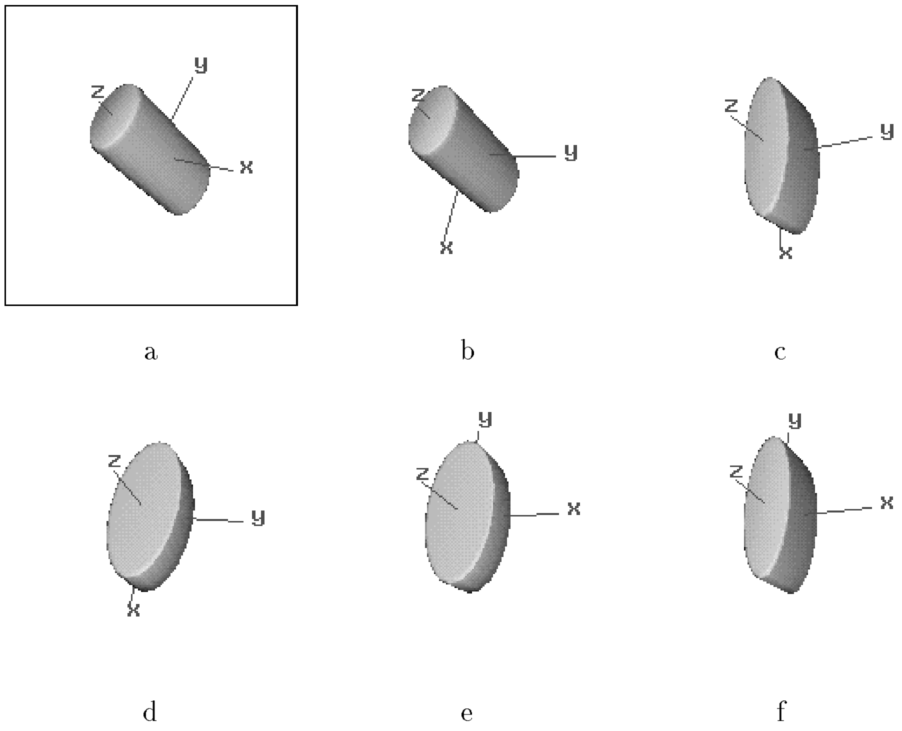


FIGURE 5.2. Permutations of the size parameters of model (a) when the shape parameters are not equal. Here, $\epsilon_1 = 0.1$ and $\epsilon_2 = 1.0$. Notice that (b) is the only model identical to (a).

the scaling of the extent parameters mentioned above. Figure 5.3 illustrates an example of this type of square/diamond equivalence class.

Notice that ϵ_1 is not involved in this type of equivalence class. The reason for this being that ϵ_1 controls the shape in two cross-sectional planes simultaneously: x–z and y–z. Using the superellipsoid description, one could never have a simultaneous square in both the x–z and the y–z planes being equivalent to a diamond in the x–z and the y–z planes. This is because cubes join at corners comprised of three edges, and diamonds are made up of corners that join four edges.

3. Multi-Modal Representation of Superellipsoid Models

Because more than one set of parameters could be used to describe the same superellipsoid model, it is best to represent each model by all of its possible equivalent forms. For this reason, we no longer limit our representation of a model to a single distribution, centered on the first minimum state settled into by the fitting procedure. We now represent each model by a multi-modal distribution, where each mode is centered on a possible canonical form.

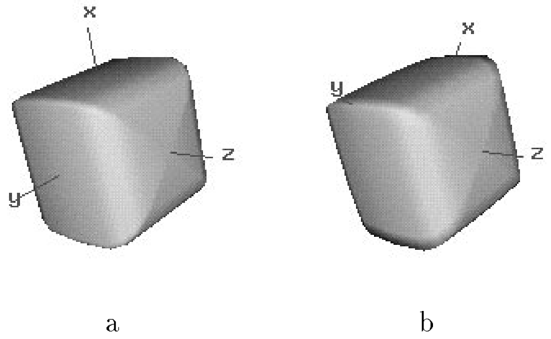
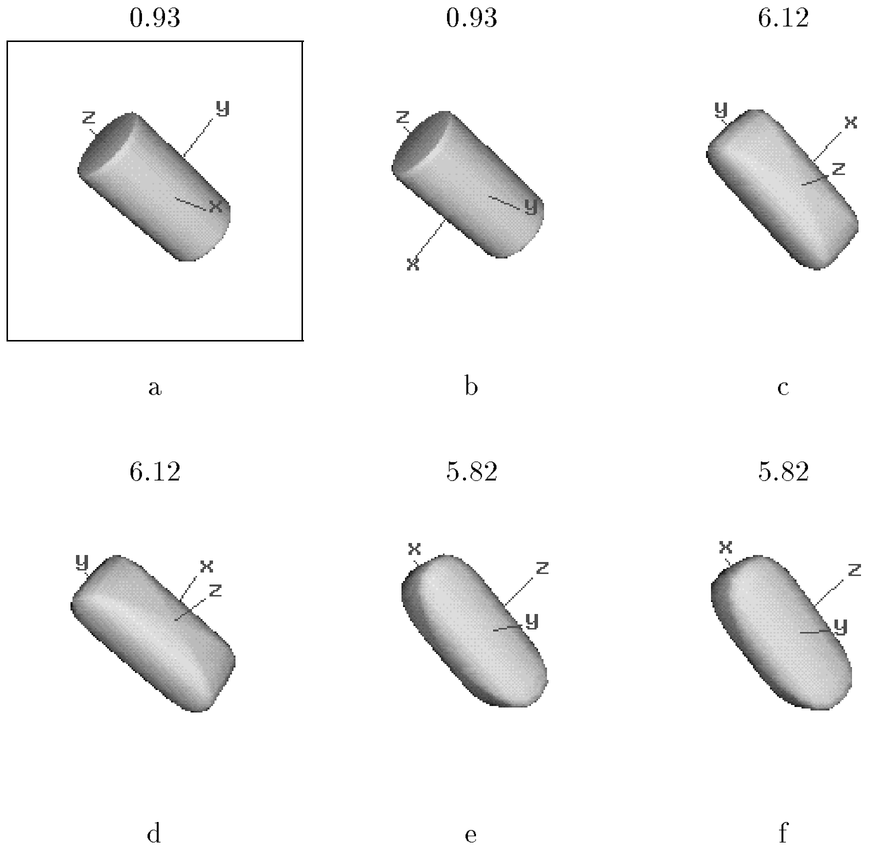


FIGURE 5.3. Square/diamond equivalences (a) Block with parameters: $a_x = 20$, $a_y = 20$, $a_z = 20$, $\epsilon_1 = 0.5$, $\epsilon_2 = 0.1$. (b) Block with parameters $a_x = 28.28$, $a_y = 28.28$, $a_z = 20$, $\epsilon_1 = 0.5$, $\epsilon_2 = 1.9$.

Since the most common degeneracies occur due to rotations, the primary focus is to ensure object representations free of rotational biases. This is ensured by enumerating, for each unidentified model, the six members of its rotational equivalence class. The first step is to fit the data to a superellipsoid model. Then, all six permutations of the extent parameters are found, resulting in six possible descriptions of the object. However, even if each of these parameter sets lies close to its appropriate minimum, we wish to find the *exact* minima corresponding to the possible rotational canonical forms. This includes accurate parameter sets as well as their corresponding covariances. Fine-tuning in this fashion is crucial in situations where discrimination between two objects is delicate. In order to attain this level of accuracy, the model is refit with each of the permuted parameters used as initial conditions. The results are six representations for the model based on all possible rotations.

However, the six representations do not necessarily produce identical model surfaces. As illustrated earlier, only models with equal shape parameters have six rotational equivalences (see Figure 5.2). Here, the results of fitting are models that strongly resemble the original, with different labelling of their axes. When the shape parameters are very different, the only surfaces that are identical are the two produced from rotations in the cross-sectional $x - y$ plane. The other four canonical models that result from fitting are different from the original. This is caused by attempting to force the fitting procedure to settle in minima that are not members of the rotational equivalence class. This leads to models that do not fit the data very well, and do not resemble the original. Figure 5.4 shows the six canonical forms of a cylinder. One can see that the only representation that is identical to the original is the one that has permuted the x and y axes. The other models are the results of permuting the axes when the shape parameters are not the equal. These no longer resemble the original.



Displayed above are the six canonical forms of a cylinder. The original model (a) is seen enclosed by a box. (b) is the model resulting from permuting the x and y axes and refitting, (c)–(f) are the results of refitting the model, with the other extent parameters permuted. Above each model is the residual error resulting from the fit.

FIGURE 5.4. The six canonical forms of a cylinder.

Since other equivalences exist for the superellipsoid, future work will concentrate on enumeration of all possible equivalences, each represented by a new mode in the normalized distribution of the model. Since these other equivalences occur less frequently, they are not included for now. As a result, recognition attempts still encounter some difficulties where these equivalences need to be taken into account.

4. Recognition of Multi-Modal Superellipsoid Models

Recognition of an unknown model represented by a multi-modal distribution is now performed. Here, a *belief vector* in a reference model is calculated by passing its single-mode distribution over the six-modal distribution of the unidentified object, and determining the belief in each mode. This is performed for each reference model. The unidentified object assumes the canonical form with the highest belief in one of the reference distributions. For the majority of the cases, this system would work well. This section will illustrate the problems that can arise with this strategy and will propose some practical solutions to these problems.

4.1. Reducing Misfit Problems. When calculating the canonical forms of a model, we permute the extent parameters, and send these as initial conditions to the fitting procedure. However, in cases where the shape parameters are not equal to each other, we are forcing inappropriate initial parameters onto the fitting procedure. This leads to higher degrees of misfits in some canonical forms. From Figure 5.4, one can see that those canonical forms that are not members of the rotational equivalence class do, in fact, produce much higher residual errors of fit. In these cases, there is a risk that the resulting distributions would fall closer to the wrong reference model's distribution than to any others. The results are false-positive identifications.

In order to reduce the number of incorrect identifications, we assign weights to the beliefs generated by each model based on the amount of misfit detected. These weights are inversely proportional to the residual error returned by the fitting process: Large errors produce small weights, decreasing all the beliefs produced by that mode. Small errors enhance the beliefs. The weight function decided on is:

$$W = \exp\left(-\frac{1}{2}\,\hat{\sigma}^2\right) \tag{29}$$

where $\hat{\sigma}^2$ represents an unbiased estimate of the sensor noise variance given by the current residual errors. In this fashion, little credibility is given to representations associated with large misfits.

As well, there are other ways in which misfit problems can be avoided. When fitting the data to a model, the fitting process can settle into different minima, depending on its starting point. This is especially true when collecting data from one viewing position, because the level of misfit is increased by the lack of constraint on the fitting. In order to ensure some level of consistency in the initial model representations, appropriate initial conditions are given to the fitting process. These starting points give the process a rough estimate of the shape of the object, as well as an acceptable pose (see (Ferrie, Lagarde & Whaite 1993)). This was done to reduce the level of misfit, and to lead the process towards a member of an appropriate rotational equivalence class. It is necessary to perform this step on the models used in training because these do not include all possible canonical forms.

4.2. Representation of the Reference Models. In the current scheme, each reference model is represented by a single-mode normalized distribution. The fitting procedure is given an appropriate starting point to ensure uniform canonical forms for the models involved in the training process. The fitting process creates a single distribution centered on the parameters at the closest minimum. Since the reference models are created from data collected from three views, the fitting procedure is well-constrained.

However, one the problems associated with using only a single mode distribution for the reference model is that, due to uncertainty during training, the system may choose a canonical form for an instance of that model differing from that of the mean. This outlier would bias the distribution of the model class. This would result in an inaccurate representation of the object, falsely diminishing its certainty in its parameters.

Ideally, one would want to represent the reference models by a six-modal normalized distribution, permitting the representation of all possible canonical forms. In this fashion, the recognition procedure would attempt to find the greatest overlap in multi-modal normal distributions. Multi-modal representation of the reference model is not employed due to the fact that training models, each represented by multi-modal distributions, is a difficult clustering problem not yet solved.

In this chapter, we have shown how to avoid the degeneracies associated with the superellipsoid model, by explicitly enumerating all equivalence classes for each model, and encompassing them into the model description. This lead to a multi-modal distribution for each model. We have also indicated how the recognition strategy described in Chapter 4 can be extended to include multi-modal superellipsoid models. In Chapter 7, we will illustrate that recognition experiments based on these representations prove successful.

CHAPTER 6

Informative Views and Active Recognition

1. Introduction

In earlier chapters (Chapters 3,4), we have described how one can cast the recognition problem into a probabilistic framework. We have shown how we can describe what we know about the world by representing all prior knowledge as probability density functions. As well, we have illustrated the way in which we can combine the information to obtain the solution in the form of a conditional probability density function, by application of a generalized inverse theory.

Now, consider an active agent charged with the task of roaming the environment in search of some particular object. It has an idea of what it is looking for, at least at some generic level, but resources are limited so it must act purposefully when carrying out its task (Aloimonos 1992). In particular, the agent needs to assess what it sees and quickly determine whether or not the information is useful so that it can evolve alternate strategies, the next place to look for example. Key to this requirement is the ability to make and quantify assertions while taking into account prior expectations about the environment. In this chapter we will show how the resulting belief distributions can be used to (i) assess the quality of a viewpoint on the basis of the assertions it generates and (ii) sequentially recognize an unknown object by accumulating evidence at the probabilistic level.

2. Determining Which Viewpoints are Informative

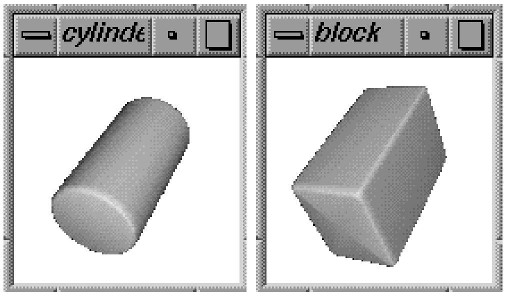
In Chapter 7, we will show that recognition based on complete information produces perfect results in all cases. Since complete information is not always available, and potentially expensive to acquire, recognition schemes based on single viewpoints are required. However, recognition based on one view will not prove to be consistently reliable. In fact, the degree of reliability depends upon the amount of information available. For example, some viewpoints capture enough of the unique characteristics of the object to sufficiently distinguish it from the others in the database. We will refer to these viewpoints as *informative viewpoints*. Other viewing positions, where it is impossible to say which object in the database the unknown is closest to, are called *uninformative viewpoints*. By determining if a viewpoint is informative or not, we can establish if further sampling is necessary to be able to recognize the object well.

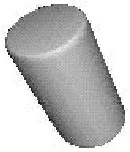
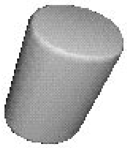
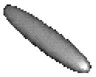
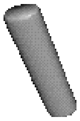
The question becomes: how can we use the inverse solution to distinguish between informative and uninformative viewpoints? We have shown an important result. Rather than establish an absolute identity for the unknown object, the method communicates the belief in each of the models in the database. Furthermore, uncertainty serves to condition prior expectations such that the shape of the resulting belief distribution can vary greatly. The results will indicate (Chapter 7) that the distribution becomes very delta-like as the interpretation tends towards certainty. In contrast, ambiguous or poor interpretations consistently tend towards very broad or flat distributions. We will exploit this characteristic to define the notion of an *informative viewpoint*, i.e. a view with a clear winner, in terms of a significantly higher belief in one model than the others. From these positions, the system is able to capture the attributes of the model that distinguish it from the others. The important contribution of this work is to be able to recognize these viewpoints, and use them in the determination of object identity.

