ACTIVE OBJECT RECOGNITION CONDITIONED BY PROBABILISTIC EVIDENCE AND ENTROPY MAPS

Tal Arbel

Department of Electrical Engineering

McGill University, Montréal

November 1999

A Thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

Abstract

This thesis introduces a novel method for sequentially accumulating evidence as it pertains to an active observer seeking to identify an object in a known environment. First, a probabilistic framework is developed, based on a generalized inverse theory, where assertions are represented by conditional probability density functions. In order to resolve ambiguous assertions from single view measurements, a sequential recognition strategy is developed in which evidence is accumulated over successive viewpoints until a definitive assertion can be made. The main contribution of the thesis is a strategy for conditioning the inference and the measurement processes with feedback from prior information.

The problem of interest is that of model-based recognition, where the task is to identify an unknown model from a database of known objects on the basis of parameter estimates. The robustness of the algorithm is illustrated through its application to two very different domains: (1) recognition of 3-D parametric models estimated directly from laser rangefinder data, (2) recognition of objects based on signatures extracted from optical flow images that they generate as they move with respect to a camera. The latter approach is completely novel and presents a major contribution to the field. Experimental results verify the strength of the approach at overcoming difficulties encountered in both contexts, as rapid convergence to the correct solution occurs in most cases.

With this framework in place, it is further shown how an active recognition strategy could be built by using *entropy maps* to guide an active observer along an optimal trajectory for confidently inferring object identity and pose, while minimizing the amount of data that must be gathered. Specifically, these maps are used to encode prior knowledge about the discriminability of objects as a function of viewing position. The thesis applies the strategy to the context of recognition based on optical flow signatures, and shows how a gaze-planning strategy can be formulated by using entropy minimization as a basis for choosing a next best view. Experimental results are presented which show the strategy's effectiveness at converging to the optimal solution in a minimum short number of steps.

Résumé

Cette thèse présente une nouvelle méthode pour accumuler séquentiellement des évidences dans le contexte d'un observateur actif cherchant à reconnaître un object dans un environnement connu. Premièrement, un cadre probabilistique est développé, basé sur une théorie inverse généralisée, où les assertions sont représentées par des fonctions de densité de probabilité. De manière à résoudre les assertions ambigües obtenues à partir de mesures uniques, une stratégie de reconnaissance séquentielle est développée et permet d'accumuler les évidences pour les différents points de vues visités en utilisant une chaîne d'accumulation bayesienne jusqu'à ce qu'une décision définitive puisse être faite. La contribution principale de cette thèse est une stratégie de mise en forme des procédés de déduction et d'acquisition de donn'ees par une contre-réaction utilisant l'information acquise apriori.

Le problème d'intérêt est celui de la reconnaissance d'objets basée sur leur modélisation, où la tâche consiste à identifier un nouveau modèle à partir d'une banque de données de modèles d'objets connus sur la base de l'estimation de paramètres. La robustesse de l'algorithme est illustré par son application à deux domaines très différents: (1) la reconnaissance de modèles paramtriques 3D obtenus directement de données télémétriques, (2) la reconnaissance d'objets basée sur la signature extraite du flot optique quils génèrent losqu'ils se déplacent par rapport à une caméra. Cette dernière approche est complètement nouvelle et représente une contribution majeure dans ce domaine. Des résultats expérimentaux permettent de juger de la puissance de cet approche pour aider à surmonter les difficultés rencontrées dans les deux applications, puisqu'une convergence rapide à la bonne solution est vérifiée dans la plupart des cas.

Avec ce cadre de reconnaissance d'objets établi, il est démontré comment il peut être utilisé pour établir une stratégie utilisant des cartes d'entropie pour guider un observateur actif le long d'une trajectoire optimale pour identifier l'objet de manière fiable tout en minimisant la quantité de données qui doit être acquise. Spécifiquement, ces cartes sont utilisées pour pré-encoder l'information disponible sur la discriminabilité des objets en fonction du point de vue. La thèse applique cette stratégie dans le contexte de la reconnaissance basée sur les signatures de flot optique et démontre comment une stratégie d'exploration visuelle peut être formulée en utilisant la minimization d'entropie pour choisir la prochaine meilleure vue. Des résultats expérimentaux montrent l'efficacité de la stratégie pour converger á la solution optimale dans un nombre minimum d'étapes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Computer vision research has always focused its efforts on solving a fundamental inverse problem: given a set of measurements, determine the underlying structure that generated it. An inference problem of particular interest is *object recognition*, the focus of a tremendous amount of research for over thirty years. Object recognition involves identifying objects in a given scene based on acquired sensor data. Most strategies attempt to solve the recognition problem from data acquired from single viewpoints. The primary difficulty in solving this problem lies its under-determined nature: ambiguous cases exist where more than one reasonable interpretation is possible. In order to attain a unique solution, many recognition approaches embed implicit constraints in their strategies. These may include constraints on the image formation parameters, or even on the shape of the solution space. As a result, these strategies may work in one situation, but cannot be easily modified to work elsewhere.

The thesis introduces a novel solution to the ambiguous recognition result, by describing a *sequential* recognition strategy, that chooses to improve uncertain interpretations by gathering evidence in all possible hypotheses over a sequence of views and *conditioning* the inference process with prior information from previous views until convergence to a plausible solution set is attained. The claim is that this leads to a more robust interpretation. As identification based on individual measurements can be ambiguous and even erroneous, a major contribution of the thesis is the development of a general Bayesian inference strategy to *qualify* assertions at each iteration. This way, recognition results can be assessed and actions can be planned based on them. In fact, this thesis illustrates how to condition the measurement process as well, by introducing an active recognition strategy that makes maximal use of *a priori* information in order to plan an optimal gaze trajectory in on-line recognition experiments. As this approach can certainly be used in other areas of computer vision, this strategy would be of interest to the field in general.

1. Motivation

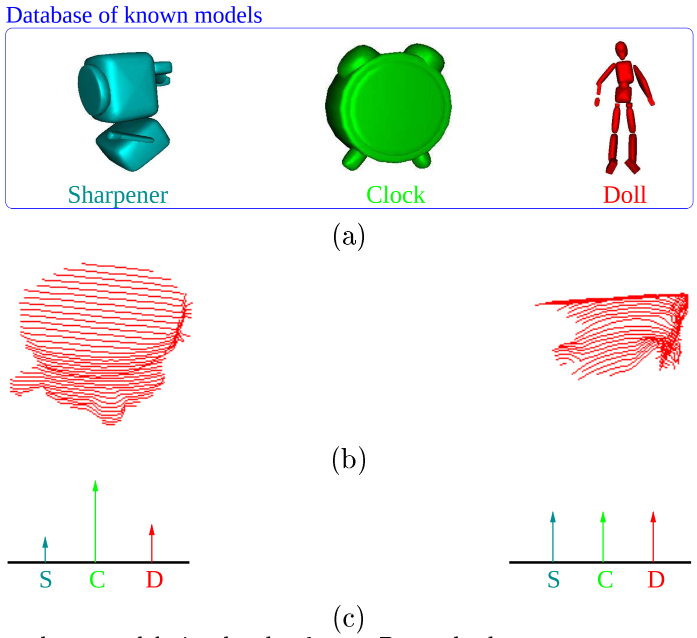
Object recognition has been one of the central problems addressed in the field of computer vision. Traditionally, it has been defined as the task of assigning a label to an unknown object, given a set of sensor measurements acquired during on-line experiments. The most common approach is to match the unknown to one of a predetermined set of known objects residing in a database computed off-line [**9**].

Despite the unifying goal of the research, the strategies developed over the years have varied dramatically in the types of features used, the way in which objects have been represented, as well as the matching strategies employed. Historically, earlier work made use of low level features, such a corners and edges, for matching objects. In later work, mid-level features extracted from reconstructed object surfaces became available for recognition. With the maturation of the field, high level models became available for object representation, from which features could be extracted for identification. Still, the scope of the types of objects that the models could represent was quite narrow. Additionally, industrial applications, such as assembly line automation, lead to the development of many specialized vision systems. A clear element in the slow rate of development of a general-purpose recognition system was the shortage of stable, high level object representatives.

Over time, pressure mounted for the development of vision systems that performed specific tasks in environments where modeling was not always feasible. This led to the advent of appearance-based recognition strategies, where rather than build intermediary object models, the raw data measurements were used as inputs for recognition. Once again, demands for a wide variety of applications, such a face and gesture recognition, spawned much research in this area. To date, there are still few general, high level computer vision systems capable of recognizing a wide variety of objects using one framework.

Despite varying contexts, the problem of recognition remains an *inverse problem* in the classic sense: given a set of measurements, determine the object that generated those measurements. Like many inverse problems, the recognition problem is ill-posed primarily due to the fact that several objects can give rise to identical measurements, generating a potential lack of uniqueness in the solution. In addition, the measurements themselves are uncertain.

The ill-posedness of the problem is illustrated in Figure1.1. Here, the task is to match the unknown object to one of the models in the database, based on laser-rangefinder measurements (seen in Figure 1.1(b)). The figure illustrates that, from some viewpoints, it is impossible to say which model generated the data. Figure 1.1(c) illustrates the recognition result of likelihoods of matches with each of the models in the database. In the second case in Figure 1.1(b), there is no clear winner as all the likelihoods are equal. This case illustrates the ill-posedness of the recognition problem, as a unique solution cannot be guaranteed.



On the top row are three models in the database. Beneath these, are two range images of one of the objects in the database taken from two different viewpoints. In the bottom row, one can see the probability distributions resulting from two recognition experiments, i.e. the likelihood of each model in the database: S (sharpener), C (clock) and D (doll) given the measurement. Notice that the system is unable to make a clear assertion in the second case.

FIGURE 1.1. Ill-posedness of recognition problem. (a) Database of objects, (b) Laser-rangfinder data from two viewpoints, (c) Resulting probability distributions.