We would also like to use the beliefs for the converse, i.e. to label a viewpoint as *uninformative*. This indicates that results from the current viewing position do not tell us much about the object's identity. This situation occurs when the unnormalized belief in each of the models is very low (or zero). Here, it is impossible to say which reference model the unknown might correspond to. This situation occurs when the distribution of the unknown model does not significantly overlap with any of the reference distributions. There are two possible reasons for this to occur. The first is the case where the distribution of the measured model is very wide due to large uncertainties in its parameters. The result is low beliefs in all the reference models in the database. This case occurs when scanning has occurred from a viewpoint where insufficient data was collected. The second case occurs when there is a breakdown in some of the prior assumptions. In this case, the issue is not one of insufficient data. Here, the parameters determined from that particular viewpoint differ significantly from any of the models in the database. The resulting distribution could actually be quite sharp, but simply does not overlap with any of the reference model distributions. In this case, it could be that the linearity assumption breaks down, implying that perhaps the assumption of a normal distribution is not valid. Zero belief cases exist when the values of the a posteriori probability density functions are extremely low. Due to numerical underflow, the procedure produces beliefs of zero for each of the reference models.

Figure 6.1 illustrates the difference between informative and uninformative viewpoints for the case of a cylinder. Here, one can see that the system is able to distinguish the cylinder from a block with great ease, if the cylinder is measured from an informative viewpoint. However, if measured from an uninformative viewpoint, there is little confidence in either model. In this case, the beliefs are in fact below the numerical precision of the system, and therefore become zeros.

Database Models



Measured Model	View 1	View 2	View 3	View 4
				
Belief in cylinder	2.237	0.009181	0.0	0.0
Belief in block	0.0	0.0	0.0	0.0
	a) Informative		b) Uninformative	

At the top of this figure are the two reference models in the data base: the cylinder and the square block. Beneath these are measured models of the *cylinder* obtained after scanning its surface from 4 different viewing positions. Below each model one can find the unnormalized belief distributions obtained when attempting to recognize each of the measured models.

FIGURE 6.1. (a) Informative and (b) uninformative views of a cylinder.

The problem of distinguishing between the two kinds of states becomes one of determining the threshold below which one can safely state that the beliefs are in fact insignificant. It is obvious that cases where the beliefs in all the models are zero are uninformative. However, this threshold depends on the numerical precision of the system. In this sense, it is chosen externally (and is, therefore, a random cutoff point). We therefore feel justified in raising this threshold to one that excludes other low confidence states. The expectation is that this will eliminate false positive states, as they are thought to occur with low belief. (We will establish this empirically in Chapter 7.) One can determine this cutoff point empirically, by observing the belief distributions from different viewpoints, and noting if there is a clear division between the clear winner states and the low confidence states. A bi-modal distribution would indicate that an application of a predefined threshold can easily distinguish between these states. In Chapter 7, we will illustrate the results of plotting the belief distributions resulting from recognizing six objects from different viewing positions.

There are at least two applications for a method that can assess the quality of the information from a particular viewpoint. First, in the case of an active observer, viewpoints can be chosen so as to maximize the distribution associated with an object of interest. This does not specify *how* to choose an informative viewpoint[1], but can be used as a figure of merit for a particular choice. Second, in the case of an off-line planner, it is often advantageous to be able to pre-compute a set of characteristic views to aid in recognition (Koenderink 1976, Koenderink 1979, Sripradisvarakul & Jain 1989, Eggert & Bowyer 1989, Eggert et al. 1992, Kriegman & Ponce 1989, Bowyer & Dyer 1990). A good strategy here would be to select the n best views of an object ranked according to its belief distribution.

3. Incremental Recognition

Provided that the low belief states have been identified, we wish to make a statement about the remaining beliefs. Even though the majority of the cases can be clearly divided into informative and uninformative states, there are still ambiguous cases where a "significant" belief in more than one model exists. Because of these situations, it becomes apparent that evidence from more than one viewpoint is needed. But at what level of representation should this evidence be accumulated? The autonomous exploration procedure that we use to generate the set of database models, for example, sequentially constructs a complete 3D representation at the level of surface geometry (Whaite & Ferrie 1994). One could follow a

[1] Strategies for gaze planning are usually operationally defined (Whaite & Ferrie 1991, Whaite & Ferrie 1994).

similar approach at the recognition phase, i.e. recalculate each belief distribution as the explorer adds new data to its representation of the unknown object. Unfortunately this would be computationally prohibitive, largely due to the expense of data fusion (Soucy 1992). A better approach would be to process each view independently and avoid the fusion problem at the data level by seeking instead to combine information at the level of the belief distribution. In Chapter 3, the inverse theory outlined how to do this by defining the operation of conjunction of states of information, i.e. the belief distributions. That is, we denote belief distributions corresponding to each model hypothesis, $\mathcal{H}_i$, given the parameters of the unknown model, $\mathcal{M}$, computed from the measurement, D_j, by $P(\mathcal{H}_i|\mathcal{M}_{D_j})$. Then, given two data sets D_j and D_{j+1} corresponding to different viewpoints we seek a conjunction of $P(\mathcal{H}_i|\mathcal{M}_{D_j})$ and $P(\mathcal{H}_i|\mathcal{M}_{D_{j+1}})$ that is equivalent to $P(\mathcal{H}_i|\mathcal{M}_{D_j+D_{j+1}})$. An active agent would then gather sufficient evidence in this fashion until the composite belief distribution associated with a particular hypothesis exceeds a predefined level of acceptability.

Although the theory formally defines conjunction, such an operation requires knowing how a change in viewpoint conditions the respective belief distributions, as they are not normalized with respect to a global frame of reference. (As we have seen in Chapter 5, the normalizing factor is some unknown function of viewpoint, and is difficult to obtain analytically.) As a result, relative values between the views are meaningless. Hence, it becomes difficult to match a belief of 500, for example, from one view, with a value of 50 from another. Each of these values may reflect the strongest possible belief from their respective views, however it is difficult to compare them in a sensible fashion. As well, in situations where there is a belief of 50 in one model and 40 in another, it becomes impossible to establish a clear winner.

For this reason, we have chosen not to choose a "winner" in ambiguous situations, and state that all positive beliefs indicate equally likely hypotheses. We illustrate this philosophy by binarizing the conditional probability density function values at each view, such that all beliefs above the threshold become ones. In this fashion, we have divided the possible results to include:

(i) *Informative states*: states with one clear winner (a single positive value).

(ii) *Uninformative states*: states without a clear winner. This includes:

a) *Ambiguous states*: states with more than one possible winner (more than one single positive value).

b) *Undetermined states*: states with no winners (all zero values).

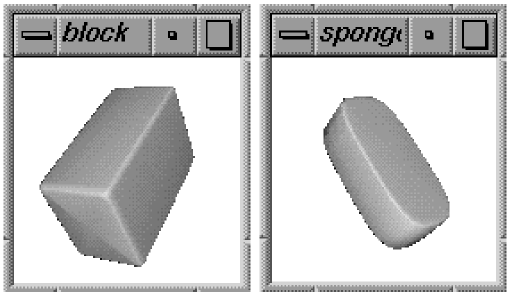
It is important to note that ambiguous states are, in fact, undetermined states that lie above the chosen threshold. In theory, careful choice of cutoff level should eliminate these states as well (without eliminating a large number of informative states). Figure 6.2 illustrates these different states in the case of a square block. Here, the system is asked to identify a square block from different views, and correctly distinguish it from a similar rounder one. This example indicates that the results match human intuition. The clear winners, or informative states, in Figure 6.2a indicate that the system is able to identify the block despite wide variations in its three dimensions. The ambiguous cases (Figure 6.2b) occur when the resulting models are rounder in shape. Here, the system has trouble differentiating between the models. In fact, these models resemble the rounded block more than the square one. In the third case (Figure 6.2c), the system does not have significant belief in any of the models. Intuitively, one can see that these models are not similar to either reference model.

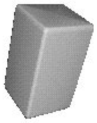
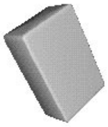
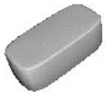
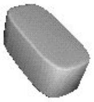
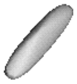
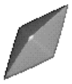
Using this method of representation, rather than base conclusions on maximum likelihood methods from independent viewpoints, methods that combine evidence from single viewpoints would consider all models whose beliefs are above a threshold to be equally significant. In accordance with Marr's "Principle of Least Commitment" (Marr 1982), all possible hypotheses, rather than just one are communicated to the external processes.

By normalizing our confidence values in this manner, combining them from different viewpoints becomes straightforward. Should the maximum likelihood hypothesis prevail in a largely view-invariant manner, then after a sequence of trials, a robust interpretation can be made by tabulating the votes for each one, represented by the binarized beliefs, and picking the hypothesis with the highest score. In this fashion, a clear winner should emerge. As well, the confidence in the incorrect models should become insignificant. In Chapter 7, we will verify this empirically by attempting to recognize a series of real objects from sequential viewpoints. We will also show that the view-invariance is maximized by applying the threshold to filter out the uninformative hypotheses.

Figure 6.3 illustrates an attempt at sequentially recognizing the square block at 40° increments. As in the previous example, the square and round blocks are used as reference models. The raw beliefs are binarized by imposing a threshold of 10^{-13}. Notice that the ambiguous case quickly becomes insignificant with the increase of evidence in the correct model. After only 9 iterations, the clear winner emerges, casting all doubt aside.

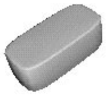
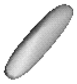
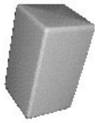
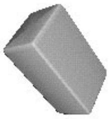
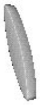
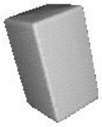
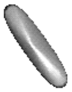
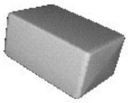
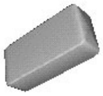
In the next chapter, we will test the recognition procedure on real single-part objects, for models created from complete (3D) data and from partial (2D) data. The possibility of applying a threshold to distinguish between informative and uninformative viewpoints will be tested, by observing the belief distributions resulting from recognition from different viewpoints. Also, Sequential recognition experiments will be performed. Finally, the ability of the system to recognize parts of articulated models from single viewpoints will be assessed. Effects of applying an external threshold to eliminate uninformative viewpoint hypotheses will be seen as well.



Measured Model	Belief in Block		Belief in Round Block	
	Unnormalized	Binarized	Unnormalized	Binarized
	0.2	1	0	0
	0.007	1	0	0
	2.0×10^{-13}	1	5.8×10^{-6}	1
	3.4×10^{-13}	1	0.002	1
	0	0	0	0
	0	0	0	0

Above are the two reference models: a block and a rounded block. In the left column of the table are the models of the block measured from informative (first pair), ambiguous (middle pair) and undetermined (last pair) viewpoints. To their right, one can find the unnormalized, and binarized (threshold of 10^{-13}) belief distributions obtained when attempting to recognize each of the measured models.

FIGURE 6.2. Informative, ambiguous, and undetermined States for the Block.

View Angle	Measured Model	Belief in Block		Belief in Round Block	
		Unnormalized	Binarized	Unnormalized	Binarized
0°		2.0×10^{-13}	1	5.8×10^{-6}	1
40°		0	0	0	0
80°		0.2	1	0	0
120°		0.03	1	0	0
160°		0	0	0	0
200°		0.1	1	0	0
240°		0	0	0	0
280°		0.03	1	0	0
320°		0.001	1	0	0
		Final Score	6		1

Displayed above are the 9 models resulting from sequentially measuring the square block at 40° increments. From left to right, one can see the viewing angle, the measured model, the unnormalized and binarized (threshold of 10^{-13}) belief distribution resulting from attempting to recognize each of the measured models. The final distribution is the histogram of the binarized distributions.

FIGURE 6.3. Incremental recognition of a block.

CHAPTER 7

Experimentation and Results

1. Introduction

In the previous chapters, we have introduced the inverse theory, and indicated how it can be used within the context of a part recognition problem. As well, we have illustrated how the results can be used to assess the quality of the information from a particular viewpoint, and an incremental recognition scheme was proposed. Solutions to problems with the superellipsoid model were presented in order to be able to use this volumetric model as an object descriptor for recognition.

In order to test the proposed methodology on real objects, several experiments are performed. Section 2 begins with the description of the system used to acquire the object descriptions. Section 3 describes the first set of experiments which tested the algorithm on several single part objects. Maximum likelihood (or Winner-takes-all) schemes were tested on models fit to data acquired all around the object (*complete* or 3D data). In addition, the tests were performed on models generated by data acquired from one viewpoint only (*partial* or 2D data). The results of these tests indicated the possibility of distinguishing between informative and uninformative viewpoints by application of an external threshold. Experiments using an incremental recognition scheme were performed, whereby evidence in the form of belief distributions was accumulated from different viewpoints sequentially. Finally, in Section 4, both single-view and incremental recognition of parts of articulated models was tested. This provided the basis for a multiple-part object recognition strategy.

2. System Overview

Throughout the experiments, object representations were created through the bottom-up system developed by the 3D Vision Group at CIM. In the system we have constructed, articulated, volumetric models are created by successive probes of a laser-rangefinder through a process of *autonomous exploration* (Whaite & Ferrie 1991, Whaite & Ferrie 1993*b*, Whaite

& Ferrie 1994). The flowchart for the bottom-up stages for the pencil sharpener can be found in Figure 7.1. It corresponds to the classical model of bottom-up vision in which sensor data are transformed into various levels of representation though successive stages of processing (Ferrie & Lagarde 1989). The additional feature is the inclusion of feedback from the fitting procedure, which is used to determine the new gaze position that will reduce model uncertainty. Because object recognition represents the highest level of processing, it relies not only on its discriminating power, but on all the lower level processes that contribute to the stability and accuracy of the object representation needed for recognition. This section will describe the system that generated the volumetric models used by the recognition scheme.

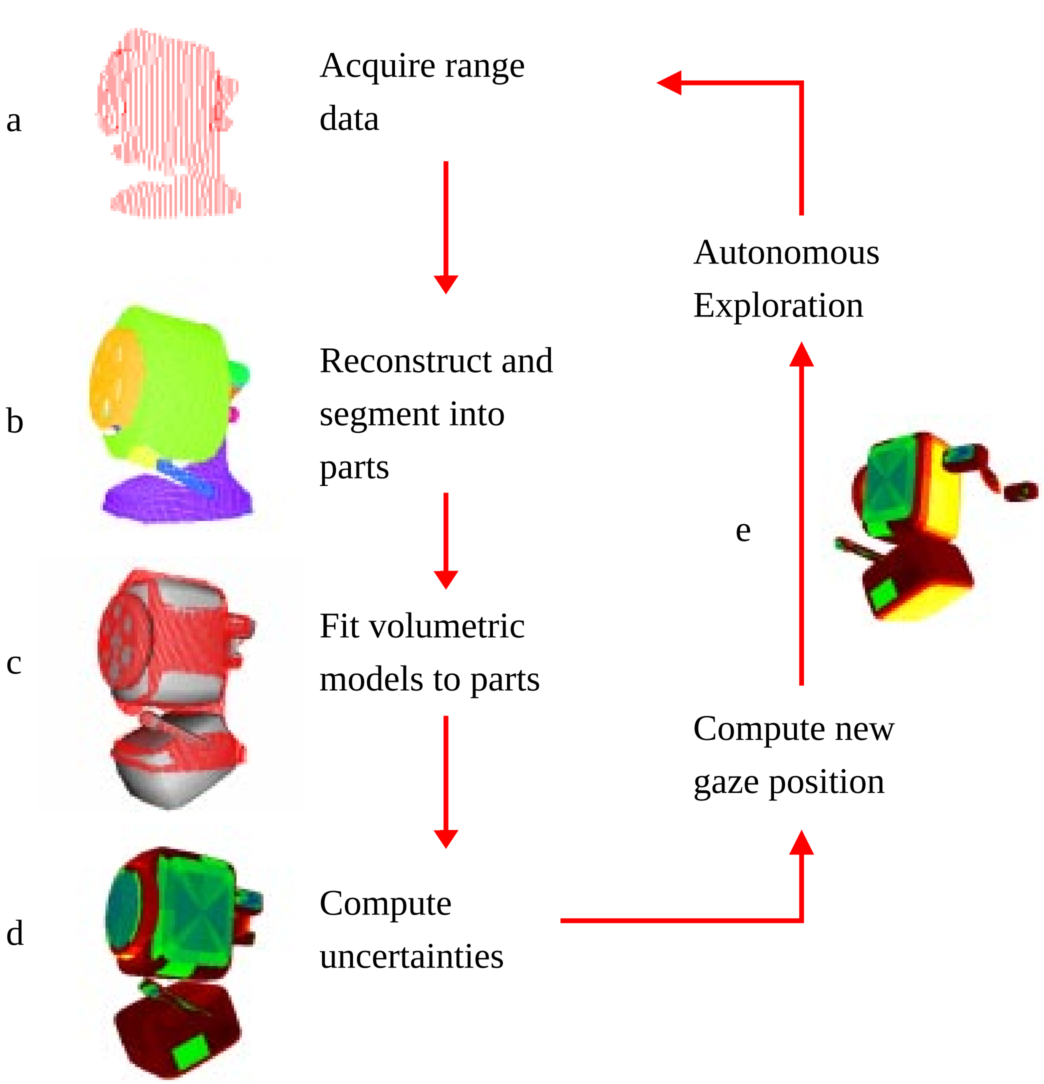
2.1. Data Acquisition. Objects are scanned using a 2-axis laser rangefinder mounted on the end of an inverted PUMA robot arm. The scanner is capable of scanning at a range of 1 *meter* (Soucy & Ferrie 1992). Its field of view is approximately 40° in the x direction, and 28° in the y direction. Each of these spans can be divided into at most 256 positions. The precision of the scanner is approximately 1 *mm* at a distance of 1 *meter*, and improves non-linearly as the distance decreases. In the experiments described, the density of scanning is such that each pixel of an $85 \times 85\ pixel^2$ image represents $3mm^2$.

In order to obtain calibrated data, i.e. real x, y, and z coordinates in the camera frame (in *mm*), a calibration procedure is applied. Here, look-up tables are created, providing the translation from points in the image to the x and y coordinates in the camera frame.

In addition, a set of precision stages, controlled by stepper motors, is used to expose different faces of the object to the laser rangefinder. The rotary table permits four degrees of freedom (two rotations, and two translations). The theoretical precision obtained is approximately 79 steps per *mm* in displacement in x and in y, 100 steps per degree for the rotation about the $z - axis$ and 0.56 step per degree for the rotation about the $x - axis$. However in reality, the precision is slightly lower if one were to take into account the mechanical play of the gears (i.e. backlash).

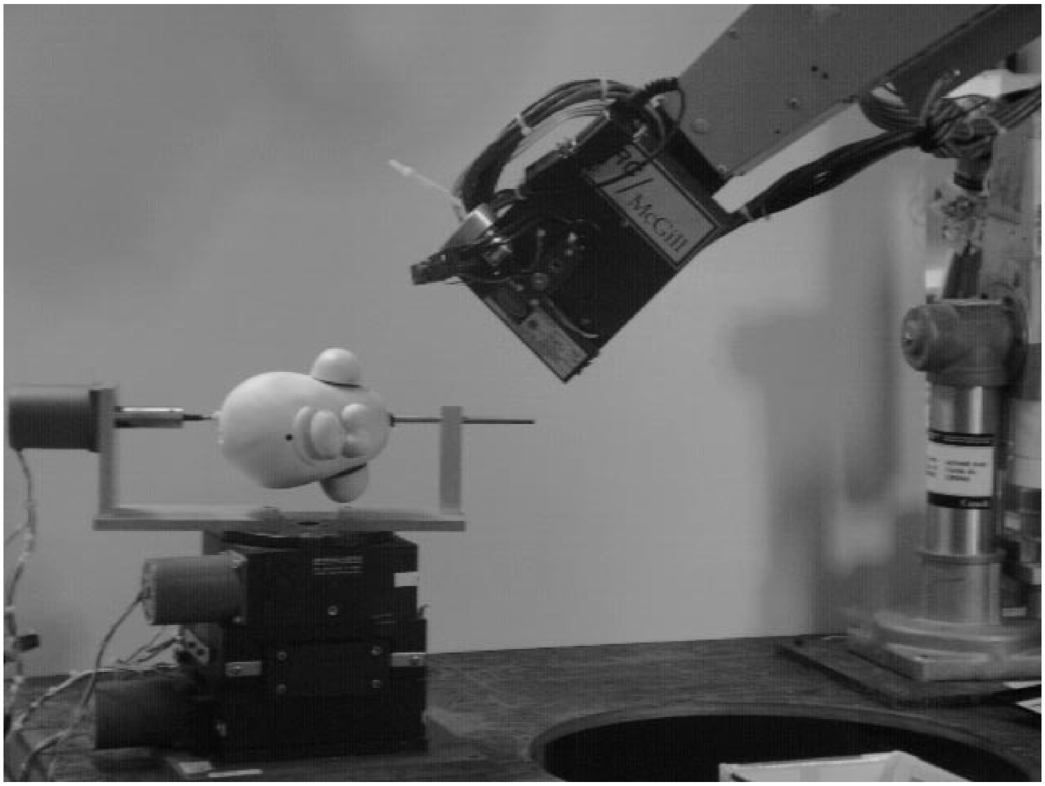
Using this set-up, different views of an object are obtained by keeping the scanner fixed and by moving the stages to which the object is attached. The data acquisition set-up can be seen in Figure 7.2. An example illustrating the data lines resulting from using the set-up to scan the pencil sharpener can be seen in Figure 7.1a.

2.2. Surface Reconstruction. The purpose of this stage is to transform the discrete range data into piecewise smooth representations of the surface (Ferrie, Mathur & Soucy


a
Acquire range data
b
Reconstruct and segment into parts
c
Fit volumetric models to parts
d
Compute uncertainties
Compute new gaze position
e
Autonomous Exploration

Here we see the classical bottom-up strategy used to obtain a parametric model of an object in the scene. Notice that the loop is closed with the addition of feedback which uses the parametric uncertainty to choose a new gaze position that will reduce model ambiguity. The process is referred to as *autonomous exploration*. See text for details.

FIGURE 7.1. Flowchart of the bottom-up system.



The set-up includes a laser rangefinder mounted on the end-effector of an inverted PUMA manipulator. The object itself is placed on a rotary table, permitting four DOFs.

FIGURE 7.2. Set-up used to scan objects.

1993). It consists of a diffusion algorithm based on surface curvature properties. The effect of the operator is to remove noise and to smooth out convex surface regions. Points along a boundary, marked by negative local minima and concave discontinuities are left undisturbed. The diffusion algorithm results in bringing out the convex surface patches in the image (Ferrie, Lagarde & Whaite 1993, Lagarde 1989, Lejeune & Ferrie 1993).

2.3. Part Decomposition. The reconstructed surface is segmented into regions corresponding to object parts. This is done by growing the labelled surface regions until they reach the previously labelled boundary points. Regions are merged using a relaxation labelling network that ensures resulting boundary contours that are consistent with predefined boundary points (Ferrie, Lagarde & Whaite 1993, Lagarde 1989, Lejeune & Ferrie 1993).

Figure 7.1b illustrates the surface patches resulting from reconstructing and segmenting the surface of the sharpener. The different colors refer to different part regions.

2.4. Data Fusion. If data are acquired from various viewing positions, they are merged using a scheme which calculates the correspondence between surfaces from neighbouring views. The motion parameters between views are calculated under the assumption that curvature is preserved. In this fashion, local motion estimates map data points from one frame to another. In order to constrain the local match, global motion consistency is enforced, where variations in velocity between frames are assumed to be piecewise-smooth. Therefore, choosing the motion parameters becomes a minimization problem, where the differences in relative position and orientation between points are minimized. In this fashion, the algorithm is tolerant of local errors in correspondence. In addition, it serves to smooth out local noise, and blend neighbouring surface patches (Soucy & Ferrie 1992, Soucy 1992).

2.5. Volumetric Modelling. At the highest level of abstraction, a volumetric model is fit to each part region. Descriptors of this nature provide the basis for the characterization of uncertainty. As well they maintain correspondence at the part level. Most importantly, they describe general shape properties, which is useful for the recognition task.

For the purposes of this thesis, the model chosen was the superellipsoid model (Solina & Bajcsy 1990). Calculating the parameters $\mathbf{a}$ is performed using an iterative, least squares minimization technique, the Levenburg-Marquardt algorithm (Luenberger 1984, Press, Flannery, Teukolsky & Vetterling 1988, Whaite & Ferrie 1991). Here, a metric $D(\mathbf{x}, \mathbf{a})$ is defined that measures the distance between each data point $\mathbf{x}$ and the superellipsoid surface described by the parameters $\mathbf{a}$. From an initial guess, the parameters are changed incrementally in a steepest descent manner to minimize the squared sum

$$\chi^2(\mathbf{a}) = \sum_{i=1}^{N} \frac{D^2(\mathbf{x}_i, \mathbf{a})}{\sigma_i^2} \tag{30}$$

of the metric over all data points. Each distance is weighted by its error, σ_i^2, in order to increase the importance of the low error terms. The procedure iterates until there is a negligeable improvement in the squared error. Currently, the five superellipsoid parameters describing object size and shape, as well as their associated covariances, are used as part descriptors for object recognition.

Figure 7.1c illustrates the results of fitting superellipsoid models to each of the part regions in Figure 7.1b.

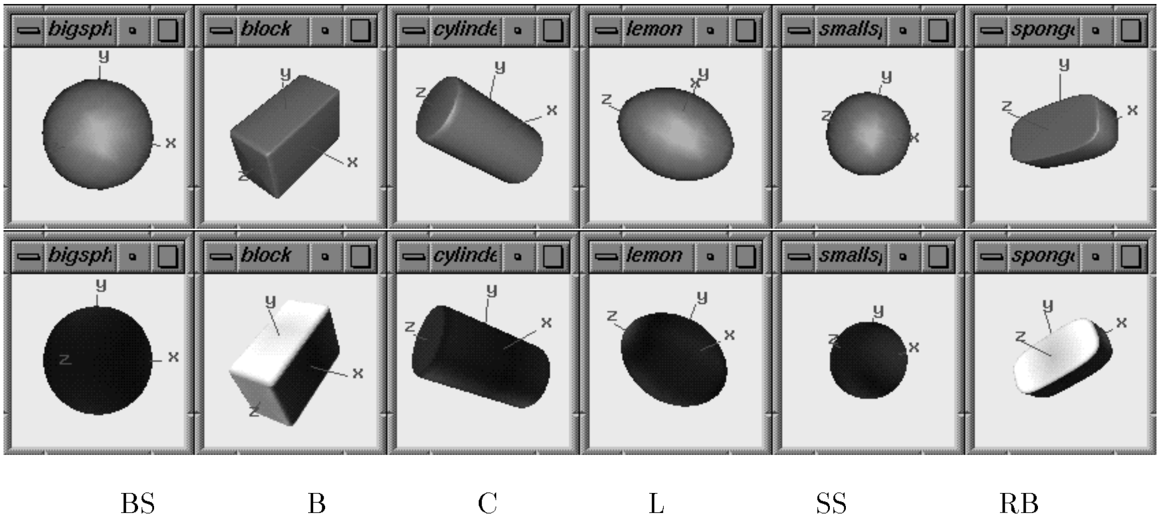
2.6. Feedback. Because of the noise in the model, and because the data are often incompletely sampled, e.g. only one side of the model is visible from a single viewpoint, the parameters will often be under-constrained and exhibit large estimation errors. In order to reduce the error, the system calculates a new gaze position where additional data can be collected. This is accomplished by using the estimated model as a predictor of the surfaces in the scene. The error is quantified in terms of an interval around each point on the predicted surface. We refer to this interval as the *surface prediction error interval*, which refers to an "error bar" protruding from a point on the estimated model's surface. The interval is coded such that "hotter" colors (such as yellow, or red) represent higher uncertainty in surface positions as predicted by the model. Figure 7.1d illustrates this color coding for the sharpener. The resulting prediction can extend beyond the visible surfaces and can thus serve as a basis for planning the next gaze direction. This is accomplished by directing the scanner to the viewpoint corresponding to the highest uncertainty of prediction. This can be seen in Figure 7.1e, where the scanner is moved to the back of the sharpener where the uncertainty is greatest. It has been shown that updating the model parameters with the additional data obtained from the new view will minimize the determinant of the parameter covariances. This process is referred to as *autonomous exploration* (Whaite & Ferrie 1991, Whaite & Ferrie 1993*b*, Whaite & Ferrie 1993*c*, Whaite & Ferrie 1994).

3. Single-Part Object Recognition

Having established the means to obtain object descriptions, the purpose of the first set of experiments was to test the recognition procedure on a series of real objects. In order to focus on this task, and to ensure results that were free of errors from the segmentation process, these experiments included only single-part objects[1]. Several experiments were performed. The first tested the ability of the system to recognize based on complete, 3D information. The second set tested the more practical problem of recognition from single viewpoints. Here, the system's ability to distinguish informative from uninformative viewpoints was assessed, by application of an external threshold. Finally, an incremental recognition scheme was invoked.

With this in mind, six objects were chosen for these experiments: two spheres (rad = $20mm$, rad = $25mm$), a block, a cylinder, a lemon, and a block with rounded edges. The objects were selected because they consisted of single parts that conformed well to superellipsoids. They varied in size and shape, so as not to be clustered together too tightly in five-dimensional feature space. However, their distributions overlapped sufficiently in several dimensions so that the recognition procedure was challenged in its discrimination task.

[1] In Section 4, we will examine the capabilities of the system in recognizing parts of articulated models.