The difficulty in solving the recognition problem, combined with its ill-posed and ill-conditioned nature, leads to shortcomings in existing strategies:

1. **Shortage of general solutions.** Despite its obvious importance, the issue of the ill-posedness of the recognition result is hardly addressed in the literature. The main strategy for overcoming it has been to impose strong, often implicit, constraints on the solution space in order to achieve a single object identification. Due to the hidden constraints, the solution may work well in a particular situation, but cannot be easily modified to work elsewhere. This is illustrated by the enormous variety

of approaches in existence, ranging from model-based to appearance-based, that are geared towards the specific task at hand and do not generalize easily. There is a lack of a general solution that applies to a large class of recognition problems.

2. **Most strategies are deterministic.** Most recognition strategies are deterministic in nature. This implies that the uncertainties of the problem are not explicitly represented in the solution. Further, these solutions often return a single object label from a set of measurements. Due to the possibility of an ambiguous (or even erroneous) result, it would be more instructive to qualify assertions so as to be able to assess the validity of the result. The best way to do this is through probabilistic methods.
3. **Most strategies are static.** Most recognition research focuses on the problem of identifying an object based on a data acquired from a single viewpoint of a scene which contains it. However, results based on a single image are clearly not sufficient. The world is dynamic, with the observer being able to move through it, gathering information and changing perceived notions of its constituents. Many active vision systems exist, particularly in the field of mobile robotics. However, very few active recognition strategies are reported in the literature. The philosophy of this work is that the ill-posedness of the result can, at least in part, be resolved by gathered data over time (i.e. over multiple viewpoints). The issue of how to do this is rarely dealt with in the literature.

In this thesis, I sought to develop and implement a system that addresses each of these issues in the most general manner possible, to ensure that the resulting strategies are applicable to a wide variety of inverse problems in vision, as well as in other fields. In the course of developing a solution, the following difficult problems were dealt with:

1. **Embedding Uncertainty into the Solution.** Inherent to all inference problems are the many sources of uncertainty that render them difficult (e.g. measurement uncertainty, variation in the database information). By choosing to ignore them, deterministic strategies are unwillingly biasing their solutions. Furthermore, as recognition from a single data set can lead to erroneous results, it would be instructive to assess the level of confidence in the result from a particular viewpoint. This would involve exposing the uncertainty in the recognition solution in a form that could be easily evaluated by an external agent. This leads to two major issues addressed by this thesis: *Can we enumerate and represent all sources of uncertainty as probability*

density functions? Can we represent the solution in the form of a probability density function?

2. **Temporal regularization.** From any particular viewpoint, there is a large amount of information that can be acquired for the recognition engine to process. Given the availability of a dynamic system, where data can be acquired from multiple viewpoints, there are various levels at which information can be combined over time, e.g. on the level of the data, on the surface level, etc. The amount of information available can be overwhelming. However, with limited resources, an active agent moving through a scene must be able to process information quickly and efficiently. Furthermore, prior information can serve to condition the inference in order to overcome the ill-posedness of the problem. This process can be referred to as "temporal regularization". This leads to the problem of *how to efficiently fuse information over time?*
3. **Planning gaze based on prior information.** In order to reduce the number of false assertions while processing the smallest amount of information, an efficient strategy would choose to actively gather data from viewpoints that lead to maximal object discriminability. The question is how to find these views. Many strategies attempt to compute them on-line and slowly move the sensor towards these positions. However, an interesting approach would be to find these viewpoints *a priori* and navigate based on them on-line. The question is: *How do you make maximal use of a priori information to best plan gaze (i.e. camera viewpoints) for recognition?*
4. **Recognition Based on Flow Images.** Many contexts exist where object-centered representations are not available. In these cases, one popular solution has been to apply appearance-based techniques to the problem of recognition. However, the major shortcoming of such an approach has been the lack of robustness in the face of slight changes in lighting and background. In overcoming this problem, an interesting approach would be one that considers differential image properties, which could be obtained from processing subsequent images acquired by a camera moving with respect to an object of interest. The question is: *Can a system be built that recognizes an object based on the optical flow images it generates as it moves with respect to a camera?*

In the next section, I will summarize my contribution to the solution of each of these problems, as well as to other open problems in the field.

2. Contributions of the Work

In this thesis, several contributions are made to the field of computer vision, in particular to the area of object recognition. The main results addressing the major difficulties encountered in recognition (as described in Section 1) are described here, as well as the other novelties of the work that would be of interest to the community at large.

2.1. Main Contributions.

1. ***The primary contribution of this thesis is the development of a sequential recognition system that accumulates evidence in the various object hypotheses over time by conditioning the inference process, at each iteration, with feedback from the previous view.*** Rather than force a solution from a single measurement, probabilistic evidence regarding the different object hypotheses is fused over viewpoints about the object of interest. The novelty in the approach is that the system "closes the loop" around inference by conditioning the recognition process in one view with prior information from the previous view. This leads to convergence to a correct object label in a short number of steps.
2. ***Another main contribution is the development of a general probabilistic framework for the solution of inverse problems such as sequential object recognition.*** The ill-posedness of the recognition problem is addressed by a Bayesian strategy that is used to (a) qualify the recognition result and (b) gather evidence over time. The generality of the framework is shown through its application to two very different recognition problems: 3D part recognition, and appearance-based recognition based on optical flow images. The application of the theory to either problem is novel. Presenting a unifying framework that is applicable to both is an important contribution to the field. This framework is based on an inverse theory first introduced in the field of Geophysics [**93**]. Its application to the problem of sequential recognition is novel as well.
3. ***A contribution of this thesis is an active recognition strategy that makes use of a priori information to guide a sensor to locations of maximal discriminability.*** A novel facet of the thesis is that maximal use of prior information is made, through building maps off-line that link inference ambiguity (based on entropy) to sensor position. A navigation strategy uses these maps during on-line

recognition experiments in order to guide the sensor to areas of maximal discriminability, thus converging to the correct recognition result more often and in a fewer number of steps. This strategy is completely novel and presents a major contribution to the field of recognition.

4. ***A main contribution of this thesis is the introduction of a strategy that recognizes objects based on the optical flow images they generate as they move with respect to a camera.*** Although optical flow images have been used for identifying actions, to my knowledge, they have never been used for the recognition of objects. This thesis shows how the general recognition framework can be applied to the problem of recognizing moving objects based on signatures in the optical flow images they generate.
5. ***This thesis introduces the notion of characteristic motions.*** This thesis extends the notion of characteristic views to the motion domain by defining a set of *characteristic motions* as a set of camera positions and movements (with respect to an object of interest) that capture a sufficient amount of the essential structure of the object to permit its recognition from novel motion sequences. The thesis illustrates how to gather and train on a small set of motions in practice, and empirically shows how this leads to recognition based on a wider set of movements.

2.2. Other Contributions. This thesis makes other contributions to the field of computer vision. Some of these are of particular interest to those in the area of object recognition. Others should be of interest to the field at large.

(i) This thesis contributes a strategy for determining the *informativeness* of the recognition result through information theoretic techniques applied to the resulting distribution. This is a significant result as external processes (such as active agents) can use this information in order to decide what action to take next (whether to gather more data for example). As most recognition strategies are deterministic, a probabilistic solution is rarely present in the literature.

(ii) Progress in finding solutions to 3D object recognition has been slow due, in part, to the lack of a stable bottom-up system that builds reliable models from sensor data. In this work, a solid contribution was made in the area of 3D object recognition, by building a sequential recognition module on top of a complete bottom-up system built at the Artificial Perception Lab at McGill University's Center for Intelligent Machines (see Chapter 3 for details).

(iii) Very few recognition strategies use superellipsoid models as part representatives due to problems of degeneracies in their representations. Previous iterations of this work [**1**] have shown how to overcome most of these difficulties. This thesis presents the first time superellipsoid models are used within a sequential recognition strategy.

(iv) This thesis will show how the 3D recognition framework can be easily extended to include the solution to other difficult problems that are still open topics of research such as: 3D multi-part object recognition, and verification of the existence of 3D parts in a scene.

(v) Due to the difficulty of the task, very few successful dynamic recognition systems have been implemented in practice. *This thesis builds two very different real active control systems for training and recognizing objects. The theory is demonstrated to be working in both of these vision domains.*

3. Literature Overview

As the work presented in this thesis covers various topics and does not fit into only one category of the mainstream literature, it was felt that it would be more appropriate to place specific comparisons with other work throughout the thesis where relevant. In particular, an overview of model-based recognition can be found in Chapter 3. A brief overview of existing appearance-based techniques, as well as those related to matching based on optical flow images can be found in Chapter 4. An overview of active vision, with emphasis on those approaches related to active recognition can be found in Chapter 5. Furthermore, other approaches will often be described after the presentation of the proposed approach so as to emphasize the differences in strategies.

4. Thesis Outline

The philosophy of this thesis is that by gathering evidence from multiple viewpoints, and conditioning the inference process with the results from each view, ambiguity inherent in recognition from single views will be resolved quickly and efficiently. A novel *sequential* recognition strategy is developed, where evidence in *all* possible hypotheses is fedback to the system at each iteration, leading to convergence to a plausible solution set in a short number of steps.

The problem of interest is object recognition based on a set (or sets) of parametric object descriptors. This differs from the problem of *parameter estimation*, where the goal is to recover a parametric object representation from measurements, given prior constraints on the shape of the solution set (i.e. from a predetermined set of functions) and an optimization metric to minimize error. Within this category are the *model-based representation* schemes that attempt to constrain the search for optimal parameters by forcing a fit to one of a set of predetermined model categories, stored in a database prior to the experiment [**32, 59, 72, 74, 105**]. This is a type of top-down fitting process. For the *model-based recognition* problem considered in this thesis, the assumption is that a bottom-up measurement process will provide a parametric descriptor of the object, from which object labels are inferred.

Other matching strategies in the literature that focus on parametric model-based recognition differ in their devised metrics for measuring the distance between parametric models in appropriate parameter spaces, e.g., Mahalanobis distance [**54**] and dot products [**76**]. Most strategies rarely include both the uncertainties in the parameters of the measured models and the ambiguities of the representations in the database. However, when fitting a model to noisy data, there is an inherent lack of uniqueness in the parameters that describe the model. In many cases, it is impossible to make a definitive statement as to which model fits the data best [**98**]. For this reason, rather than choose external constraints that would force the choice of one model over another, it would be more instructive to embed the uncertainty of the chosen description into the representation. This is precisely the approach that is taken in this thesis in computing the recognition solution.