Displayed above are the reference objects that result from training on complete surface data: a big sphere (BS), a block (B), a cylinder (C), a lemon (L), a smaller sphere (SS), and a rounded block (RB). Below, the same models are shaded according to the projection of parameter uncertainties into 3D space. White reflects large uncertainties, and black indicates parameters that are tightly constrained. For example, the light face of the block shows that the y size parameter is more uncertain than the x.

FIGURE 7.3. Six representatives that result from training.

Training (see Section 2) automatically produced object class representatives, by measuring the object numerous times. Each individual model was created by scanning the object from several views using a laser range-finder, then a superellipsoid model was fit to the data, and the resulting parameters stored (see previous section). For the purposes of creating a stable database for recognition, it was established that three views of each object, 120° apart were sufficient to constrain the fitting procedure. Each sample was scanned from a random scanning position, producing 24 samples of each object. Figure 7.3 illustrates the six representative models of each object that result from training.

For all the experiments, the model of the unidentified object was created using the bottom-up system described in the previous section. Whether data were collected from one view or from several views, in order to use the resulting superellipsoid model as a descriptor for recognition, the system had to calculate the six possible equivalence classes corresponding to it (as discussed in Chapter 5). These parameter sets were incorporated into the overall model by representing the object with a multi-modal distribution. During the matching stage, the system then chose the representation from the equivalence class that had the highest belief in one of the reference models.

3.1. Matching Using Complete Information. In the first experiment, recognition was performed using an unknown model computed from a sequence of views covering the visible surfaces of an unknown object. The intent of this experiment was to validate the recognition procedure against models produced by the autonomous exploration process on running to completion (Whaite & Ferrie 1991). Twenty-four samples of each object, each scanned from three different viewpoints, were presented to test the invariance of recognition against variations in sampling and viewpoint. Using maximum likelihood as the basis for recognition, i.e. choosing the model with the highest confidence value, the results shown in Figure 7.4 were obtained.

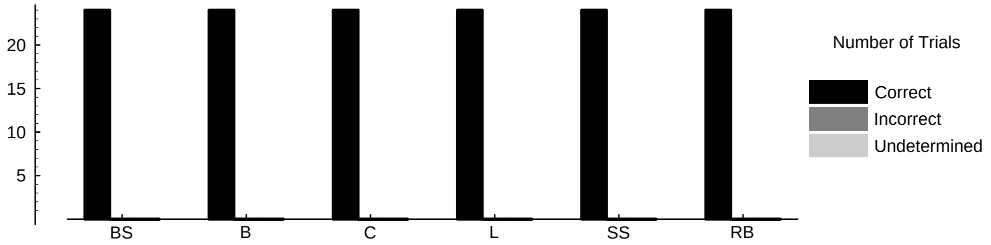


FIGURE 7.4. Matching samples taken from multiple viewpoints.

The results indicate that the system can successfully recognize an instance of any object in the database with perfect results, provided that its surfaces are accessible, independently of viewpoint and sampling order. In addition, the identifications are made with a high degree of certainty. This is to be expected given that the probability density functions of each of the unidentified objects exhibit small variations in parameter space due to the relatively complete information available. Training produces reference models that are also "delta-like"and well separated from each other. The distribution of the unidentified object would necessarily overlap that of the correct reference model much more than the others. Examples of the non-normalized belief distributions of the lemon and block can be found in Table 1.

Examination of the resultant beliefs shows that complete information allows the system to correctly identify objects with a high degree of certainty. The high beliefs reflect the fact that both the measurement distributions and the reference model distributions are "delta-like" and close together.

3.2. Matching Using Partial Information. Since complete information is not always available (and potentially expensive to acquire), a more realistic test would be to determine the parameters of an unknown model from partial information. In the limit this would consist of attempting to base recognition on data acquired from a single viewpoint and would clearly violate the multiple-view assumptions implicit in the training process.

Trial	BS	B	C	L	SS	RB
1	0	0	0	5.11	0	0
2	0	0	0	6.53	0	0
3	0	0	0	12.66	0	0
4	0	0	0	70.70	0	0
5	0	0	0	42.32	0	0
6	0	0	0	27.13	0	0

a) Belief distributions of the lemon

Trial	BS	B	C	L	SS	RB
1	0	6.09	0	0	0	0
2	0	6.24	0	0	0	0
3	0	9.87	0	0	0	0
4	0	1.58	0	0	0	0
5	0	15.21	0	0	0	0
6	0	11.67	0	0	0	0

b) Belief distributions of the block

TABLE 1. Results of several iterations of recognition of a)lemon and b)block viewed from multiple viewpoints.

Furthermore, it has been shown elsewhere that the resulting model parameters would be inherently less stable (Whaite & Ferrie 1991). However, should the procedure still retain some of its earlier selectivity — as evidenced by a low degree of false positive matches — then an incremental recognition procedure becomes a possibility. This would involve accumulating evidence from the belief distributions of sequential viewpoints until a clear winner emerges.

In this second set of experiments, recognition was performed on thirty-six single-view samples of each object. Here, data were collected at 40° intervals along 4 different great circle routes. The same methodology as in the first experiment was applied in the recognition of the unknown model parameters. The results obtained are shown in Figure 7.5.

As expected, recognition based on partial information is less certain than in the previous case where the complete surfaces of the unknown object were accessible. Here, undetermined states exist in situations where the unnormalized values of the posterior probability density

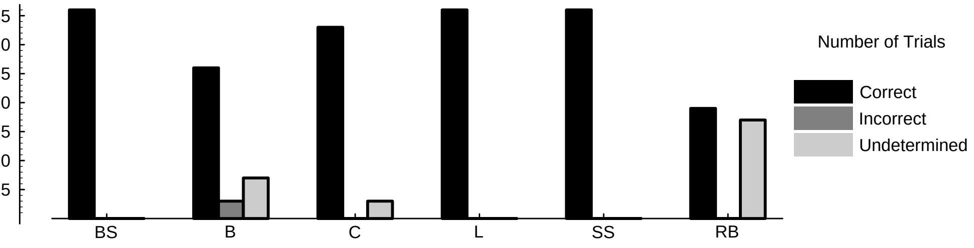
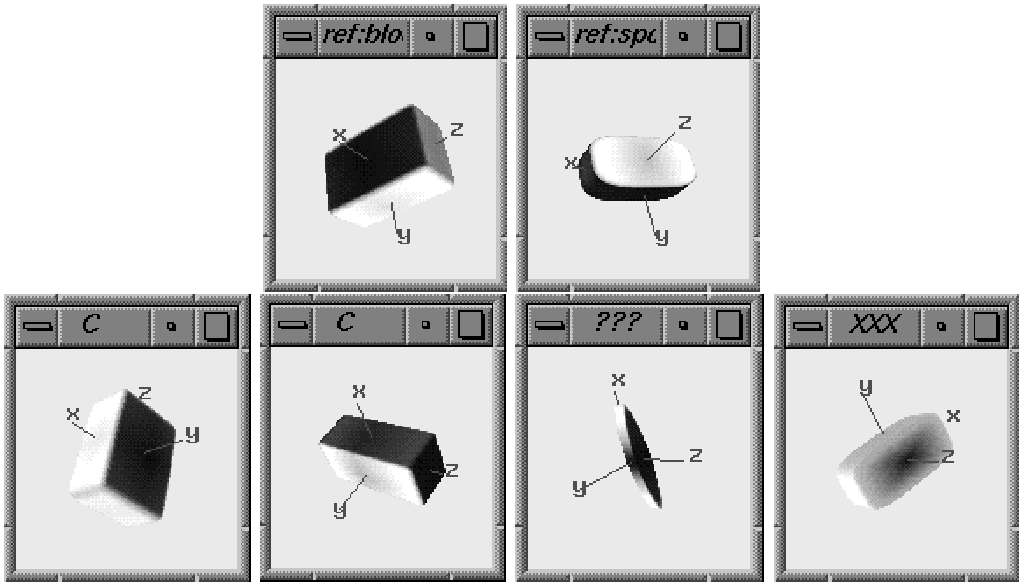


FIGURE 7.5. Matching samples taken from single viewpoints.

functions are extremely low (on the order of 10^{-60}). Due to numerical underflow, the procedure produces beliefs of zero for each of the reference models. We refer to viewpoints such as these, that do not tell us much about the object's identity, as *uninformative* (see Chapter 6).



In the top boxes are the square block and rounded block reference models. Below these are four different attempts at recognizing the square block from different viewing positions. In each case the model is compared to the each of the six references in turn, and beliefs in each are computed. Above each model one can see the result of running a maximum likelihood algorithm on the results. *C* indicates a correct recognition, *???* indicates an undetermined state, and *XXX* refers to a false recognition. Here, the system identifies the square block as being the rounded one. The objects are shaded according to their uncertainties (see figure 7.3).

FIGURE 7.6. Examples of recognition of the block from single views.

Figure 7.6 shows some specific examples of recognition attempts on the block from different viewing positions. In the first two cases, the procedure correctly identified the objects as corresponding to the block despite wide fluctuations in their size parameters. This is due to the fact that the models encompass the uncertainties corresponding to these parameters in their representations. The reference model also learned of these possible variations during training, incorporating them into its representation. Therefore, the distributions were close enough to that of the reference block to make a correct identification. This reinforces the hypothesis that objects need not be represented by extremely accurate descriptions. Rough size and shape representations are sufficient as long as the reference object has learned about these possible fluctuations in the training stage.

In the third case, the system could not identify the object as being any of the known models. In this case, this model does not visually resemble any of the references in size or shape. This is a situation where there is insufficient data from that viewpoint to produce a good model of the object. Further sampling of the object should provide better results.

In the final case, the system incorrectly identified the block as being the rounded block. (As well, the model is visually closer to the rounded block.) The reason for the match is that, although the reference block is not very certain about all of its size parameters, as indicated by the white shading on its sides, it is quite certain about its shape parameters. This is indicated by the black shading around the block reference model's edges. Therefore, measurements that are rounded in shape do not sufficiently overlap in its distribution. In this case, despite the high uncertainty in the parameters of the unknown model (causing its distribution to be quite flat), there was sufficient overlap in the distribution of the reference rounded block to cause a false identification.

Table 2 shows the belief distributions resulting from incremental attempts at recognizing the lemon and the block. The data were collected from single views at 40° intervals in an equatorial plane. One can see that the beliefs are considerably weaker than in Table 1 where complete information was used. The first iteration in the recognition of the block produced a false-positive identification. In this case, the system identified the block as being the rounded block, despite the fact that the resulting distribution overlapped with the distribution of the reference block as well. The belief in both models was quite low, indicating that the system is quite uncertain about the identification. In fact, in many cases, a false-positive identification is associated with low beliefs. This suggests that if the threshold for undetermined states were raised, the incorrect identifications would become undetermined states.

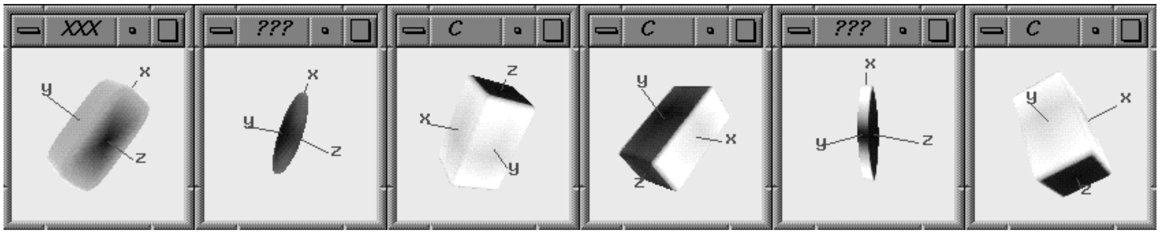
In order to justify raising this threshold, the beliefs resulting from the experiment described above were plotted on a logarithmic scale graph. The expectation in observing these results was that the scatter of the beliefs was bi-modal. This would imply that a distinct separation between informative and uninformative cases exists, permitting the

Viewpoint	BS	B	C	L	SS	RB
0°	0	0	0	2.97×10^{-21}	0	0
40°	0	0	0	6.93×10^{-15}	0	0
80°	0	0	0	0.18	0	0
120°	0	0	0	2.44×10^{-5}	0	0
160°	0	0	0	8.07×10^{-3}	0	0
200°	0	0	0	3.38×10^{-4}	0	0
240°	0	0	0	1.10×10^{-16}	0	0
280°	0	0	0	0.31	0	0

a) Belief distributions of the lemon

Viewpoint	BS	B	C	L	SS	RB
0°	0	4.00×10^{-13}	0	0	0	1.16×10^{-5}
40°	0	0	0	0	0	0
80°	0	0.33	0	0	0	0
120°	0	0.05	0	0	0	0
160°	0	0	0	0	0	0
200°	0	0.21	0	0	0	0
240°	0	0	0	0	0	0
280°	0	0.05	0	0	0	0

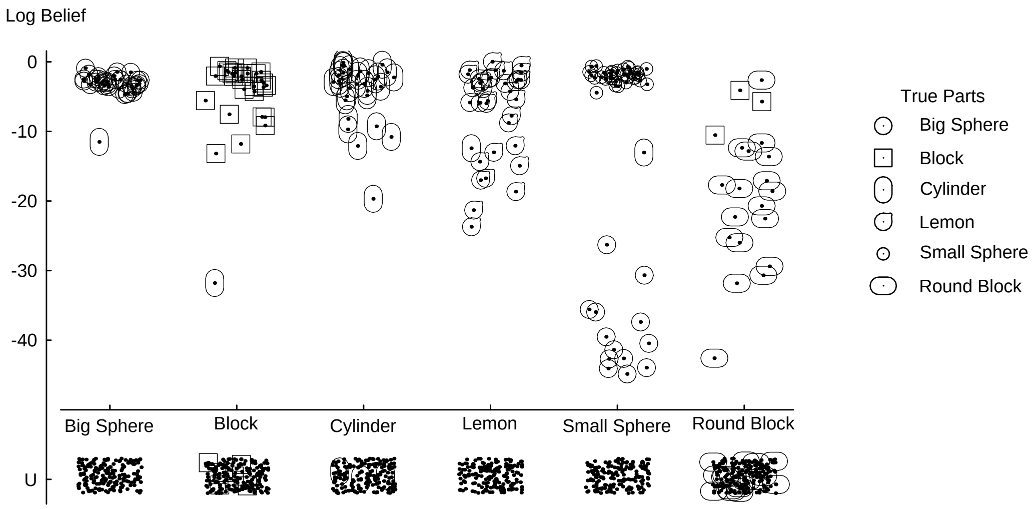
b) Belief distributions of the block



Displayed above are the first six attempts at successively recognizing the block at 40° increments. Shading is in accordance with parameter uncertainties (see figure 7.3). The results of running a maximum likelihood algorithm are found above each box (see figure 7.6).

TABLE 2. Results of incremental recognition of a)lemon and b)block viewed from 40° single viewpoints.

application of a threshold to distinguish between the two. The results can be found in the plot in Figure 7.7.



Above are the results from attempting to recognize 36 different single-view samples of each of the models in the database. The beliefs in the different models are represented by different symbols, each symbol indicating the true model used during that trial.

The level of numerical underflow of the system is represented by a "U" on the $y - axis$. Because so many trials fall into this category they are marked with a simple point, *except* when the belief is for the true model used in the trial.

By observing the log of the beliefs, one can see the bi-modality in the results.

FIGURE 7.7. Log of beliefs in the Big Sphere, Block, Cylinder, Lemon, Small Sphere, and Round Block.

The results illustrate a clustering effect in the beliefs. The first large cluster indicates that the highest degree of confidence lies in the correct model hypotheses. Beneath this group, is a scatter of beliefs in the incorrect model. The degree of evidence of these hypotheses varies from model to model. This second large cluster occurs for beliefs in models that lie below the numerical precision of the system (denoted the "U" level). The distinct bi-modality of the results justifies the application of an external threshold differentiating between the high confidence informative views and the low confidence uninformative views. In addition, they indicate that the value of this threshold is not critical. For example, for the Big Sphere model, the cutoff point can lie anywhere from 10^{-5} to 10^{-60} (above the "U" level). However, the desire is to choose this threshold so as to eliminate the majority of false positive cases. Although the plot does not illustrate the maximum likelihood results, making it impossible to tell where false positive indications occur, one can see that by placing the cutoff above the scatter of incorrect hypotheses, one can ensure a minimal amount of incorrect maximum likelihood indications. Furthermore, one can see from the results that one does not necessarily need to choose a universal threshold level for all the models. By examining the difference in the Big Sphere and the Rounded Block distributions, one can see that choosing individual cutoff levels would render the results more accurate. For maximal efficiency, these levels can be computed off-line prior to experimentation, and then used in the recognition stage.

For the purposes of testing the hypothesis that an external cutoff would divide the results into informative and uninformative cases (and eliminate the majority of false-positive cases), the threshold for undetermined states was uniformly raised to 0.00001. Figure 7.8 shows the results of imposing this threshold on the belief distributions. One can see that all but one incorrect state (B) has become undetermined. However, several correct identifications have become undetermined as well. This is to be expected since setting this threshold causes all uncertain identifications to be removed. We therefore make the empirical observation that, by raising the threshold, states that are not undetermined are accompanied by a high accuracy in recognition.

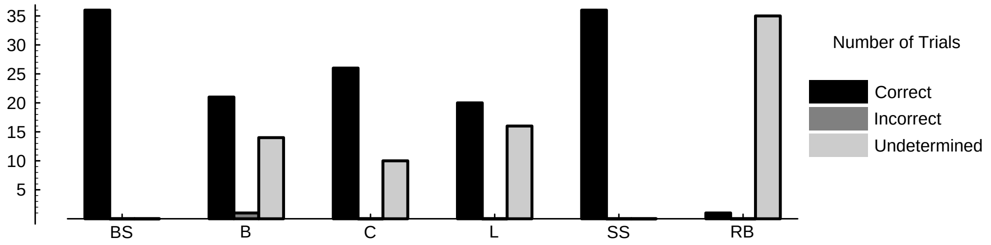


FIGURE 7.8. Matching samples taken from a single viewpoint while imposing a threshold of 0.00001.

3.3. Incremental Recognition. The described experiments suggest the possibility of an incremental recognition procedure. It is based on the following observations obtained empirically over successive trials:

i) Viewpoints that provide very little information, or *uninformative* views, generally can be detected by their low confidence levels (beliefs). Because of the bi-modality of the belief spread, these can be discovered by application of a threshold. Detection of such events is a clear indicator that further sampling is required.

ii) *Informative* views are generally accompanied by high beliefs, but with the possibility of a false-positive indication. These can also be detected by threshold application.

iii) The likelihood of successive false-positive indications is very small. First, this is a consequence of the high selectivity of the reference distributions which result in low frequencies of false-positive indications in the first place (e.g. Figure 7.8). Second, it is unusual for observer motion to result in similar viewpoints in two successive views (general position assumption).

To illustrate these observations by example, Table 2 shows a sequence of single-view recognition attempts, corresponding to the first 6 entries in the second half of the table. Iteration 1 is inconclusive, the object is either a square or rounded block (However the results of running a maximum likelihood algorithm indicate that the object is a rounded block). In iterations 2 and 5 the object is undetermined. Iterations 3, 4, and 6, on the other hand, consistently support the correct classification of the unknown object as the square block.

To explore the possibility of an incremental scheme, an experiment was performed whereby evidence from single-views was accumulated. The method described in Chapter 6 was employed, whereby the system binarized the beliefs above the predefined threshold at each view. Evidence at each stage was computed by histogramming the binarized beliefs accumulated thus far. Table 3 displays the result of accumulating evidence after 36 single-view iterations. Table 3a illustrates the results when the zero states were established by the numerical limitations of the system, whereas in b, a threshold of 0.00001 was imposed externally. One can see from these results that, after several iterations, choosing a winner based on a maximum likelihood scheme on the accumulated beliefs gave the correct answer in all cases. The false-positive cases became insignificant due to insufficient evidence. In fact, Table 3b illustrates that hardly any evidence in incorrect models remained after applying the threshold of 0.00001. However, in the case of the rounded block, the majority of the evidence in the correct model was also eliminated, indicating that perhaps this choice of threshold was too high in this case. Its belief values were, in fact, significantly lower than the rest of the objects. In these cases, this choice of threshold seems to be appropriate in that it removes the false-positive cases, while maintaining a high degree of confidence in the correct hypotheses. This justifies using independent threshold levels for each of the models in the database.

4. Multiple-Part Object Recognition

We have shown that recognition of single-part objects based on partial information retains some of the selectivity of systems based on complete information. However, these objects are less complex than most found in the real world, so we are interested in the natural extension to recognizing objects that consist of several articulated parts. Our current focus is “recognition by parts”, whereby measured objects are segmented into their constituent

	BS	B	C	L	SS	RB
BS	36	0	0	0	13	0
B	0	28	0	0	0	3
C	1	1	33	1	1	0
L	0	0	0	36	0	0
SS	0	0	0	0	36	0
RB	0	0	0	0	0	18

a) Threshold = computational underflow

	BS	B	C	L	SS	RB
BS	36	0	0	0	0	0
B	0	21	0	0	0	1
C	0	0	26	0	0	0
L	0	0	0	20	0	0
SS	0	0	0	0	21	0
RB	0	0	0	0	0	1

b) Threshold = 0.00001

Displayed above are the tables describing the accumulation of evidence from 36 single-view experiments. Each row describes the histogram of the binarized belief distributions for a particular measured model. The columns refer to the reference models. Zero values are defined by a) numerical underflow of system and b) a threshold of 0.00001.

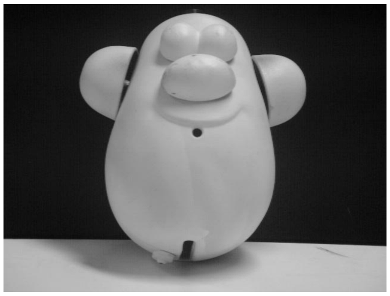
TABLE 3. Histogram of binarized belief distributions after single-view iterations.

parts, each of which is compared to the parts in the database. The task of recognizing these parts is much more challenging than recognizing single-part objects due to problems of self-occlusion and segmentation. Objects are seen as collections of independent parts, where topological relationships are not yet considered in this thesis[2].

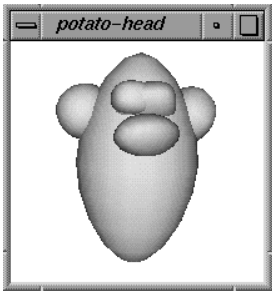
A toy potato-head consisting of two ears, two eyes, a nose and a head was chosen for the purposes of testing the part recognition algorithm on complex objects. In order to scan the object from all possible viewing positions, the head was scanned as described in Section 2. A picture of the set-up used to scan the head is found in Figure 7.2.

Figure 7.9a displays the actual potato-head toy used in the experiment. Most of constituent parts conformed well to non-deformable superellipsoid models, with the exception of the head whose shape was tapered. The potato-head toy was chosen because its parts were

[2] Recognition strategies that take topology into account are currently being investigated.



a) Original potato-head toy.



b) Reference potato-head model created by training.

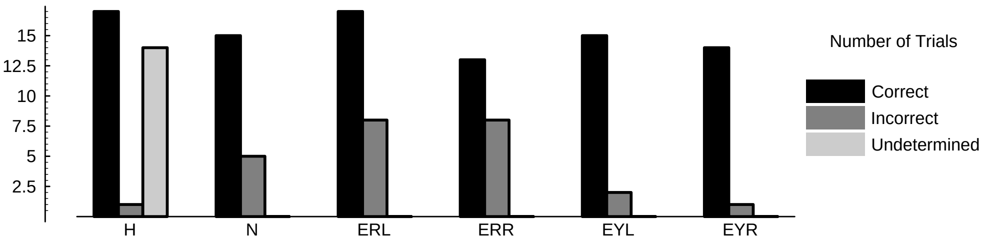
FIGURE 7.9. Potato-head: a) real object and b) reference model.

similar to each other as well as to the reference spheres making discrimination a challenging task.

Ten samples of the potato-head were used in the training procedure. Each sample was produced by scanning the object from several viewpoints in an exploration sequence. The reference model resulting from training can be found in Figure 7.9b.

4.1. Matching Using Partial Information. Since the more interesting task is to recognize an object with only partial information available, an experiment was devised whereby the potato-head was measured from 32 independent viewing positions. Recognition was performed on each of these samples in turn, using a database consisting of the parts of the potato-head as well as the single-part reference models used earlier as distractors. The results of the using maximum likelihood on the beliefs can be seen in Figure 7.10.

The results indicate that the system was able to successfully recognize instances of articulated parts of a complex object with only partial information available. The system was able to maintain its selectivity even with very little information available from single viewpoints, compounded by the added effects of self-occlusion. In fact, even with complete



Displayed above are the tables describing the belief distributions of the potato-head measured from single viewpoints. The parts of the potato-head are: a head (H), a nose (N), a left ear (ERL), a right ear (ERR), a left eye (EYL), and a right eye (EYR). Here, identifying one eye as the other, or one ear as the other was considered to be a correct identification. Zero values are defined by the numerical underflow of system.

FIGURE 7.10. Matching samples of the potato-head taken from single viewpoints.

data gathered from all around the object surface, most parts were embedded within others and thus part of their surfaces were not visible. The results were models that were unconstrained in several directions. This caused the reference parts to be created without complete information. Therefore training no longer ensured models with parameters that were close to the true values. This added to the difficulty of the recognition task.

For the purposes of the maximum likelihood experiments, the left and right eyes were considered to be two instances of the same object. The same applied to the ears. This is because a "recognition by parts" strategy considers objects that are identical in size and shape to be the same model, as is the case with the eyes and ears of the potato-head. In future research, when topological relationships will be included into a solution for recognition of complex objects, different instances of the same part will be distinguished by position and orientation.

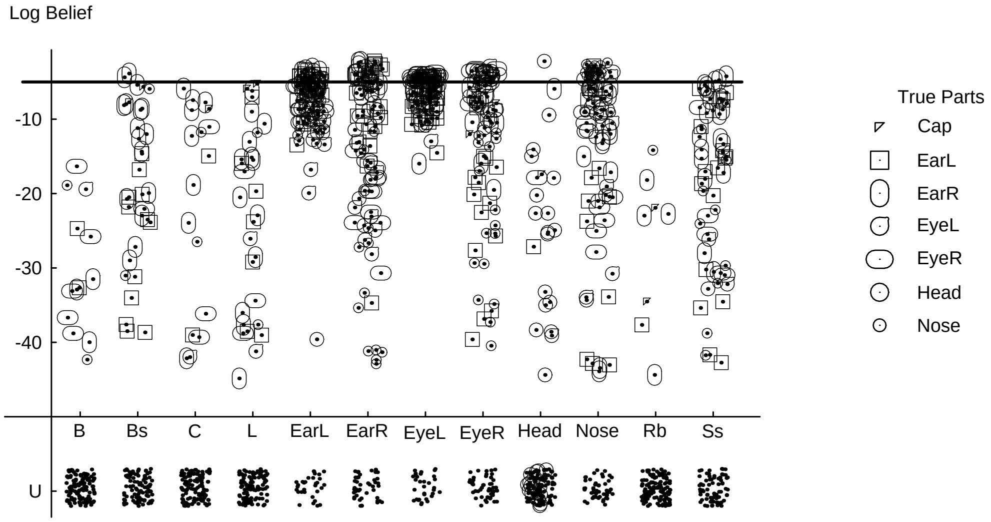
The results show a high number of undetermined states for the head. This is because the head is tapered, breaking the assumption that the objects *can* in fact be accurately modeled by non-deformable superellipsoids. Different single-view samples of the head produce very different superellipsoids depending on where the data were collected from. Similar to the problem caused by self-occlusion, the reference head was described by one particular superellipsoid, whose parameters were tightly constrained (due to the fact that data were gathered all around the object to create each sample used in training). In the current scheme, the reference description did not encompass all possible superellipsoid models describing the tapered part. Therefore other equally viable descriptions that result from single view measurements were not recognized correctly. This lead to undetermined states.

Other potential problems occur because the recognition process relies heavily on the accuracy of the segmentation process. Because of this, errors in the segmentation of the range data can lead to errors in recognition. In these experiments, there were several cases where the head was divided into two distinct parts: a “head” and a “cap”. Because the database allowed for only one part for the head, the system identified the cap part as being as lemon or some other reference model. This was understandable as the cap was similar in size and shape to these models.

However, most of the incorrect states arose due to the similarity of the reference models. For example, the eyes resembled the smaller sphere, the nose and the ears. Similarly, the ears were extremely close to the bigger sphere in size and shape. As a result, their distributions overlapped significantly, making it difficult to distinguish between them. Yet, in the majority of cases, these incorrect identifications occurred with low beliefs. This lead to the hypothesis that that most of these states actually arose from uninformative viewpoints, and could be eliminated by raising the threshold for undetermined states.

In order to justify application of an external threshold to distinguish between uninformative and informative viewpoints, the beliefs in the potato-head parts as well as the beliefs in the single-part objects were plotted on a logarithmic scale graph. Once again, a bi-modal distribution was anticipated, whereby a clear division between the informative and uninformative states would permit the use of a threshold to distinguish between the two. The results can be found in Figure 7.11.

As hypothesized, the results indicate a bi-modal distribution for the beliefs in the potato-head parts. For each of these parts, there lay a top cluster, representing relatively high beliefs in the correct models. Beneath this, a thin scatter of beliefs in other models can be seen. Finally, the bottom cluster occurred for those beliefs that were below the numerical precision of the system (producing zero beliefs). However, the majority of the beliefs were concentrated in the the top cluster illustrating that, most of the time, the system had high confidence in the correct part. However, some scattered beliefs in the single-part distractors occurred as well. It is important to note that the majority of these cases lay below the top cluster of correct identifications, indicating that by application of a threshold anywhere from 10^{-10} to 10^{-5} should eliminate the majority of the false-positive cases. Once again, the exact value of the cutoff level is not critical. Figure 7.11 illustrates the results that can be achieved by applying a threshold of $10^{-}5$. This would lead to minimal false-positive indications accompanying a high number of correct votes. The case of the head, however, emphasizes the possibility of individual threshold levels for maximal efficiency. Here, a much lower threshold would ensure the highest number of correct matches.