Chapter 2 shows how to cast the recognition problem in probabilistic terms from the point of view of an inverse theory [**93**]. The strategy makes each of the sources of information available to the recognition engine explicit. These include: (i) information from measurements, (ii) it a priori information about models, and (iii) information from the physical theory relating measurements and models. Each of these sources of knowledge are then represented as probability density functions that can be easily combined using standard Bayesian techniques. The result is an *a posteriori* (discrete) conditional density function, which describes the degree of likelihood of the unknown object matching each of the objects in the database, given a set of measurements. Application of the theory to a particular inverse problem becomes the straight-forward task of defining the appropriate form of each of the component probability density functions.

Based on this result, the thesis shows how the resulting distributions can be used to sequentially recognize an unknown object by accumulating evidence in the various hypotheses at the probabilistic level. Chapter 2 illustrates how the probabilities are conditioned based on the recognition result from the previous view. The method is a Bayesian chaining strategy whereby the posterior from one view is fed in as the prior for the next view.

The robustness of the strategy is shown through application of the theory to two very different, and very difficult recognition problems: (i) 3D parametric object recognition (Chapter 3), and (ii) appearance-based object recognition based on optical flow images (Chapter 4). The method will be used to resolve uncertainties associated with the recognition of objects in both of these challenging contexts. This will be verified through experimentation with two *real* vision systems.

Chapter 3 focuses on the application of the theory to three-dimensional object recognition, where objects are represented by parametric shape descriptors (i.e. models) such as superellipsoids [**12, 13, 39, 79**], deformable solids [**27, 76**], and algebraic surfaces [**91**]. In this context, models are constructed through a process of *autonomous exploration* [**98, 102, 103**] in which a part-oriented, articulated description of an object is inferred through successive probes with a laser range-finding system. The bottom-up system that was used to build the representations consists of several stages of computation. These include: data acquisition, surface reconstruction, segmentation into convex parts, and fitting the models to the individual data patches (here, superellipsoids were used). The inherent noise uncertainties associated with each stage of the bottom-up system renders recognition based on the resulting parameters difficult, and often unpredictable. Empirical evidence indicates that applying the sequential recognition strategy to this problem leads to the quick resolution of ambiguous results through the accumulation of evidence over different viewpoints.

Recognition in cases where an object-centered model representing the data is not readily available is discussed in Chapter 4. In addition, theses cases are more general, the scene more complex, the data noisier and the conditions less controlled. This leads to application of the theory to an appearance-based recognition strategy that builds on the works of Nayar et al. [**71**] and Turk et al. [**95**]. Appearance-based recognition has gained tremendous popularity in the Computer Vision community over the past few years due to its fast indexing time and its ability to deal with complex scenes where model-building becomes infeasible. Most of the existing strategies [**16, 19, 53, 56, 11, 71, 75, 87, 95, 66**] focus on applications (e.g. face recognition), where using only the raw grey-scale images as input to an appearance strategy has worked quite well. The major drawback is that tight control over lighting and background has to be enforced in order to ensure similar appearance.

In order to overcome these drawbacks, inputs for an appearance-based recognition strategy are taken to be the optical flow images induced on the retina by the relative motion between camera and object. This approach offers some advantages, as the optical flow field is relatively invariant to scene illumination (provided it stays constant during the acquisition). In addition, it provides some figure/ground separation in the case of a moving object and a stationary observer.

Recognition based on flow images presents its own series of problems. It is well-known that optical flow images are composed of a mixture of information regarding object shape, object texture, camera geometry, and relative motion. By imposing constraints on the camera geometry (orthographic projection), and on the possible relative motions, it will be shown empirically that it is indeed possible to extract a signature that is largely correlated with object shape from the instantaneous optical flow for the expected range of motions. The resulting noisy and relatively low resolution images are perfect candidates for testing the sequential recognition system, in that a single such image is rarely sufficient to make assertions about object identity. The hypothesis is that a sequential strategy should resolve ambiguous interpretations over time, should a maximum *a priori* (MAP) solution prevail. The system shows empirical robustness in its ability to recognize based on previously unseen motion patterns.

Both recognition problems were tested in two sets of real recognition experiments (Chapter 6). Two different control systems were built and the strategy was tested in a series of experiments. The results in both scenarios indicate the system's ability to converge to the correct object hypothesis in a short number of iterations. The strategy was proven to be general, working well in both widely different systems. This generality will permit easy adaptability to other recognition contexts.

The concept of sequential recognition was then extended by the introduction of an *active recognition* strategy that chooses an optimal trajectory in order to recognize an unknown object. Chapter 5 considers the case where a mobile agent is moving through a scene with the task of identifying the objects within it. Limits are placed in the number of steps it can take. The specific application involves a monochrome television camera mounted on the end effector of a gantry robot. The gantry is free to move about the workspace where different test objects are placed (i.e. stationary environment). As the camera moves relative to an object, an optical flow pattern, induced on the image plane, results in a discrete image sequence. The task that the system must perform is to generate an optimal trajectory (i.e. the shortest sequence) that will result in the correct object identity.

The first step in addressing this problem is to build a model of how measurement parameters influence the confidence for the resulting assertions. In Section 1, the concept of Shannon entropy is used to define a measure of ambiguity for the resulting a posteriori probability distributions called an *entropy map*. Here the entropy map is used to relate ambiguity to camera (viewing) position and serves as the basis for an active vision system. By choosing viewpoints that minimize ambiguity, the system seeks out locations that are maximally informative. Hence fewer observations are required to arrive at a confident assertion. Other strategies use entropy on-line in order to maximize information gain during navigation [**22, 23**]. The key difference in this work is that the a *priori* information available is maximized, by building the entropy maps *off-line* and using them to guide the on-line navigation. The hypothesis, which is supported by the experiments presented in Chapter 6, is that the correct assertion will become apparent over time. It is shown that gaze planning using the entropy map further increases overall robustness (i.e. the probability of making correct assertions) by avoiding viewpoints that are inherently ambiguous.

CHAPTER 2

Sequential Bayesian Object Recognition

The problem of object recognition consists mainly of inferring from measurements of an unknown object, that model (or models) which most closely represents it in a database of known objects. Hadamard [**44, 45**]) defines a problem to be well-posed if a solution: (a) exists, and (b) is unique, and (c) varies continuously with the data. Like many inverse problems, the recognition problem is ill-posed primarily due to the fact that that several models can give rise to identical measurements. As a result it is not possible to identify the unknown object uniquely, and several solutions may be equally viable. Imposing a closed-world assumption (i.e. only the database objects can exist) overcomes the first condition. However, requiring a single object identity to be chosen often leads to instability in the solution, and a violation of the condition of continuity. In addition, the solution process can be ill-conditioned (i.e. not robust against noise), and experimental uncertainty gives rise to highly uncertain measurements. 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, is not portable to other contexts.

For these reasons, a solution to the recognition problem is sought that satisfies the following requirements: (i) The solution should be robust. (ii) The solution should be general enough so that its application to a wide variety of problems becomes possible. (iii) The approach should place emphasis on the evidence from the data, rather than on strong *a priori* constraints. (iv) Rather than generate a unique object label from a single data set (which might be impossible with competing hypotheses), qualification of recognition assertions is required. This will permit higher level processes to assess the result and decide what action to take next (This might include deciding whether a particular viewpoint is informative or not in terms of recognition, or deciding where to move the sensor next to gather more data). (v) The strategy needs to be efficient and work at minimal computational expense.

This thesis addresses these requirements through the introduction of a *sequential recognition* strategy that seeks to improve recognition results by accumulating evidence in the various hypotheses over a sequence of viewpoints. Rather than having to make a decision based on limited information from one data set alone, ambiguous results can be resolved *over time*, a concept we refer to as *temporal regularization*. In this fashion, convergence to a correct winner can be achieved in relatively few steps and a more robust solution can be attained.

For a sequential recognition process to work, the strategy should be able generate a measure of confidence in various hypotheses at each iteration. This can indeed be accomplished through application of the inverse theory first presented by Tarantola [**93**] in the field of geophysics. This theory provides a solution to the general inverse problem, and earlier work [**7, 1, 3, 8**] has shown that it can be easily applied to the problem of parametric shape recognition. The beauty of the theory lies in the fact that all the sources of knowledge used to obtain inverse solutions are made explicit, so if conditioning is required, the necessary assumptions about that knowledge are apparent and can be examined to see if they are realistic. The result is expressed in the form of a posterior conditional probability density function, which, in the context of object recognition, assigns confidence in each of the objects in the database. This complies with the standard definition of *statistical inference*, which refers to the assignment of probabilities to hypotheses, given a defined hypothesis space and a particular data set [**67**]. By rendering the solution explicit, the results are available to higher level processes for assessment. Decisions about the usefulness of the solution can only be made in the domain of the task at hand. Because of experimental uncertainties, there is always the possibility that an object will be identified incorrectly. Only the task can know if the likelihood of errors is acceptable.[1]

This leads to the interesting issue of what to if the level of errors is not acceptable. Because the sources of knowledge are explicit they are not only visible to the operational

[1]Solutions to such problems fall under the category of *decision theory*, which refers to choosing between alternative actions on the basis of the probabilities in order to minimize the expectation of a loss function.

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. Whaite et al. [**100, 101**] demonstrated that *autonomous exploration* functions well by providing feedback at the model building level. The intention of this chapter is to show that, with the aid of this theory, one can incorporate feedback from the recognition task. Chapter 5 will illustrate that the optimal sensor position for recognition can be incorporated into the feedback process as well.

The question then arises: At what level of representation should the evidence in the various models be accumulated? One approach would be to merge data on the level of features (e.g. surface geometry) as the agent moves through a scene. At each iteration, the degree of confidence in each hypothesis could be recomputed based on the total data gathered thus far. Unfortunately this would be computationally prohibitive, largely due to the expense of data fusion [**88**]. A more efficient strategy would be to perform low-level processing on each data set separately and avoid the fusion problem at the data level altogether by seeking instead to combine information at the highest level of inference, at the level of belief in the various hypotheses. In accordance with Marr's "Principal of Least Commitment" [**69**], all decision-making is deferred to the highest level of processing. An active agent could then gather evidence until the composite belief associated with a particular hypothesis exceeds a prescribed figure of merit. This chapter will illustrate how such evidence can be accumulated efficiently through a Bayesian chaining strategy, whereby the entire posterior distribution from one viewpoint is fed back as a prior for the next view.

Section 1 of this chapter will describe the Bayesian recognition strategy which is based on the inverse theory, and illustrate its pertinence to the problem of parametric object recognition. For clarity, the inverse theory will be described using standard probabilistic notation, rather than the more unfamiliar notation introduced by Tarantola. In order to illustrate that the final inverse solution degenerates to the same result as those obtained through standard Bayesian techniques, the analogous Bayesian solution will be derived. The result naturally leads to a strategy, described in Section 2, for sequentially recognizing objects through Bayesian chaining of the posterior beliefs of one stage as priors for the next stage. Later chapters will illustrate how the framework can be applied to two different real-world recognition problems.

1. Bayesian Recognition

The classic definition of the recognition result is the assignment of a single label to an unknown object in a scene [**9, 25**]. The majority of the approaches taken in obtaining this solution are deterministic in nature, but the benefits of statistical approaches to inference over deterministic ones are many-fold. First, all sources of uncertainty entering the formulation are explicitly represented within the probability density functions. This is advantageous as many approaches ignore the effects of uncertainty in their solution. Second, a mathematically sound recipe for combining the information exists, therefore removing the need for artificially imposed heuristics that are not based on mathematical foundations. Subjective prior information that is available can still be factored into the formulation, with emphasis on the information from the data sets as they are collected. Finally, statistical strategies permit qualification of the recognition result, permitting assessment of the results by the higher level processes which rely on them to make decisions. It is precisely this qualification that permits the extension to a sequential recognition strategy.

In this work, the recognition problem is re-defined as follows: Given a set of measurements of an unknown object, compute the degree of likelihood of it matching each of the objects in a stored database. In order to outline the solution to this problem, some notation must be introduced. The set of K possible object hypotheses is defined as $\{O_i\}|_{i=1..K}$. A Bayesian recognition strategy is introduced whereby the posterior beliefs over the entire set of K object hypotheses, given a set of measurements $\mathbf{d}$ in a scene, is represented by a posterior discrete (conditional) probability density function $p(O_i|\mathbf{d})|_{i=1..K}$. A framework for obtaining this result applying Tarantola's general inverse theory to the problem of shape recognition was first introduced in [**7, 1, 3, 8**]. The advantage of applying the theory to the recognition problem is that it defines a formal recipe for explicitly enumerating all the information available in solving this inverse problem, defining each as a probability density function, and introducing a recipe for combining them in order to obtain the inverse solution.

In this section, it will be shown how this formalism can be used to build a general framework for the inverse problem of inferring object labels based on measurements. For clarity and simplicity, the theory is presented using standard probability notation, rather than the notation introduced by Tarantola. In many cases, the final result reduces to a standard Bayesian formulation. In order to illustrate this point, the analogous Bayesian result will be derived.

1.1. Tarantola's Inverse Theory. Tarantola's inverse theory begins with the description of a general physical system that can be conveniently visualized as a mapping between two different spaces: the *model space* M and the *data space* D. It is assumed that M and D are vector spaces with a finite number of real valued parameters. M is defined 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 the theory is that knowledge of a physical parameter (model or measurement) is subjective in that it varies from observer to observer depending upon the data gathered and on the observer's prior model. This subjective knowledge can be quantified by a probability, which Tarantola defines as a *state of information*, assigning a positive number that reflects our belief that the true value of the parameter lies within some given range. For a vector space the rule is represented by a probability density function.

Thus, the first postulate of the inverse theory is that knowledge about a set of parameters is described by a probability density function over the parameter space. This involves devising appropriate density functions in order to represent what we know about the world. However, probability theory does not specify 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. Deciding whether the chosen form is appropriate requires experimentally confirming predictions.

The general inverse problem illustrates how to compute a probabilistic representation of the posterior information on model parameters. In attaining this solution, the theory makes explicit each of the relevant sources of knowledge about the world, and states that each should be represented by a probability density function. Each of the sources of information specified by the theory will be described next, with intentional omission of its notation for clarity.

1. **Information Obtained from Physical Theories.** A physical theory is the solution to the *forward problem*. It describes how to predict the values of the observed

data given a model. This includes information on the physical correlation between observable parameters and model parameters.

2. ***A Priori* Information on Models.** It is often the case that other specific information is available about the models that can be useful. This can include knowledge that only a finite number of models can exist. Such information can provide a powerful constraint on the solution set, eliminating the need to examine all possible 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. This type of knowledge is referred to as *a priori* information on models.
3. **Information From Measurements.** Much of the knowledge about a problem comes in the form of experimental measurements of observed parameters. All instruments are subject to varying degrees of uncertainty so knowledge of the observable parameters is imperfect. The theory requires the explicit specification of the information from measurements, and its associated uncertainties.

Each of these sources of information can be represented by probability density functions, whose form will vary depending on the context of the problem to be solved. Tarantola then describes a simple framework for combining each of these sources of information by defining a logical operation of *conjunction* of states of information [**93**, pages 29–31]. The definition involves combining the joint prior information with the joint theoretical information, to get the a posteriori state of information on models. The flowchart of the inverse theory can be found in Figure 2.1.

In the next section, the inverse theory will be described within the context of object recognition.

1.2. The Inverse Theoretic Approach to Recognition. With this framework in place, the general inverse theory is applicable to any inverse problem where the physical system can be characterized by a set of parametric descriptors. As such, the theory is applicable to the problem of parametric shape recognition where raw data extracted from a scene is represented by a set of features in parametric vector form. In the remainder of this section, an inverse theoretic framework is built for the problem of parametric shape recognition. However, it is important to note that this framework is generalizable to a wide variety of such inverse problems.

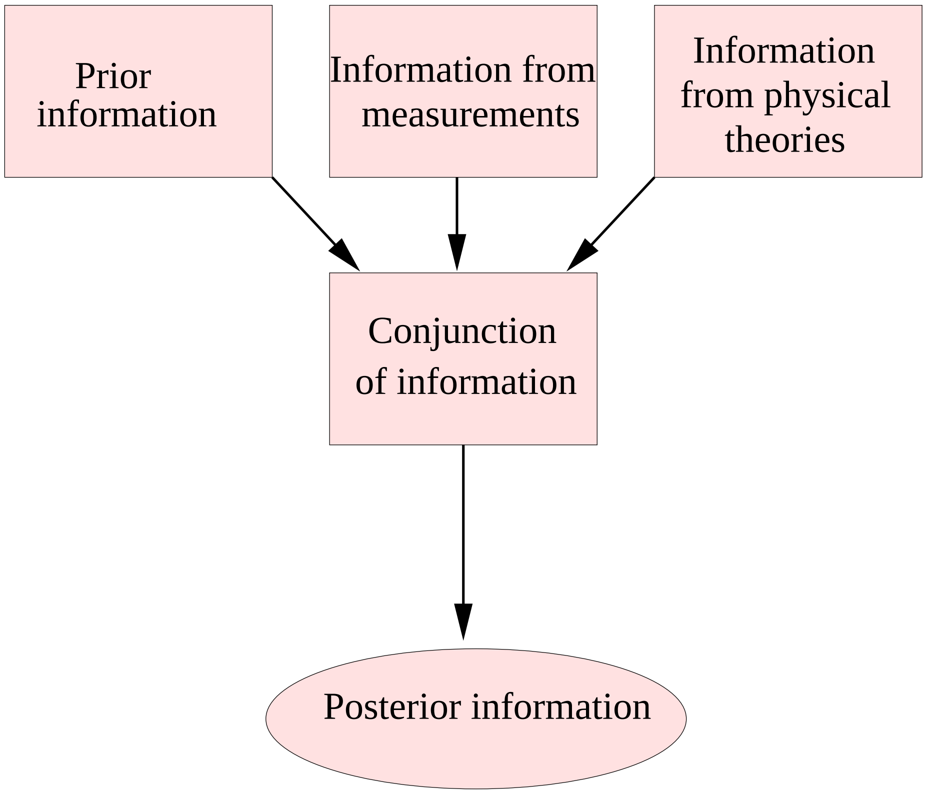


FIGURE 2.1. Flowchart of inverse theory.

The solution to the general inverse problem is a probability density function describing the beliefs in a set of parametric models. For the problem of object recognition, what is desired is a probability density function describing the belief in a set of object *labels*. The general inverse theory can easily be used to attain this goal. This is accomplished through an additional level of inference, involving a mapping between the model space M and the space of all possible model labels, denoted O. This space can be visualized as a separate space, where each point is assigned a particular label i (i.e. the object corresponding to this label is O_i), or as simply a subspace of the set of all possible models parameters M, i.e. object labels can be linked to a particular set of model parameters. For the particular case of object recognition, the space O is discrete, with each point in O denoting a particular object in the database.

The general inference chain for this type of inverse problem can be represented by:

$$\mathbf{d} \rightarrow \mathbf{m} \rightarrow O_i \tag{2.1}$$

This implies a level of inference from raw data, $\mathbf{d}$, to model parameters, $\mathbf{m}$, and one from parameters, $\mathbf{m}$, to labels, O_i.

Each source of knowledge enumerated by the inverse theory can then be described within this context:

1. Information Obtained from Physical Theories. Within this context, the forward problem consists of predicting the values of the observed parameters, $\mathbf{m}$, given a particular model label, O_i. In the case of parametric shape recognition, a measurement of an object produces a vector description of object features. The theory is rarely exact due to modelization uncertainties. This information is represented by a conditional probability density function, $p(\mathbf{m}|O_i)$, describing how the observed parameters $\mathbf{m}$ vary given a particular object, O_i. This information is usually obtained during a training or learning phase of the recognition process. The general form of this function over the entire space of objects is $p(\mathbf{m}|O)$.