Above are the results from attempting to recognize 32 different single-view samples of each of the parts of the potato-head: the Left Ear (EarL), Right Ear (EarR), Left Eye (EyeL), Right Eye (EyeR), Head (Head), and Nose (Nose). The single-part reference models were also included as distractors for the recognition process. These included the Block (B), Big Sphere (Bs), Cylinder (C), Lemon (L), Round Block (Rb), and Small Sphere (Ss). (For an explanation of the plot, see Figure 7.7).

One can see the bi-modality in the log of the beliefs in the potato-head models. The beliefs in the distractors appear much more scattered, the majority lying beneath the top cluster of the potato-head parts. The top horizontal line indicates the results achieved by applying a threshold of 10−5. This would lead to minimal false-positive indications accompanying a high number of correct votes.

FIGURE 7.11. Log of beliefs in the Potato-Head parts, as well as the Big Sphere, Block, Cylinder, Lemon, Small Sphere, and Round Block.

To investigate that the hypothesis that an external cutoff can divide the results into informative and uninformative states, and remove the majority of incorrect identifications, the cutoff point was raised to 0.00001. The results are shown in Figure 7.12. On can see that, in the most of cases, the external threshold retained most of the correct states, confirming that the system had high confidence in the correct identifications. The exception was the case of the head, where low beliefs caused almost all of the correct identifications to become undetermined states.

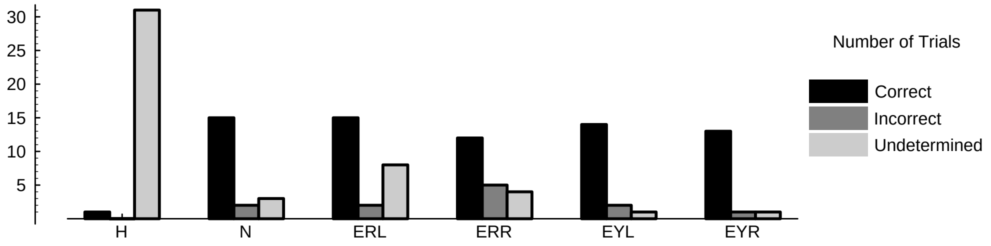


FIGURE 7.12. Matching samples of the potato-head model while imposing a threshold of 0.00001.

	H	N	ERL	ERR	EYL	EYR	BS	B	C	L	SS	RB
H	17	0	1	0	0	0	0	0	0	0	0	0
N	0	20	20	15	20	20	2	2	2	2	10	1
ERL	1	15	25	24	25	15	12	2	2	7	18	1
ERR	1	15	21	21	21	13	16	3	8	13	20	4
EYL	1	16	17	12	17	17	0	1	1	2	4	0
EYR	1	15	15	14	15	15	1	5	3	3	5	0

a) Threshold = computational underflow

	H	N	ERL	ERR	EYL	EYR	BS	B	C	L	SS	RB
H	1	0	0	0	0	0	0	0	0	0	0	0
N	0	16	2	0	12	1	0	0	0	0	0	0
ERL	0	1	14	12	0	0	0	0	0	0	0	0
ERR	0	1	9	12	0	1	2	0	0	0	3	0
EYL	0	3	0	0	14	8	0	0	0	0	0	0
EYR	0	3	0	0	13	11	0	0	0	0	0	0

b) Threshold = 0.00001

TABLE 4. Histogram of binarized belief distributions for the potato-head after 32 single-view iterations (For explanation, see Table 3).

4.2. Incremental Recognition. In order to explore the possibility of an incremental recognition strategy for complex objects, an experiment was devised whereby evidence from single-views of the potato-head toy was accumulated. Similar to the single-part object case, the belief distributions were binarized at a predefined threshold at each viewing position. At each stage, a histogram of the binarized distributions produced the evidence accumulated thus far. The results of accumulating evidence after 32 single-views can be seen in Table 4. In Table 4a, the cutoff point was determined by the numerical precision of the system. In Table 4b, a threshold of 0.00001 was imposed externally.

Table 4a illustrates that the distributions from single-views were relatively “wide” in that a measured model produced a degree of belief in several reference models at once. The result is that, in most cases, the accumulated binarized evidence points to several models at once. Attempting to choose a single winner after several iterations would therefore be a difficult task. The choice would however be limited to a few candidates as some false-positive indications have become insignificant due to insufficient evidence. For example, in the case of the Nose, lack of evidence in the Big Sphere, Block, Cylinder, Lemon and Rounded Block has caused the belief in these models to become unsubstantiated. The hypothesis was that the evidence in the true model was so much stronger than the evidence in the other models that, by raising the threshold to an appropriate value, one could eliminate the majority of the false indications. The result would be an accumulation of evidence in the true models.

Table 4b validates the hypothesis by illustrating that the majority of the evidence in the incorrect models were removed after application of the external threshold. In fact, if one were to choose a winner based on a maximum likelihood scheme of the accumulated evidence, the results would be correct for all models[3]. In the case of the head, however, the majority of the evidence in the correct model was eliminated as well. This indicates the possibility that the choice of threshold was not appropriate for the head.

The problem of merging the belief distributions from different viewpoints of complex objects is quite difficult. The difficulty lies in establishing correspondence between parts from different views. The problem is much more difficult than in the single-part object case which encompassed the strong prior assumption that the object measured does not change from view to view. This assumption no longer holds, and a theory providing the correspondence is needed. Methods that provide part correspondence based on geometry (Soucy & Ferrie 1992, Soucy 1992) were used for these experiments, however they are restrictive in that the different viewpoints must be close enough to contain overlapping data. As well, merging data on the level of geometry is computationally expensive. Therefore, a new scheme for merging the belief distributions, based on the models themselves, their associated beliefs, and the relationships between them will be the focus of future research.

We have demonstrated that system is able to recognize parts of articulated objects with only partial information available. Extension to recognition of multiple-part objects will involve incorporating topological information into the solution. The rotation and translation parameters of the superellipsoid models provide this information as they can be used to infer the distance and angle between the parts. Once belief in each of the parts is established, graph matching techniques can be employed to calculate the belief in the entire object. Current work in our lab is concentrated on the solution to this problem.

[3] We have treated the left and right eyes as being the instances of the same object (similarly for the ears).

CHAPTER 8

Conclusions

In this thesis, we have presented a new framework for parametric shape recognition based on a probabilistic model of inverse theory first introduced by Tarantola (1987). We have shown how a Bayesian recognition strategy can be derived automatically by applying the theory and have demonstrated its implementation in a system for recognizing 3D objects based on superellipsoid parameters (As well, see Arbel, Whaite & Ferrie 1994*b*, Arbel, Whaite & Ferrie 1994*a*).

Casting the problem into a general inverse theory framework introduces several important contributions to the field of object recognition. The first is that the method explicitly enumerates all sources of prior knowledge. This way, if conditioning is necessary, the sources of knowledge are apparent, and can therefore be examined. This is important in that many recognition systems include implicit, hidden assumptions about the nature of the world. As a result, these methods may work well in specific situations, but cannot be easily modified to work elsewhere. By representing knowledge as a probability density function, both the information and the ambiguities associated with them are incorporated into the solution. This permits the recognition engine to make well-informed decisions. As well, the method is not dependent on the exact nature of the information, but rather provides a general recipe for merging any group of contextual priors. Finally, the solution to the inverse problem is presented in the form of a conditional probability density function. The importance of this result is that it provides a *qualification* of the assessments made by the recognition procedure. This is vital in that no problem in vision works in complete isolation, but rather must communicate descriptions of results to external processes. In order to do so, it is important to inform these processes of the uncertainties in the descriptions as well. Most recognition schemes do not provide this information. Instead, they make absolute assessments about the identity of the unknown object. This provides the external processes with only partial information, biased to their notion of what constitutes a clear “winner”.

We have developed a method of avoiding degeneracies in the superellipsoid representation, which permits the use of this convenient parametric form without incurring undue computational overhead. We have determined empirically that there are only a finite number of possible equivalence classes for the superellipsoid, and rather than restrict ourselves to one particular model description, we proposed a method that represents each model by a multi-modal description encompassing all possible degeneracies. Our current representation only includes the rotational equivalent forms, but future work will include all possible forms in the representations.

The experimental results indicate that the strategy is quite robust, not only in situations where complete surface information is available but also in those cases where it is only partially accessible. In this and other works (Arbel, Ferrie & Whaite 1994), we have demonstrated that it is indeed possible to differentiate between *informative* and *uninformative* viewpoints, and have shown how the resulting belief distributions can be used to assess the quality of the interpretation, by assessing the beliefs associated with a particular set of assertions based on this data. The importance of this result is that it provides a basis by which an external agent can assess the quality of the information from a particular viewpoint, and make informed decisions as to what action to take using the data at hand. The bi-modal nature of the resulting belief distributions have indicated that this can be easily accomplished by application of an external threshold.

We have also demonstrated that some viewpoints can give rise to ambiguous information, where the system has confidence in more than one hypothesis. Similar to the motivation behind *autonomous exploration* in the model-building phase (Whaite & Ferrie 1994), ambiguous views have spawned the development of an incremental recognition scheme, where we seek information from a new viewpoint to reduce the overall ambiguity. We have shown how evidence, in the form of the belief distributions, can be accumulated from a sequence of views. The experiments have demonstrated that the maximum likelihood hypothesis is largely viewpoint-invariant, implying that merging votes for the different hypotheses over a sequence of views should lead to a clear winner. Because the beliefs are not normalized, we have given equal weighting to all hypotheses by binarizing the values above a threshold. We have illustrated that by histogramming the binarized beliefs and picking the highest score of the result, we choose the correct winner in all cases.

By qualifying the recognition results, the method provides potential for a wide variety of applications. For example, an active recognition agent can choose viewpoints that will maximize the belief distribution associated with an object of interest. We have not specified *how* to choose this viewpoint, but the method can be used to determine if the particular choice leads to a sufficient level of information. Another important application of the methodology is a strategy for off-line computation of a pre-computed set of characteristic views. One can rank these views according to the belief distributions, and then store the n best views. Predefining these views speeds up on-line computations by directing the active agent's attention to informative viewpoints, thereby reducing the search space of viable hypotheses. These and other topics are currently under investigation in our laboratory.

Some observations are in order regarding the autonomous explorer, the system used to automatically generate the database models used for recognition. In the numerous trials performed during the course of this research we were able to consistently obtain stable parametric descriptions of the model database. These were largely independent of viewpoint, variations in sampling, and the trajectory chosen by the mobile laser scanner. The generation of stable, salient object models is clearly an essential ingredient in the implementation of a successful object recognition system. Future work will involve exploration guided by feedback from the recognition system. This is possible because all sources of knowledge are made explicit within the framework described. Therefore, the system could actively acquire information needed to correctly classify the objects.

The system described exhibits a high degree of selectivity in matching object primitives, paving the way for recognition of articulated objects. Current work includes a scheme for multiple-part object recognition involving a graph-matching procedure. It is based on the the work presented in this thesis, which outlines a sound, statistical method for comparing the nodes. Given its success in discriminating based on partial information, the search-space for the graph-matching problem should be considerably reduced.

Finally, we conclude by noting that although, in this thesis, we have concentrated on the problem of recognizing particular parametric models, the general inverse theory can be used to solve many problems in vision. One such application is the problem of object classification. Here, rather than represent the database knowledge as a series of delta functions, one for each prototype in the database, one might represent a database of classes by a series of normal distributions. The effect would be to spread out the database prototypes from points in parameter space to clouds of points. Another option might be to represent each class by a sum of delta functions. An example of which may be to include three possible sizes for each reference model. It is important to note that although these applications differ from classical object recognition, they do not involve a change in the methodology, but rather a modification in the shape of the distributions representing the sources of knowledge. This type of flexibility, made possible because all of the sources of knowledge are made explicit, is one of the prime advantages of using the general inverse theory to solve problems in vision.

APPENDIX A

Combining Normal Distributions

This appendix chapter will present the mathematical details involved in the proof that the convolution of normal distributions is itself a normal distribution (as was required in Chapter 4). Section 1 will provide the proof that the convolution of multivariate Gaussian distributions is itself a multivariate Gaussian distribution. Section 2 will use this result to show that the integral of the product of two normal distributions (the convolution) is also a normal distribution.

1. The Convolution of Gaussians

In this section, we wish to illustrate the useful result that the convolution of a multivariate Gaussian function (or a normal density function) with another is itself Gaussian. We will denote a Gaussian function over the space X as

$$\begin{aligned} G(\mathbf{x}, \mathbf{C}) &= \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x})\right) \\ &= \exp\left(\mathbf{x}^T\mathbf{H}\,\mathbf{x}\right) \end{aligned} \tag{31}$$

where where $\mathbf{x}$ is a vector in the n-dimensional vector space X, and $\mathbf{C}$ is a linear covariance operator on the space X (an $n \times n$ matrix) . The covariance operator defines the spread or dispersion of the function on the different parameter directions. In matrix form it is symmetric ($\mathbf{C}^T = \mathbf{C}$), and positive definite. Where convenient we will also use the alternate form with $\mathbf{C}^{-1} = 2\mathbf{H}$ where $\mathbf{H}$ is the Hessian of the quadratic form $\mathbf{x}^T\mathbf{H}\ \mathbf{x}$. $\mathbf{H}$ is also symmetric and positive definite ($\mathbf{x}^T\mathbf{H}\,\mathbf{x} > 0$ for $\mathbf{x} \neq \mathbf{0}$). Let

$$G_a(\mathbf{x}) = \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{C}_a^{-1}\mathbf{x}\right) = \exp\left(-\mathbf{x}^T\mathbf{H}_a\mathbf{x}\right)$$

and

$$G_b(\mathbf{x}) = \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{C}_b^{-1}\mathbf{x}\right) = \exp\left(-\mathbf{x}^T\mathbf{H}_b\mathbf{x}\right)$$

denote two such Gaussian functions. The convolution of these is defined as

$$\begin{aligned}(G_a * G_b)(\mathbf{x}_c) &= \int_X G_a(\mathbf{x}_c - \mathbf{x})\ G_b(\mathbf{x})\ d\mathbf{x} \\ &= \int_X \exp\Big(- \big((\mathbf{x}_c - \mathbf{x})^T \mathbf{H}_a (\mathbf{x}_c - \mathbf{x}) + \mathbf{x}^T \mathbf{H}_b \mathbf{x}\big)\Big)\ d\mathbf{x} \qquad (32) \\ &= \int_X \exp\left(-Q(\mathbf{x})\right)\ d\mathbf{x}\end{aligned}$$

When expanded, the quadratic exponent is

$$\begin{aligned}Q(\mathbf{x}) &= (\mathbf{x}_c - \mathbf{x})^T \mathbf{H}_a (\mathbf{x}_c - \mathbf{x}) + \mathbf{x}^T \mathbf{H}_b \mathbf{x} \\ &= \mathbf{x}^T(\mathbf{H}_a + \mathbf{H}_b)\mathbf{x} - 2(\mathbf{H}_a\mathbf{x}_c)^T\mathbf{x} + \mathbf{x}_c^T \mathbf{H}_a \mathbf{x}_c \\ &= \mathbf{x}^T \mathbf{A}\ \mathbf{x} - 2\mathbf{b}^T\mathbf{x} + c; \qquad (33)\end{aligned}$$

where $\mathbf{A} = \mathbf{H}_a + \mathbf{H}_b$, $\mathbf{b} = \mathbf{H}_a\mathbf{x}_c$, and $c = \mathbf{x}_c^T\mathbf{H}_a\mathbf{x}_c$. Note that because $\mathbf{H}_a$ and $\mathbf{H}_b$ are symmetric positive definite then their sum $\mathbf{A}$ its inverse $\mathbf{A}^{-1}$ are as well.