2. *A Priori* Information on Models. Here, information from unspecified sources about the kinds of models which exist in the world is explicitly represented by the probability density function $p(O)$. The theory does not specify the form of the function. For the particular case of object recognition, this implies *a priori* belief in each of the objects in the database, $p(O_i)$, i.e. a discrete distribution.

3. Information From Measurements. In the first stage of inference, data extracted from the scene lead to a set of parametric descriptors. In fact, estimating the parameters of the underlying model that generated the data set is an inverse problem in and of itself. Solution to such a problem using an inverse theory framework was addressed in [**101**]. The result of such a measurement is represented by the probability density function $p(\mathbf{m}|\mathbf{d})$.

With this framework in place, the next section will illustrate how the inverse theory can be directly invoked to obtain the desired probabilistic solution to the inverse problem.

1.3. The Inverse Theory Solution to Recognition. In order to obtain the desired solution, each of the sources of information are combined using the definition of the conjunction of states of information. This involves combining the joint prior information, $p(\mathbf{d}, \mathbf{m})$, with the joint theoretical information, $p(\mathbf{m}, O)$ to get the a posteriori state of information, $p(O, \mathbf{m}|\mathbf{d})$. Previous work [**1, 8**] has shown how to derive this result using the inverse theory notation. Using the standard probabilistic notation defined above, the conjunction of states of information yields:

$$p(O, \mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{m}, O)\; p(\mathbf{d}, \mathbf{m})}{p(\mathbf{d})\; p(\mathbf{m})}. \tag{2.2}$$

The a posteriori marginal information about the objects can be obtained by marginalizing out the parametric information:

$$p(O|\mathbf{d}) = \int_M p(O, \mathbf{m}|\mathbf{d})\; d\mathbf{m}. \tag{2.3}$$

Substituting the terms in (2.2) gives:

$$p(O|\mathbf{d}) = \int_M \frac{p(\mathbf{m}, O)\; p(\mathbf{d}, \mathbf{m})}{p(\mathbf{d})\; p(\mathbf{m})}\; d\mathbf{m}. \tag{2.4}$$

Standard probability techniques indicate that:

$$\begin{aligned} p(\mathbf{m}, O) &= p(\mathbf{m}|O)\; p(O), \\ p(\mathbf{d}, \mathbf{m}) &= p(\mathbf{m}|\mathbf{d})\; p(\mathbf{d}). \end{aligned} \tag{2.5}$$

This leads to:

$$\begin{aligned} p(O|\mathbf{d}) &= \int_M \frac{p(\mathbf{m}|O)\; p(O)\; p(\mathbf{m}|\mathbf{d})\; p(\mathbf{d})}{p(\mathbf{d})\; p(\mathbf{m})}\; d\mathbf{m}, \\ &= \int_M \frac{p(\mathbf{m}|O)\; p(O)\; p(\mathbf{m}|\mathbf{d})}{p(\mathbf{m})}\; d\mathbf{m}. \end{aligned} \tag{2.6}$$

The final recognition result using this framework becomes:

$$p(O|\mathbf{d}) = p(O) \int_M \frac{p(\mathbf{m}|O)\; p(\mathbf{m}|\mathbf{d})}{p(\mathbf{m})}\; d\mathbf{m}. \tag{2.7}$$

This is the *inverse solution* to the general inverse problem as it applies to the problem of object recognition.

In the context of object recognition, the *existence* of the solution to the inverse problem simply means that $p(O|\mathbf{d})$ is not identically null. If it were then this would indicate incompatibility between the theory, the experimental results, and *a priori* assumptions.

An important result of the applying this framework to the recognition problem is that the final solution, $p(O|\mathbf{d})$, is *unique*, in that there is only one uniquely defined solution. In this sense, the ill-posedness of the problem is alleviated. However, this is not the standard consideration for the term ill-posed. It is actually the shape of the distribution that determines the ill-posedness of the recognition result in the classic sense. For instance, many peaks in the distribution may lead to the conclusion that the identification is non-unique. As such, the usual notion of ill-posedness is left as an operational issue, dependent on the task at hand.

The flowchart of the inverse theory applied to recognition can be found in Figure 2.2.

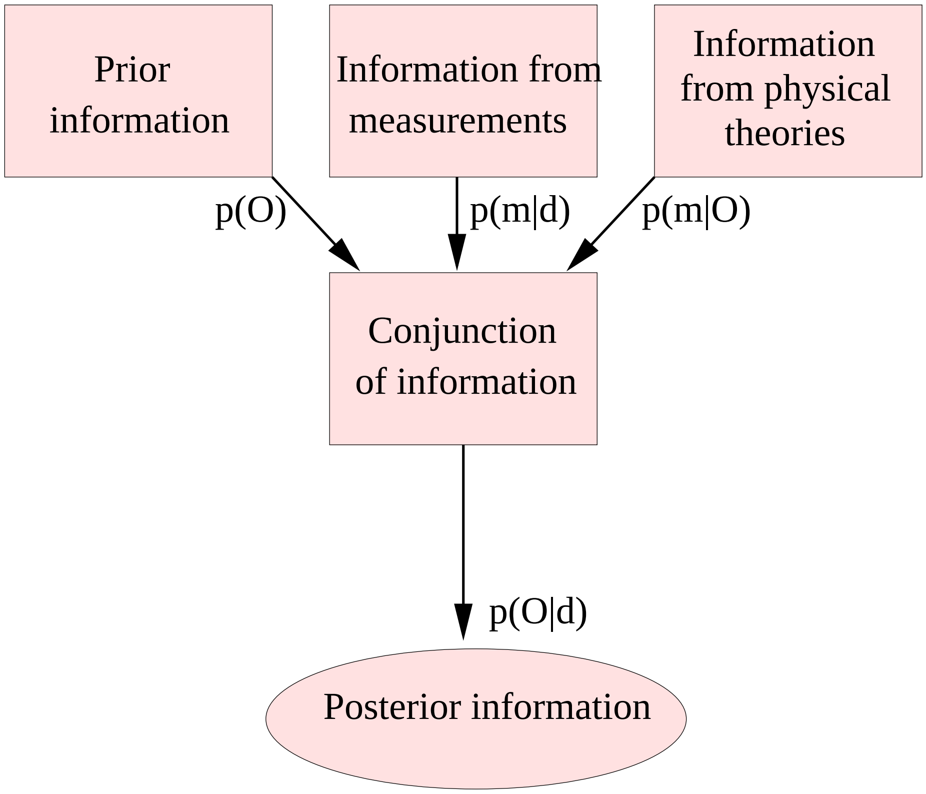


FIGURE 2.2. Flowchart of inverse theory applied to recognition.

1.4. The Bayesian Recognition Solution. The inverse theory provides a general framework for solving inverse problems. Under the conditions described in this chapter, the final inverse solution degenerates to the Bayesian solution [**93**, page 61]. Although the framework is a comprehensive strategy for solving inverse problems, at times it may be more convenient to describe it in strict Bayesian terms. In this section, the analogous Bayesian solution will be described, in order to illustrate that the result could have been alternatively derived using standard Bayesian techniques.

The goal of the system is to compute a degree of confidence in a set of object labels in a predetermined database, given a set of measurements. The a posteriori information about the object labels is given by the marginal probability density function:

$$p(O|\mathbf{d}) = \int_M p(O, \mathbf{m}|\mathbf{d})\ d\mathbf{m}. \tag{2.8}$$

The equation for the marginal a posteriori density function becomes

$$p(O|\mathbf{d}) = \int_M p(O|\mathbf{m}, \mathbf{d})\ p(\mathbf{m}|\mathbf{d})\ d\mathbf{m}. \tag{2.9}$$

Simplifications can be attained based on the following postulate:

POSTULATE 1 (sufficient statistic). $\mathbf{m}$ *is a sufficient statistic for* O*, therefore:*

$$p(O|\mathbf{m}, \mathbf{d}) = p(O|\mathbf{m}). \tag{2.10}$$

Thus $p(O|\mathbf{d})$ reduces to:

$$p(O|\mathbf{d}) = \int_M p(O|\mathbf{m})\ p(\mathbf{m}|\mathbf{d})\ d\mathbf{m}, \tag{2.11}$$

which, by invoking Bayes' rule:

$$p(O|\mathbf{m}) = \frac{p(\mathbf{m}|O)\ p(O)}{p(\mathbf{m})}. \tag{2.12}$$

becomes:

$$\begin{aligned} p(O|\mathbf{d}) &= \int_M \frac{p(\mathbf{m}|O)\ p(O)\ p(\mathbf{m}|\mathbf{d})}{p(\mathbf{m})}\ d\mathbf{m}, \\ &= p(O) \int_M \frac{p(\mathbf{m}|O)\ p(\mathbf{m}|\mathbf{d})}{p(\mathbf{m})}\ d\mathbf{m}. \end{aligned} \tag{2.13}$$

Notice that the solution using Bayesian techniques led to the same final solution as in the case of the inverse solution (Equation (2.7)).

Should the posterior information from measurements, $p(\mathbf{m}|\mathbf{d})$, not be available, then the solution can be equivalently expressed in terms of the physical theory for measurements, $p(\mathbf{d}|\mathbf{m})$, as:

$$p(O|\mathbf{d}) = \frac{p(O)}{p(\mathbf{d})} \int_M p(\mathbf{m}|O)\ p(\mathbf{d}|\mathbf{m})\ d\mathbf{m}, \tag{2.14}$$

where $p(\mathbf{d})$ is a constant of proportionality such that:

$$p(O|\mathbf{d}) \propto p(O) \int_M p(\mathbf{m}|O)\ p(\mathbf{d}|\mathbf{m})\ d\mathbf{m}. \tag{2.15}$$

Equation (2.15) is the *Bayesian solution* to the general parametric recognition problem. In fact, its form is general and can be applied to inverse problems where the posterior belief in (parametric) model hypotheses is desired.

In general terms, the *evidence* for the hypotheses is determined by the $p(\mathbf{d}|O)$ term. In this formulation, the evidence refers to:

$$p(\mathbf{d}|O) = \int_M p(\mathbf{m}|O)\, p(\mathbf{d}|\mathbf{m})\; d\mathbf{m}. \tag{2.16}$$

The evidence, $p(\mathbf{d}|O)$, is the Bayesian's transportable quantity for comparing alternate models [**67**]. As data are gathered, the subjective part of inference, the original prior $p(O)$, will typically be overwhelmed by the objective term, the evidence. This further illustrates that the strategy emphasizes the evidence from the data in determining the posterior results, rather than depending on strong subjective priors. Later chapters will show how the form of the evidence will change based on the context of the problem to be solved.

Although the final Bayesian solution was equivalent to that obtained using the inverse theory, following the formulation of the inverse theory offers an advantage in that it provides an easy recipe for solving inverse problems. In general, the problem with obtaining the solution using the standard Bayesian approach is that the formalism begins with a general statement and requires the expansion of this solution into its components using standard probability rules. As there are many ways to expand the equation, this requires foresight into what terms will be of interest in the final solution. The beauty of using a formalism such as one developed in this thesis is that all the components of the solution are explicitly specified first, and a recipe for their combination provides an easy solution for a wide set of inverse problems. As a result, the strategy can easily be followed by other vision researchers with similar inverse problems.

The result of applying the theory to the context of parametric object recognition is that rather than a solution in the form of a single object identity (as is popular, see [**9, 25**]), the system produces a probability density function describing the likelihood of the various objects in the database having generated the measurements. Later chapters will show how solving the inverse problem in different contexts becomes a simple matter of defining the form of each of the probability density functions, and representing them in this final solution.

1.5. Other Work Related to Statistical Recognition. Other statistical strategies for object recognition exist in the literature. Work by Subrahmonia, Keren, and

Cooper [**54, 91**] involved the application of Bayesian techniques to the problem of recognition of objects modeled by fourth order polynomials. The key difference in their work is in the methods used to attain the solution. The inverse solution forces all information to be made explicit, and provides a general recipe for its combination in a formal and structured fashion. In addition, they were interested in generating a single object identity. Here, a fundamental goal is the propagation of evidence over time, requiring a qualification of the recognition results at each stage of processing.

Other strategies use statistical methods to perform recognition through alignment [**77, 97**]. This implies attempting to find the transformation that takes the unknown image features into model coordinates. This is a different problem than the one considered here, as it is closer to a view registration problem, using probabilities to refine a search for features, rather than a general recognition problem.

Moghaddam et al. [**11**] present a Bayesian approach to a different type of problem: face recognition. Here, statistical methods are used to model the probability of two different classes of face variations: intra-personal (variation in appearance of the same person) and extra-personal (variation in appearance due to difference in identity).

Nair and Aggarwal [**70**] also use Bayesian strategies for a part-based recognition strategy. An aspect graph approach is taken, and the strategy is to train on the appearance of parts from various viewpoints. Parts are represented as EM mixture models, combined from data from different views. Prior beliefs in each part are computed through a saliency measure that computes how far part is from the rest in the database. Unfortunately, they do not specify the details of the workings of their approach.

Shimshoni and Ponce [**86**] use local geometric descriptors, such as angles and length ratios, in a probabilistic recognition strategy that links identification to pose estimation. The idea is to find peaks in the probability of correct match surface, where the probability of a match is increased if spatial coherence (i.e. pose) is verified.

Other strategies that use statistical methods in a sequential recognition domain will be described in the next section.

2. Accumulation of Evidence

The majority of the recognition strategies in the literature have a static approach to recognition, and therefore would be satisfied with basing an interpretation on a single data set. This would imply that the final recognition solution outlined in Equation (2.7) would be satisfactory for most recognition problems. The key difficulty with such a solution is that even with strong *a priori* knowledge, there are still ambiguous cases where a "significant" belief in more than one model exists. Operationally, this is further complicated by the noise and quantization error inherent in most practical algorithms. This will be particularly evident later in Chapter 4 in the context of recognizing an object from its optical flow patterns. Because of the confounding of motion, structure, and imaging parameters it is unlikely that the identity of an object can be resolved from measurements taken from a single viewpoint. Even with direct surface measurements provided by a laser rangefinder (Chapter 3), ambiguities still exist as a result of occlusion, projective singularities, and errors in parameterization and modeling. Robustness should increase if decisions could be deferred until a sufficient confidence interval had been established. For these reasons, it becomes apparent that evidence from more than one viewpoint is needed. The question becomes: how do we accumulate evidence from different views, when the evidence is in the form of a conditional probability density function?

In previous stages of this work [**3**], a "winner" was not selected in ambiguous situations, but rather all beliefs above a threshold were labelled as indications of equally likely hypotheses. In order to communicate the validity of all hypotheses above a particular threshold, the beliefs were binarized at this threshold value. By normalizing our confidence values in this manner, combining them from different viewpoints became straightforward: Should the resulting distributions be bi-modal, and the maximum likelihood hypothesis prevail in a largely view-invariant manner, then after a sequence of trials, a robust interpretation could 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. In addition, the confidence in the incorrect models should become insignificant.

Although the voting strategy provides an interesting method for accumulating information from a sequence of views, its weakness lies in the fact that valuable information is lost by not quantifying the *degree* of confidence in each of the hypotheses. In addition, the results from each particular viewpoint were attained independently, discarding valuable knowledge gained prior to each experiment in the calculations. Finally, the previous strategy considered only the likelihoods (unnormalized beliefs) in their tabulation and used the binarization as their method of normalization. The method sought here works directly on the posterior probabilities themselves.

2.1. The Sequential Recognition Framework. Through the process of merging information over a sequence of images, evidence can be accumulated for each model hypothesis, O_i, until a prescribed confidence is attained. In order to attain an inexpensive solution to this problem, a rather strong assumption is made that each data set sampled at a time interval is statistically independent of all other data sets acquired thus far, i.e. knowledge about a data set in one iteration adds no specific information about the raw data acquired in the next iteration. In general, this assumption is considered to be valid and simplifies computations significantly. Let us denote data acquired at time intervals t and $t+1$ as $\mathbf{d}_t$ and $\mathbf{d}_{t+1}$ respectively. The independence assumption can be expressed as follows:

Assumption 1. *Data Independence.*

$$p(\mathbf{d}_{t+1}|\mathbf{d}_t) = p(\mathbf{d}_{t+1}) \quad \textit{iff } \mathbf{d}_{t+1} \textit{ is independent of } \mathbf{d}_t. \tag{2.17}$$

This leads to the main result of the thesis: should the data sets attained at each interval be statistically independent of each other, then information can easily be merged at the level of probabilities by using a recursive Bayesian chaining strategy. This implies that the posterior distribution at time t, $p(O|\mathbf{d}_t)$, can be fed back to the system as the prior at time $t+1$. This can be expressed by re-writing equation (2.13) as:

$$p(O|\mathbf{d}_{t+1}) \propto p(O|\mathbf{d}_t) \int_M p(\mathbf{m}|O)\ p(\mathbf{d}|\mathbf{m})_{t+1}\ d\mathbf{m}. \tag{2.18}$$

This result can be informally derived as follows:

$$p(O|\mathbf{d}_t, \mathbf{d}_{t+1}) = \frac{p(O)}{p(\mathbf{d}_t, \mathbf{d}_{t+1})} \int_M p(\mathbf{m}|O)\ p(\mathbf{d}_t, \mathbf{d}_{t+1}|\mathbf{m})\ d\mathbf{m}. \tag{2.19}$$

Given Assumption 1,

$$\begin{aligned} p(O|\mathbf{d}_t, \mathbf{d}_{t+1}) &= \frac{p(\mathbf{d}_t, \mathbf{d}_{t+1}|O)\ p(O)}{p(\mathbf{d}_t, \mathbf{d}_{t+1})}, \\ &= \frac{p(\mathbf{d}_t|O)\ p(\mathbf{d}_{t+1}|O)\ p(O)}{p(\mathbf{d}_t)\ p(\mathbf{d}_{t+1})}, \\ &= \frac{p(\mathbf{d}_{t+1}|O)}{p(\mathbf{d}_{t+1})}\ p(O|\mathbf{d}_t). \end{aligned} \tag{2.20}$$

(2.21)

By definition, we know that:

$$\frac{p(\mathbf{d}_{t+1}|O)}{p(\mathbf{d}_{t+1})} = \frac{p(O|\mathbf{d}_{t+1})}{p(O)}. \tag{2.22}$$

Substituting into (2.20), gives:

$$p(O|\mathbf{d}_t, \mathbf{d}_{t+1}) = \frac{p(O|\mathbf{d}_t)}{p(O)}\, p(O|\mathbf{d}_{t+1}). \tag{2.23}$$

By Equation (2.15), we know that:

$$p(O|\mathbf{d}_{t+1}) \propto p(O) \int_M p(\mathbf{m}|O)\, p(\mathbf{d}_{t+1}|\mathbf{m})\; d\mathbf{m}. \tag{2.24}$$

This gives us the result we want as:

$$\begin{aligned} p(O|\mathbf{d}_t, \mathbf{d}_{t+1}) &\propto \frac{p(O|\mathbf{d}_t)}{p(O)}\; p(O) \int_M p(\mathbf{m}|O)\, p(\mathbf{d}_{t+1}|\mathbf{m})\; d\mathbf{m}, \\ &\propto p(O|\mathbf{d}_t) \int_M p(\mathbf{m}|O)\, p(\mathbf{d}_{t+1}|\mathbf{m})\; d\mathbf{m}. \end{aligned} \tag{2.25}$$

Equation 2.18 indicates that the sequential recognition strategy is, in fact, a Markov process in the classical sense [**26**], in that computations at each interval depend only on the results from the previous iteration. Formulating the problem as a Markov process stands in contrast to strategies that perform computations, at each iteration, based on the entire data collection gathered thus far. This leads to a more efficient and inexpensive approach. The notation is chosen such that the term $p(O|\mathbf{d}_{t+1})$ actually refers to $p(O|\mathbf{d}_t, \mathbf{d}_{t+1})$. The term $p(O|\mathbf{d}_{t+1})$ is used in order to emphasize that only the data from the current interval, $\mathbf{d}_{t+1}$, is considered directly. The data from the previous intervals are considered implicitly through the information contained in the posterior density function $p(O|\mathbf{d}_t)$. The flowchart of the sequential recognition system can be found in Figure 2.3.