Because $\mathbf{A}$ is symmetric, the terms in $\mathbf{x}$ can be collected by rewriting the quadratic form about the location of its minimum $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, such that

$$\begin{aligned}(\mathbf{x} - \mathbf{A}^{-1}\mathbf{b})^T\mathbf{A}(\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}) &= \mathbf{x}^T\mathbf{A}\mathbf{x} - 2(\mathbf{A}^{-1}\mathbf{b})^T\mathbf{A}\mathbf{x} + (\mathbf{A}^{-1}\mathbf{b})^T\mathbf{A}(\mathbf{A}^{-1}\mathbf{b}) \\ &= \mathbf{x}^T\mathbf{A}\mathbf{x} - 2\mathbf{b}^T(\mathbf{A}^{-1})^T\mathbf{A}\mathbf{x} + \mathbf{b}^T(\mathbf{A}^{-1})^T\mathbf{A}\mathbf{A}^{-1}\mathbf{b} \\ &= \mathbf{x}^T\mathbf{A}\mathbf{x} - 2\mathbf{b}^T\mathbf{x} + \mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}\end{aligned}$$

or that

$$\mathbf{x}^T\mathbf{A}\mathbf{x} - 2\mathbf{b}^T\mathbf{x} = (\mathbf{x} - \mathbf{A}^{-1}\mathbf{b})^T\mathbf{A}(\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}) - \mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}. \qquad (34)$$

After substituting this into (33), we get that

$$\begin{aligned}Q(\mathbf{x}) &= \mathbf{x}^T\mathbf{A}\mathbf{x} - 2\mathbf{b}^T\mathbf{x} + c \\ &= (\mathbf{x} - \mathbf{A}^{-1}\mathbf{b})^T\mathbf{A}(\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}) + c - \mathbf{b}^T\mathbf{A}^{-1}\mathbf{b} \\ &= (\mathbf{x} - \mathbf{A}^{-1}\mathbf{b})^T\mathbf{A}(\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}) + Q_{min}. \qquad (35)\end{aligned}$$

Expanding the value at the minimum

$$\begin{aligned}Q_{min} &= c - \mathbf{b}^T\mathbf{A}^{-1}\mathbf{b} \\ &= \mathbf{x}_c^T\mathbf{H}_a\mathbf{x}_c - (\mathbf{H}_a\mathbf{x}_c)^T(\mathbf{H}_a + \mathbf{H}_b)^{-1}(\mathbf{H}_a\mathbf{x}_c) \\ &= \mathbf{x}_c^T\mathbf{H}_a\mathbf{x}_c - \mathbf{x}_c^T\left(\mathbf{H}_a(\mathbf{H}_a + \mathbf{H}_b)^{-1}\mathbf{H}_a\right)\mathbf{x}_c \\ &= \mathbf{x}_c^T\left(\mathbf{H}_a - \mathbf{H}_a(\mathbf{H}_a + \mathbf{H}_b)^{-1}\mathbf{H}_a\right)\mathbf{x}_c.\end{aligned}$$

This can be simplified further by factorizing out $\mathbf{H}_a$ on the left and $(\mathbf{H}_a + \mathbf{H}_b)^{-1}\mathbf{H}_a$ on the right

$$\begin{aligned}
Q_{min} &= \mathbf{x}_c^T \mathbf{H}_a \left(\mathbf{H}_a^{-1}(\mathbf{H}_a + \mathbf{H}_b) - \mathbf{I}\right)(\mathbf{H}_a + \mathbf{H}_b)^{-1}\mathbf{H}_a \mathbf{x}_c \\
&= \mathbf{x}_c^T \mathbf{H}_a \left(\mathbf{H}_a^{-1}\mathbf{H}_b\right)(\mathbf{H}_a + \mathbf{H}_b)^{-1}\mathbf{H}_a \mathbf{x}_c \\
&= \mathbf{x}_c^T \mathbf{H}_b (\mathbf{H}_a + \mathbf{H}_b)^{-1}\mathbf{H}_a \mathbf{x}_c \\
&= \mathbf{x}_c^T \left(\mathbf{H}_a^{-1}(\mathbf{H}_a + \mathbf{H}_b)\mathbf{H}_b^{-1}\right)^{-1} \mathbf{x}_c \\
&= \mathbf{x}_c^T \left(\mathbf{H}_a^{-1} + \mathbf{H}_b^{-1}\right)^{-1} \mathbf{x}_c
\end{aligned} \tag{36}$$

With this, the convolution (32) is separable into two Gaussians, only one of which is a function of $\mathbf{x}$, that is

$$\begin{aligned}
(G_a * G_b)(\mathbf{x}_c) &= \int_X \exp\bigl(-Q(\mathbf{x})\bigr)\, d\mathbf{x} \\
&= \int_X \exp\bigl(-\left((\mathbf{x} - \mathbf{A}^{-1}\mathbf{b})^T \mathbf{A}(\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}) + Q_{min}\right)\bigr)\, d\mathbf{x} \\
&= \exp\bigl(-Q_{min}\bigr) \int_X \exp\bigl(-(\mathbf{x} - \mathbf{A}^{-1}\mathbf{b})^T \mathbf{A}(\mathbf{x} - \mathbf{A}^{-1}\mathbf{b})\bigr)\, d\mathbf{x}.
\end{aligned} \tag{37}$$

The integral of the Gaussian over the space X has a known solution — that used to normalize the multivariate normal probability distribution. Let us first change variables to $\mathbf{y} = \mathbf{x} - \mathbf{A}^{-1}\mathbf{b}$, then $d\mathbf{x} = d\mathbf{y}$, so

$$\begin{aligned}
\int_X \exp\bigl(-(\mathbf{x} - \mathbf{A}^{-1}\mathbf{b})^T \mathbf{A}\ (\mathbf{x} - \mathbf{A}^{-1}\mathbf{b})\bigr)\, d\mathbf{x} &= \int_X \exp\left(-\frac{1}{2}\mathbf{y}^T(2\mathbf{A})\mathbf{y}\right) d\mathbf{y} \\
&= (2\pi)^{\frac{n}{2}} \left|(2\mathbf{A})^{-1}\right|^{\frac{1}{2}} \\
&= (2\pi)^{\frac{n}{2}} \left(|2\mathbf{A}|^{-1}\right)^{\frac{1}{2}} \\
&= (2\pi)^{\frac{n}{2}} |2\mathbf{H}_a + 2\mathbf{H}_b|^{-\frac{1}{2}} \\
&= (2\pi)^{\frac{n}{2}} \left|\mathbf{C}_a^{-1} + \mathbf{C}_b^{-1}\right|^{-\frac{1}{2}}.
\end{aligned}$$

When it and (36) are substituted into (37), we get that the convolution of two multivariate Gaussians

$$\begin{aligned}
(G_a * G_b)(\mathbf{x}_c) &= \sqrt{\frac{(2\pi)^n}{\left|\mathbf{C}_a^{-1} + \mathbf{C}_b^{-1}\right|}}\ \exp\left(-\frac{1}{2}\mathbf{x}_c^T(\mathbf{C}_a + \mathbf{C}_b)^{-1}\mathbf{x}_c\right) \\
&= \sqrt{\frac{(2\pi)^n}{\left|\mathbf{C}_a^{-1} + \mathbf{C}_b^{-1}\right|}}\ G(\mathbf{x}_c, \mathbf{C}_a + \mathbf{C}_b)
\end{aligned} \tag{38}$$

is itself a Gaussian where the covariances are summed.

2. Integral of the Product of Normal Distributions

When integrated the product of two normal distributions is a normal distribution. This is to be expected as the integral is really the convolution of normalized Gaussians, and as we have shown in Section 1, this is itself a Gaussian. Because the Gaussians have been normalized we would expect the convolution to be normalized as well.

As was stated in Chapter 4, a multivariate normal probability density function over the space X is a normalized Gaussian

$$\begin{aligned} N(\mathbf{x}-\mathbf{x}_\mu, \mathbf{C}) &= \frac{1}{\sqrt{(2\pi)^n \; |\mathbf{C}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x}_\mu)^T \mathbf{C}(\mathbf{x}-\mathbf{x}_\mu)\right) \\ &= \frac{G(\mathbf{x}-\mathbf{x}_\mu, \mathbf{C})}{\sqrt{(2\pi)^n \; |\mathbf{C}|}} \end{aligned} \tag{39}$$

centered on the mean value of the distribution $\mathbf{x}_\mu$, and with a dispersion in the various parameter directions given by the covariances $\mathbf{C}$.

The integral of the product of two normal distributions can be written as the convolution of two Gaussians. To show this we first note from (39) that $G(\mathbf{x}-\mathbf{x}_\mu, \mathbf{C}) = G(\mathbf{x}_\mu - \mathbf{x}, \mathbf{C})$. This is simply a consequence of the symmetry of the distribution. Thus we have that

$$\begin{aligned} \int_X N(\mathbf{x}-\mathbf{x}_a, \mathbf{C}_a)\; N(\mathbf{x}-\mathbf{x}_b, \mathbf{C}_b)\; d\mathbf{x} &= \int_X \frac{G(\mathbf{x}-\mathbf{x}_a, \mathbf{C}_a)}{\sqrt{(2\pi)^n |\mathbf{C}_a|}} \; \frac{G(\mathbf{x}-\mathbf{x}_b, \mathbf{C}_b)}{\sqrt{(2\pi)^n |\mathbf{C}_b|}} \; d\mathbf{x} \\ &= \int_X \frac{G(\mathbf{x}_a - \mathbf{x}, \mathbf{C}_a)\; G(\mathbf{x}-\mathbf{x}_b, \mathbf{C}_b)}{(2\pi)^n \sqrt{|\mathbf{C}_a| \; |\mathbf{C}_b|}} \; d\mathbf{x} \end{aligned}$$

After a change of variable $\mathbf{y} = \mathbf{x}-\mathbf{x}_b$, it follows that $d\mathbf{y} = d\mathbf{x}$ and that this is the convolution of two Gaussians

$$\begin{aligned} &= \int_X \frac{G\left((\mathbf{x}_a - \mathbf{x}_b) - \mathbf{y}, \mathbf{C}_a\right)\; G(\mathbf{y}, \mathbf{C}_b)}{(2\pi)^n \sqrt{|\mathbf{C}_a| \; |\mathbf{C}_b|}} \; d\mathbf{x} \; d\mathbf{y}. \\ &= \frac{(G_a * G_b)(\mathbf{x}_a - \mathbf{x}_b)}{(2\pi)^n \sqrt{|\mathbf{C}_a| \; |\mathbf{C}_b|}}. \end{aligned}$$

From (38) with $\mathbf{x}_c = (\mathbf{x}_a - \mathbf{x}_b)$

$$\begin{aligned} \int_X N(\mathbf{x}-\mathbf{x}_a, \mathbf{C}_a)\; N(\mathbf{x}-\mathbf{x}_b, \mathbf{C}_b)\; d\mathbf{x} &= \sqrt{\frac{(2\pi)^n}{|\mathbf{C}_a^{-1} + \mathbf{C}_b^{-1}|}} \frac{G(\mathbf{x}_a - \mathbf{x}_b, \mathbf{C}_a + \mathbf{C}_b)}{(2\pi)^n \sqrt{|\mathbf{C}_a| \; |\mathbf{C}_b|}} \\ &= \frac{G(\mathbf{x}_a - \mathbf{x}_b, \mathbf{C}_a + \mathbf{C}_b)}{\sqrt{(2\pi)^n \, |\mathbf{C}_a| \; |\mathbf{C}_b| \; \left|\mathbf{C}_a^{-1} + \mathbf{C}_b^{-1}\right|}}. \end{aligned} \tag{40}$$

Using the well known property of determinants that $|\mathbf{C}_a|\ |\mathbf{C}_b| = |\mathbf{C}_a \mathbf{C}_b|$ we can reorder and write that

$$\begin{aligned} |\mathbf{C}_a|\ |\mathbf{C}_b|\ \left|\mathbf{C}_a^{-1} + \mathbf{C}_b^{-1}\right| &= \left|\mathbf{C}_a(\mathbf{C}_a^{-1} + \mathbf{C}_b^{-1})\mathbf{C}_b\right| \\ &= |\mathbf{C}_a + \mathbf{C}_b|\,. \end{aligned} \tag{41}$$

After substituting this into (40), we see that that the integral of the product of the two normal distributions (really the convolution of two normal distributions) is

$$\begin{aligned} \int_X N(\mathbf{x} - \mathbf{x}_a, \mathbf{C}_a)\ N(\mathbf{x} - \mathbf{x}_b, \mathbf{C}_b)\ d\mathbf{x} &= \frac{G(\mathbf{x}_a - \mathbf{x}_b, \mathbf{C}_a + \mathbf{C}_b)}{\sqrt{(2\pi)^n\,|\mathbf{C}_a + \mathbf{C}_b|}} \\ &= N(\mathbf{x}_a - \mathbf{x}_b, \mathbf{C}_a + \mathbf{C}_b) \\ &= N(\mathbf{x}_b - \mathbf{x}_a, \mathbf{C}_a + \mathbf{C}_b) \end{aligned} \tag{42}$$

which is also a normal distribution, but where the covariances are summed.

REFERENCES

Aloimonos, Y., E. (1992), 'Purposive, qualitative, active vision', *CVGIP: Image Understanding* **56**(1), 3–129. special issue.

Arbel, T., Ferrie, F. P. & Whaite, P. (1994), Informative views and active recognition, Technical Report TR-CIM-94-16, Center for Intelligent Machines, McGill University, Montréal, Québec, Canada. *Submitted to 5th International Conference on Computer Vision*, Available via ftp at `ftp.cim.mcgill.ca` in `pub/techrep/1994/CIM-94-16.ps.Z`.

Arbel, T., Whaite, P. & Ferrie, F. P. (1994*a*), Recognizing volumetric objects in the presence of uncertainty, *in* 'Proceedings 12th International Conference on Pattern Recognition', IEEE Computer Society Press, Jerusalem, Israel, pp. 470–476.

Arbel, T., Whaite, P. & Ferrie, F. P. (1994*b*), Recognizing volumetric objects in the presence of uncertainty, Technical Report TR-CIM-94-03, Center for Intelligent Machines, McGill University, Montréal, Québec, Canada. Available via ftp at `ftp.cim.mcgill.ca` in `/pub/3d/papers/tr-cim-94-04.ps.gz`.

Arman, F. & Aggarwal, J. (1993*a*), 'CAD-based vision: Object recognition in cluttered range images using recognition strategies', *Computer Vision, Graphics, and Image Processing* .

Arman, F. & Aggarwal, J. (1993*b*), 'Model-based object recognition in dense-range images - a review', *ACM Computing Surveys* **25**(1), 5–43.

Bajcsy, R. & Solina, F. (1987), Three dimensional object recognition revisited, *in* ICC (1987).

Barr, A. H. (1981), 'Superquadrics and angle preserving transformations', *IEEE Computer Graphics and Applications* **1**(1), 11–23.

Bhanu, B. (1982), Surface representation and shape matching of 3-D objects, *in* 'Proceedings, IEEE Computer Society Conference on Pattern Recognition and Image Processing', IEEE, Las Vegas, Nav., pp. 349–354.

Bolles, R. & Cain, R. (1982), 'Recognizing and locating partially visible objects', *Intl. J. Robotics Research* **1**(3), 57–82.

Bolles, R., Horaud, P. & Hannah, M. (1984), 3DPO: A three-dimensional part orientation system, *in* M. Brady & R. Paul, eds, 'The 1st International Symposium on Robotics Research', MIT Press, Cambridge, Mass., pp. 413–424.

Boult, T. & Gross, A. (1988), On the recovery of superellipsoids, *in* 'Proc. of DARPA Image Understanding Workshop', Washington, D.C., pp. 1052–1063.

Bowyer, K. & Dyer, C. (1990), Aspect graphs: An introduction and survey of recent results, *in* 'Close Range Photogrammetry Meets Machine Vision, Proc. of SPIE', Vol. 1395, pp. 200–208.

Burns, J. & Kitchen, L. (1988), Rapid object recognition from a large model base using predicition hierarchies, *in* 'Proc. of DARPA Image Understanding Workshop', pp. 711–719.

Chin, R. T. & Dyer, C. R. (1986), 'Model-based recognition in robot vision', *Computing Surveys* **18**(1), 67–108.

Darrell, T., Sclaroff, S. & Pentland, A. (1990), Segmentation by minimal description, *in* 'Proceedings, 3RD International Conference on Computer Vision', Computer Society of the IEEE, IEEE Computer Society Press, Osaka,Japan, pp. 112–116.

Dickinson, S., Pentland, A. & Rosenfeld, A. (1990), Qualitative 3-D shape reconstruction using distributed aspect graph matching, *in* 'Proceedings, 3RD International Conference on Computer Vision', Osaka, Japan, pp. 257–262.

Dickinson, S., Pentland, A. & Rosenfeld, A. (1992), '3-D shape recovery using distributed aspect matching', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14**(2), 174–198.

Eggert, D. & Bowyer, K. (1989), Computing the orthographic projection aspect graph of solids of revolution, *in* 'PROC. of IEEE Workshop on the Interpretation of 3-D Scenes', IEEE, Austin, Texas, pp. 102–108.

Eggert, D., Bowyer, K., Dyer, C., Christensen, H. & Goldgof, D. (1992), The scale space aspect graph, *in* 'Proceedings, Conference on Computer Vision and Pattern Recognition', IEEE, Champaign, Il., pp. 335–340.

Fan, T. (1990), *Describing and Recognizing 3-D Objects Using Surface Properties*, SpringerVerlag, New York.

Fan, T., Medioni, G. & Nevatia, R. (1987), 'Segmented descriptions of 3-D surfaces', *IEEE Int. J. Robot. Automat.* **3**(6), 527–538.

Fan, T., Medioni, G. & Nevatia, R. (1989), 'Recognizing 3-D objects using surface descriptions', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**(11), 1140–1157.

Faugeras, O. & Hebert, M. (1983), A 3-D recognition and positioning algorithm using geometrical matching between primitive surfaces, *in* 'Proceedings, 8th International Joint Conference on Artificial Intelligence', Karlsruhe, West Germany, pp. 996–1002.

Ferrie, F. P. & Lagarde, J. (1989), Robust estimation of shape from shading, *in* 'Proceedings 1989 Topical Meeting on Image Und. and Machine Vision', Cape Cod, Massachusetts, pp. 24–27.

Ferrie, F. P., Lagarde, J. & Whaite, P. (1993), 'Darboux frames, snakes, and super-quadrics: Geometry from the bottom up', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15**(8), 771–784.

Ferrie, F. P., Mathur, S. & Soucy, G. (1993), Feature extraction for 3-D model building and object recognition, *in* A. Jain & P. Flynn, eds, '3D Object Recognition Systems', Elsevier, Amsterdam.

Flynn, P. (1992), Saliencies and symmetries: Toward 3D object recognition from large model databases, *in* 'Proceedings, Conference on Computer Vision and Pattern Recognition', Champaign, Il., pp. 322–327.

Flynn, P. & Jain, A. (1991*a*), 'BONSAI: 3-D object recognition using constrained search', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**(10), 1066–1075.

Flynn, P. & Jain, A. (1991*b*), 'CAD-based computer vision: From CAD models to relational graphs', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**(2), 114–132.

Flynn, P. & Jain, A. (1992), '3-D object recognition using invariant feature indexing of interpretation tables', *Computer Vision, Graphics, and Image Processing* **55**(2), 119–129.

Goad, C. (1983), Special purpose automatic programming for 3-D model-based vision, *in* 'Proc. of DARPA Image Understanding Workshop', Cambridge,Mass., pp. 94–104.

Grimson, W. (1987), On the recognition of parametrized objects, *in* 'The 4th International Symposium on Robotics Research', Santa Cruz, California.

Grimson, W. (1989), 'On the recognition of curved objects', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**(6), 632–642.

Grimson, W. & Huttenlocher, D. (1990), On the sensitivity of geometric hashing, *in* 'Proceedings, 3RD International Conference on Computer Vision', Osaka, Japan, pp. 334–338.

Grimson, W. & Lozano-Perez, T. (1987), 'Localizing overlapping parts by searching the interpretation tree', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **9**(4), 469–482.

Hansen, C. & Henderson, T. (1988), Towards automatic generation of recognition strategies, *in* 'Proceedings, 2ND International Conference on Computer Vision', Tampa, Fla., pp. 275–279.

Hansen, C. & Henderson, T. (1989), 'CAGD-based computer vision', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**(11), 1181–1193.

Hutchinson, S., Cromwell, R. & Kak, A. (1989), Applying uncertainty reasoning to model based object recognition, *in* 'Proceedings, Conference on Computer Vision and Pattern Recognition', San Diego, Calif., pp. 541–548.

Huttenlocher, D. & Ullman, S. (1987), Object recognition using alignment, *in* 'Proceedings, 1ST International Conference on Computer Vision', London, UK., pp. 102–111.

Ikeuchi, K. (1987*a*), 'Generating and interpretation tree from a CAD model for 3-D object recognition in bin-picking tasks', *Intl. J. Comput. Vision* **1**, 145–165.

Ikeuchi, K. (1987*b*), Pre-compiling a geometrical model into an interpretation tree for object recognition in bin-picking tasks, *in* 'Proc. of DARPA Image Understanding Workshop', Los Angeles, Calif., pp. 321–339.

Jain, A. & Hoffman, R. (1988), 'Evidence-based recognition of 3-D objects', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10**(6), 783–802.

Kak, A., Vayada, A., Cromwell, R., Kim, W. & Chen, C. (1987), Knowledge-based robotics, *in* 'Proc. IEEE Intl. Conf. on Robot. Automat.', New York, pp. 637–644.

Keren, D., Cooper, D. & Subrahmonia, J. (1992), Describing complicated objects by implicit polynomials, Technical Report 102, Brown University LEMS, Laboratory for Engineering Man/Macine Systems, Division of Engineering, Brown University, Providence RI 021912 USA.

Kim, W. & Kak, A. (1991), '3-D object recognition using bipartite matching embedded in discrete relaxation', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**(3), 224–251.

Koenderink, J. (1976), 'The singularities of the visual mapping', *Biological Cybernetics* **24**(1), 51–59.

Koenderink, J. (1979), 'Internal representation of solid shape with respect to vision', *Biological Cybernetics* **32**(4), 211–216.

Kriegman, D. & Ponce, J. (1989), Computing exact aspect graphs of curved objects: Solids of revolution, *in* 'PROC. of IEEE Workshop on the Interpretation of 3-D Scenes', IEEE, Austin, Texas, pp. 116–122.

Kriegman, D. & Ponce, J. (1990), 'On recognizing and positionind curved 3-D objects from image contours', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**(12), 1127–1137.

Kwong, J. & Kim, S. (1993), 'Uncertainy of features in planar object recognition and a new classifier', *Pattern Recognition Letters* **14**, 591–598.

Lagarde, J. (1989), Constraints and their satisfaction in the recovery of local surface structure, Master's thesis, Dept. of E.E., McGill Univ.

Lamdan, Y., Schwartz, J. & Wolfson, H. (1988), On recognition of 3-D objects from 2-D images, *in* 'Proc. IEEE Intl. Conf. Robot. Automat.', Philadelphia, PA., pp. 1407–1413.

Lamdan, Y. & Wolfson, H. (1990), Geometric hashing: A general and efficient model-based recognition scheme, *in* 'Proceedings, 3RD International Conference on Computer Vision', Osaka, Japan, pp. 238–249.

Lejeune, A. & Ferrie, F. (1993), Partitioning range images using curvature and scale, *in* 'PROC. IEEE Computer Society Conference on Computer Vision and Pattern Recognition', New York City, New York, pp. 800–801.

Lowe, D. (1985), *Perceptual Organization and Visual Recognition*, Kluwer, Norwell, MA.

Luenberger, D. G. (1984), *Linear and Nonlinear Programming*, Addison–Wesley Publishing Company, Reading, Massachusetts.

Marr, D. (1982), *Vision*, W.H. Freeman & Co., San Francisco.

Mood, A. M. & Graybill, F. A. (1963), *Introduction to the Theory of Statistics*, McGraw-Hill Book Company Inc., New York, N.Y.

Newman, T., Flynn, P. & Jain, A. (1993), 'Model-based classification of quadric surfaces', *Computer Vision, Graphics, and Image Processing:Image Understanding* **57**(2), 235–249.

Nilsson, N. J. (1965), *Learning Machines-Foundations of Trainable Pattern Classifying Systems*, McGraw-Hill Book Co., Stanford Research Institute, Menlo Park, California.

Pentland, A. (1987), Recognition by parts, *in* ICC (1987), pp. 612–620.

Pentland, A. & Sclaroff, S. (1991), Closed form solutions for physically based shape modelling and recognition, *in* T. Kanade & K. Ikeuchi, eds, 'IEEE Transactions on Pattern Analysis and Machine Intelligence: Special Issue on Physical Modeling in Computer Vision', Vol. 13(7), pp. 715–729.

Pfeiffer, P. E. (1978), *Concepts of Probability Theory*, 2nd revised edn, Dover Publications Inc., New York.

Press, W., Flannery, B., Teukolsky, S. & Vetterling, W. (1988), *Numeric Recipes in C – The Art of Scientific Computing*, Cambridge University Press, Cambridge, U.K.

Raja, N. S. & Jain, A. K. (1992), 'Recognizing geons from superquadrics fitted to range data', *Image and Vision Computing* .

Solina, F. & Bajcsy, R. (1990), 'Recovery of parametric models from range images: The case for superquadrics with global deformations', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**(2), 131–147.

Soucy, G. (1992), View correspondence using curvature and motion consistency, Master's thesis, Dept. of E.E., McGill Univ.

Soucy, G. & Ferrie, F. P. (1992), Motion and surface recovery using curvature and motion consistency, *in* 'PROC. Second European Conference on Computer Vision', Santa Margheri-ta Ligure, Italy, pp. 222–226.

Sripradisvarakul, T. & Jain, R. (1989), Generating aspect graph for curved objects, *in* 'PROC. of IEEE Workshop on the Interpretation of 3-D Scenes', IEEE, Austin, Texas, pp. 109–115.

Stein, F. & Medioni, G. (1992), 'Structural indexing: Efficient 3-D object recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14**(2), 125–145.

Subrahmonia, J., Cooper, D. B. & Keren, D. (1992), Practical reliable bayesian recognition of 2D and 3D objects using implicit polynomials and algebraic invariants, LEMS 107, Brown University LEMS, Laboratory fo Engineering Man/Machine systems, Division of Engineering, Brown University, Providence, RI 02912, USA.

Swain, M. (1988), Object recognition from a large database, *in* 'Proc. of DARPA Image Understanding Workshop', pp. 690–696.

Tarantola, A. (1987), *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*, Elsevier Science Publishing Company Inc., 52, Vanderbuilt Avenue, NewYork, NY 10017, U.S.A.

Thompson, D. & Mundy, J. (1987), Model-directed object recognition on the connection machine, *in* 'Proc. DARPA Image Understanding Workshop', Los Angeles, CA., pp. 93–106.

Whaite, P. & Ferrie, F. (1993*a*), Model building and autonomous exploration, *in* 'SPIE - Intelligent Robots and Computer Vision XII: Active Vision and 3D Methods', Boston, Massachusetts, pp. 73–85.

Whaite, P. & Ferrie, F. P. (1991), 'From uncertainty to visual exploration', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**(10), 1038–1049.

Whaite, P. & Ferrie, F. P. (1992), Uncertain views, *in* 'PROC. IEEE Computer Society Conference on Computer Vision and Pattern Recognition', Champaign, Illinois, pp. 3–9.

Whaite, P. & Ferrie, F. P. (1993*b*), Active exploration: Knowing when we're wrong, *in* 'PROC. Fourth International Conference on Computer Vision', Computer Society of the IEEE, IEEE Computer Society Press, Berlin, Germany, pp. 41–48.

Whaite, P. & Ferrie, F. P. (1993*c*), Autonomous exploration: Driven by uncertainty, Technical Report TR-CIM-93-17, Center for Intelligent Machines, McGill University, Montréal, Québec, Canada. *Submitted to PAMI, March 1994.* Available via ftp at `ftp.cim.mcgill.ca` in `/pub/3d/papers/tr-cim-93-17.ps.gz`.

Whaite, P. & Ferrie, F. P. (1994), Autonomous exploration: Driven by uncertainty, *in* 'Proceedings, Conference on Computer Vision and Pattern Recognition', Computer Society of the IEEE, IEEE Computer Society Press, Seattle, Washington, pp. 339–346.

Wu, K. & Levine, M. D. (1994), Recovering parametric geons from multiview range data, *in* 'Proceedings, Conference on Computer Vision and Pattern Recognition', IEEE Computer Society., Seattle, WA., pp. 159–166.

Document Log:

Manuscript Version 0
Typeset by $\mathcal{A}\!\mathcal{M}\!\mathcal{S}$-LaTeX — 8 May 1995

Tal Arbel

Center for Intelligent Machines, McGill University, 3480 University St., Montréal (Québec) H3A 2A7, Canada, *Tel.* : (514) 398-7158

E-mail address: `taly@@cim.mcgill.ca`

Typeset by $\mathcal{A}\!\mathcal{M}\!\mathcal{S}$-LaTeX