By propagating information in this manner, evidence in the true model should grow over a short number of views, while the belief in the others (as defined by their posterior probabilities) should decline. The strength of the approach is that occasionally ambiguous viewpoints (in terms of several likely hypotheses) are resolved by the strong prior evidence in the correct model collected over the previous views. Furthermore, it need not be the case that the correct model have significantly higher belief than the other models over all views. Should a consistent belief in the correct hypothesis prevail over the viewpoints, then evidence in that model grows, leading to a strong prior for subsequent viewpoints.

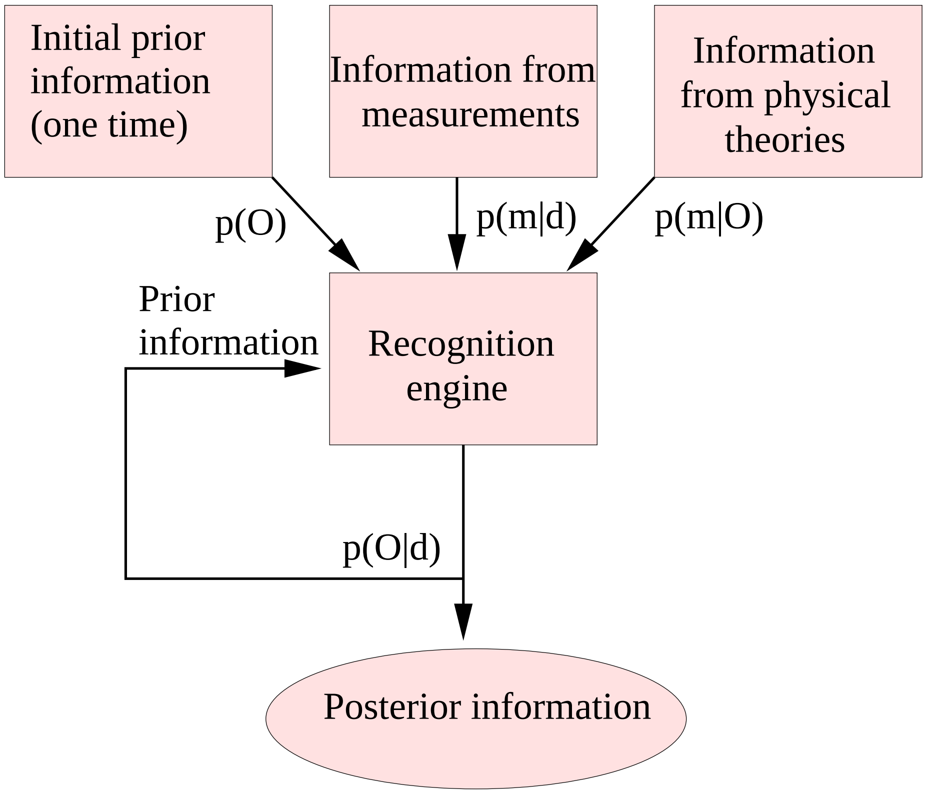


FIGURE 2.3. Flowchart of sequential recognition system.

Cascading information at this high a level permits the system to communicate to higher levels all that it has learned to date without a large computational expense. Furthermore, by quantifying the level of confidence in the various hypotheses at each stage, an active agent would then be able to gather evidence in this fashion until the composite belief associated with a particular hypothesis exceeds a prescribed figure of merit, or until a clear winner emerges.

2.2. Other Work Related to Sequential Recognition. Relatively little work in sequential recognition exists in the computer vision literature. Various strategies for accumulating evidence in various hypotheses will be discussed here. A discussion of the existing literature in active recognition, with particular emphasis on *where* to move the sensor to gather data for recognition, will be addressed in Chapter 5, where it will be more relevant.

Several strategies in combining evidence have opted for the Dempster-Shafer formalism [**29**] over standard Bayesian techniques. These include Hummel and Landy [**47**], Kittler and Hancock [**57**], and Onishi et. al [**65**] to name a few. This technique assigns an upper and lower bound on the possibility of evidence and shows how to combine them. It is designed to eliminate problems encountered by standard probability-based systems. As such, the methodology used to merge evidence does not follow standard statistical rules (i.e. Cox's axioms) and therefore the result is not a probability per say, but a *belief* in different possibilities. Consequently, the result has been criticized for inconsistencies in that changing the order of data events substantially changes the resulting beliefs.

Black et al. [**18, 17, 16**] apply a sequential probabilistic model to the problem of identifying *actions*, i.e. gesture and facial expression recognition. The structure of their solution is different from the one presented in this work, as they apply the CONDENSATION algorithm first introduced in [**51**] to sample, predict and merge information from the posterior probability distribution over model parameters (whose shape is unknown) over time. Here, Bayesian chaining of the posterior probabilities can be directly applied, due to a well-defined posterior distribution.

Herbin [**46**] also presents an iterative strategy for recognition, whereby the posterior information on objects is propagated recursively as the prior for the next iteration. The formulation of the solution is different than the one presented here in several aspects. The formulation requires the computation of conditional state (in terms of observable parameters) transition probabilities: the probability of the next state of the system given the current state. This is attained by observing the frequencies of occurrence of the different states based on different camera trajectories. In addition, at each iteration, the system rejects data based on computation of confidence intervals for the estimation. This causes the rejection of data from viewpoints that do not clearly favor one object over the others. This may lead to biased results as viewpoints that lead to several winners still contain valuable information for the recognition engine, information that could alter the final result.

Several other approaches use a model similar in flavor to that presented in previous iterations of this work [**3**], whereby the frequency of occurrence of local descriptors is used to determine the most probable overall model over time [**78, 84**]. This strategy is quite robust, however it has the disadvantage of ignoring the likelihood of the different possibilities at each iteration. This important information can alter the results substantially (see Section 2 for a discussion.).

In the robotics literature, various solutions exist for merging the acquired data as a mobile robot explores its environment. In [**33, 80**], the authors apply Bayesian updating methods in order to fuse information from different sensors. The authors apply the technique to the problem of tracking.

3. Identifying Informative Views

The byproduct of a probabilistic solution to recognition is that an external agent can asses the quality of the results from a particular viewpoint, and determine whether or not the viewpoint is *informative*. The notion of informativeness is subjective, and often depends on the goal of the task at hand. For the purposes of object identification, it usually refers to the degree to which the data supports discrimination of an object from the others in a database.

There are various reasons for wanting to be able to asses the quality of the information from a particular view. For one thing, it can act as a useful stopping criterion for an active agent gathering data and accumulating evidence in the various object hypotheses. Should the results from a particular view indicate a significant belief of one model over the rest, the agent may choose to cease gathering data (decisions regarding the degree of significance sufficient for recognition is left as an operational issue). In some situations, it might be useful to reject data from a particular viewpoint should it not lead to a clear winner. This subjective choice of rejection criteria is dangerous as omitting data sets in the recognition process could lead to biased results. Finally, it would be useful to store the degree to which each camera viewpoint is informative for recognition. Chapter 5 will illustrate an active recognition strategy that stores this information during a learning phase, and then uses it on-line during recognition in order to guide a sensor towards viewpoints that are considered maximally informative.

There are many strategies for computing the degree of informativeness of data from a particular view. One strategy would be to label a view as informative if there is a significant belief in one object over the others, as defined by a predetermined confidence threshold [**46**]. Of course this strategy is sensitive to the particular choice of threshold. Another measure of recognition ambiguity given a particular data set would be to compute the entropy of the posterior distribution. In fact, this is a natural extension of the derived recognition result, leading to a notion of informativeness described in terms of Shannon's entropy [**26**]:

$$H(P(O|\mathbf{d})) = \sum_i p(O_i|\mathbf{d})\ log\ \frac{1}{p(O_i|\mathbf{d})}, \tag{2.26}$$

A high entropy result implies high ambiguity associated with the data set, and therefore a rather uninformative recognition result. This is precisely the measure taken in Chapter 5.

4. Summary

In this chapter, a general framework for object recognition was presented. The approach is based on a sequential estimation process that accumulates evidence in the various hypotheses over time in an efficient manner, by propagating information on the highest level of inference, the level of the probabilities. The approach uses a Bayesian recognition strategy at each iteration that generates a degree of confidence in the various hypotheses in the form of a conditional probability density function. The strategy is based on an inverse theory that provides a recipe for enumerating the components of the problem, and for combining them in an easy manner. The final result is shown to be useful for determining the degree to which a viewpoint is informative with respect to object identification.

In the next chapters, the generality of the approach will be illustrated through its application to two different real-world recognition problems: Chapter 3 illustrates the problem of recognizing the articulated parts of 3-D models as computed from laser rangefinder data. Chapter 4 focuses on identifying objects moving in complex environments from their optical flow images. Both of these involve various stages of processing, each of which is noisy and uncertain, in order to obtain the required set of parametric descriptors of shape. The versatility of the approach will be showcased by showing that application of the theory to these very different cases is simply a matter of choosing the appropriate probability density functions to describe the relevant information.

CHAPTER 3

Recognition of 3D Models

The first recognition application addressed in this thesis involves identifying objects consisting of multiple parts, where a part is defined as a shape primitive. The focus is recognition of objects whose shapes are complex, where the main difficulty lies in finding adequate model representations. Adding to the challenge is the added possibility of self-occlusion. Two alternatives for representing complex objects have been explored in the literature: The first approach involves finding a single complex model to represent the entire object (for example, see [**54, 91**]). This implies fitting a complex model to an intricate data set, generated from a complicated object shape. Choosing a single model that is appropriate for the task at hand, without compromising the generality of the system, is one of the fundamental tradeoffs in vision. Furthermore, many complex models are not object-centered, thus limiting recognition to viewer-centered matching schemes. Finally, these approaches are very sensitive to partial object occlusion.

The second approach is *recognition-by-parts* [**70, 74**], where an object is represented as a collection of parts, and recognized based on identification of its constituent parts. This approach alleviates the problem of trying to find a single model for complex shapes, and degenerates to the easier problem of recognition of simple individual parts. Furthermore, this approach is not as sensitive to partial object occlusion, in that recognition based on a subset of the total set of parts is generally feasible. Finally, object-centered representations for the individual parts is still possible, rendering the solution robust with respect to changes in viewpoint. For these reasons, a part-based approach to recognition is adopted in this chapter.

The generality of the theory described in Chapter 2 will be illustrated through its application to the recognition of 3D parametric models of object parts. Part models are generated through a real bottom-up vision system developed by several researchers at the Artificial Perception Lab at McGill University. This system, described in Section 1, gathers laser rangefinder data, reconstructs and segments the data into patches, fits 3D parametric models to each of the patches and autonomously decides where to move the sensor next to gather more data. Each stage of this inference chain requires solution to an ill-posed vision problem. This leads to a difficult recognition problem where making assertions from a single viewpoint can lead to ambiguous results. Section 2 will illustrate how the Bayesian formalism introduced in the previous chapter can be used to enumerate all sources of information available and compute the degree of confidence in the various part hypotheses. Section 3 will illustrate how the recognition ambiguities can be successfully resolved by accumulating evidence in the different parts over time.

The versatility of the recognition strategy is further exemplified through its application to other related problems. Section 4 describes extension of the result to the problem of object detection. The goal of this problem is to detect a particular object among a set of objects in a cluttered scene. In a recognition-by-parts strategy, this involves locating the object through identification of its parts. A solution to this problem is shown to be quite simple using the described formulation. In fact, a Bayesian formulation improves upon the standard result by *qualifying* the degree to which the system is confident in the object's parts being among the measured parts in the scene. This is further extended to computing the belief in the object itself being in the scene.

1. Bottom-up System

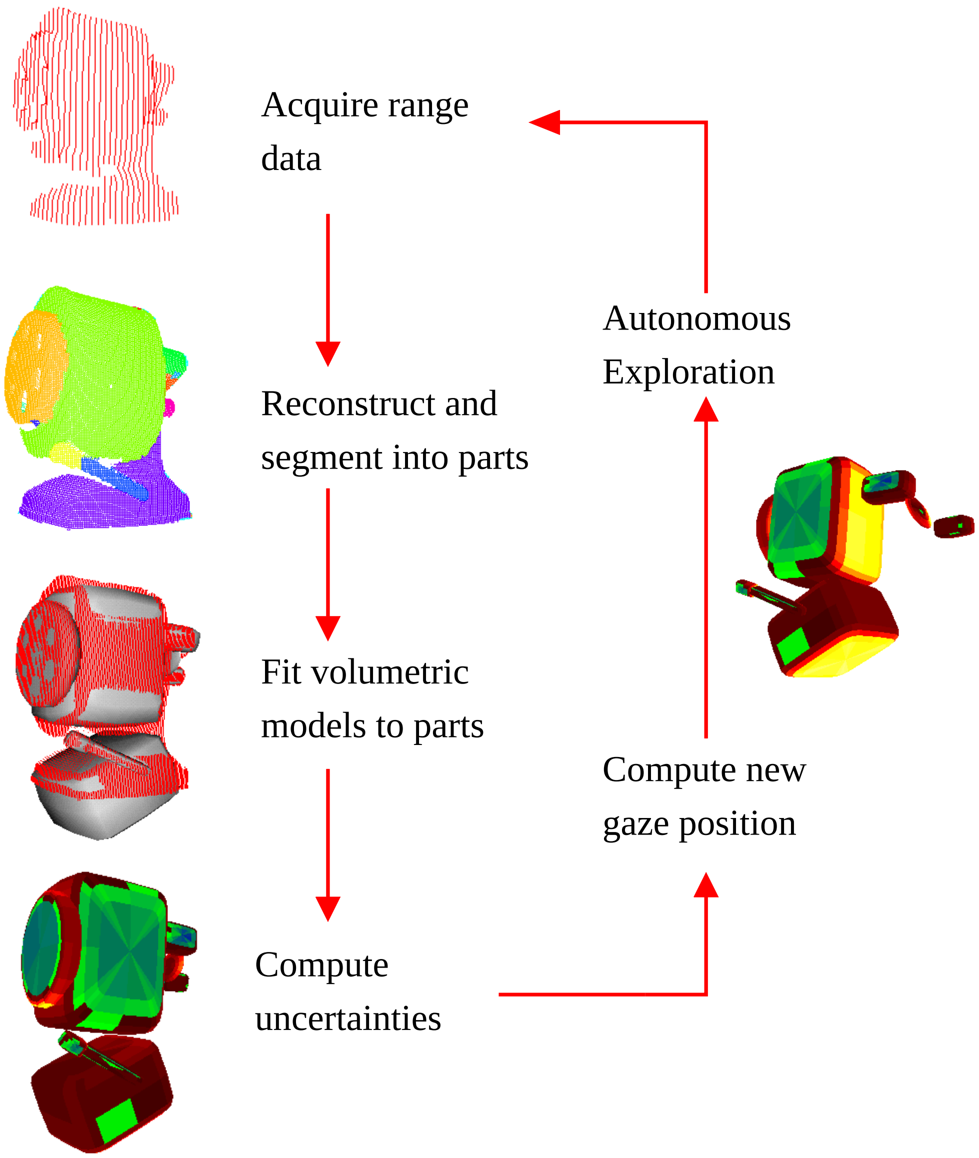
The application of interest pertains to three-dimensional object recognition, where objects are represented by parametric shape descriptors (i.e. models) such as superellipsoids [**12, 13, 39, 79**], deformable solids [**27, 76**], and algebraic surfaces [**91**]. Although the strategy is general, the current context is one in which models are constructed through a process of *autonomous exploration* [**98, 103, 102**] in which a part-oriented, articulated description of an object is inferred through successive probes with a laser range-finding system. In this section, each stage of the bottom-up system will be briefly described.

From any particular sensor viewpoint, surfaces are reconstructed based on the raw data measurements acquired. Inferring the underlying surfaces often requires smoothing the raw data, while maintaining surface structure. This is a difficult problem, one that requires a solution that is robust at many scales. From here, the surfaces are segmented into their constituent parts. The notion of what constitutes a part is not well-defined, and has been largely subjective. In this context, the strategy of Lejeune and Ferrie [**38, 62**] for reconstruction and segmenting the surfaces is adopted. Their notion of object parts consists of convex data patches separated by concave discontinuities. They apply non-linear diffusion to the range data, and generate a hybrid representation at the boundaries: both edge-based and region-based. These two complementary systems are integrated through a relaxation process. The process is generally stable even with the presence of large amounts of noise. However, due to issues of scale and to the fact that the convex patch constraint is somewhat restrictive, errors in segmentation are inevitable. The result is a set of convex parts of an articulated object.

Each convex data patch is then fit to individual parametric models. This leads to compression of the data to a few parameters per patch. Although any parametric primitive would be acceptable, non-deformable superellipsoid models were chosen because of the range of shapes they can represent, as well as their computational simplicity[**39**]. In addition, their parameters are object-centered, therefore robust to changes in viewing position. Associated with each primitive is a covariance matrix $\mathbf{C}$ which embeds the uncertainty of its representation. This covariance matrix can be used to plan subsequent gaze positions where additional data can be acquired to reduce this uncertainty further [**98, 103**]. Data registration is performed according to the strategy in [**88**] prior to the next fitting procedure. The system iterates until the resulting model meets some operationally-defined stopping criterion. A system which automatically builds object models based on this principle is reported in [**62, 102**]. In this chapter, the parameters and covariances of the models are used as features for recognition.

Figure 3.1 illustrates the entire bottom-up system that generates the parametric models used for recognition, as well as the exploration strategy used to gather more data. Each processing stage in the system presents its own challenge, and the effects of compounded noise and uncertainty will be illustrated in Chapter 6.

The benefit of having a stable model building system such as this one is that recognition based on the resulting parameters becomes feasible. In fact, very few, if any such systems exist in the field. Due to a shortage of reliable 3D model-building systems that generate repeatable parameters, very little work is reported on recognition based on the intrinsic



In this flowchart, one can see the entire bottom-up system used to generate part models: (i) The system starts by acquiring laser rangefinder data (seen as dashed red lines). (ii) Surfaces are reconstructed and segmented into part patches (each part is displayed as a different colored patch). (iii) Each patch is then fit to a superellipsoid model. These are shown with the data superimposed on them. (iv) The uncertainties in the parameters are shown superimposed onto the models, and colored such that areas that are highly uncertain are seen in "hot" colors (reds, yellows), and areas that are certain are seen in "cold" colors (blues, greens). (v) The next view is chosen based on the uncertainties of the models.

FIGURE 3.1. Bottom-up system for acquiring 3D model parameters.

parameters of 3D models. The majority of the recognition systems in the literature focus on identification either based on low level or intermediate level features, or even on global features (such as the number of edges or features) either derived from the raw image or from higher level surface or model representatives. Low level features used for recognition might include linear edge fragments, and circular arcs [**41, 42, 43**], line segments, corners, zeros of curvature, or other 2D perceptual structures [**50, 61, 64, 94**]. Intermediate schemes extract features from reconstructed surface patches. These might include surface area, or surface type (cylindrical, spherical or planar) [**40**], surface normals, centroids, direction of axes of surfaces, centers of sphere [**55**], edge adjacency types, i.e. convex, or concave [**35, 36, 37**], or the area and diameter of the surface patches [**52**].

Higher level 3D object recognition schemes often focus on building object models, and extracting externally chosen features from them (see [**31**]). They usually consist of geometrical (low or intermediate) features, or rather unstable global features. In general, extrinsic features are usually much more sensitive to noise, occlusion and viewpoint than the intrinsic ones, such as the parameters of object-centered models. Strategies that concern themselves solely with global descriptors, such as the number of object parts, are generally quite unstable. For this reason, the strategy is often used in conjunction with other types of features, only to help prune the search space.

In more recent literature, a popular strategy entails doing away with model representations altogether, and recognizing based on the appearance of the pixels of the raw greyscale [**71, 95**] or range images [**24**]. More will be said about appearance-based approaches in the next chapter.

It has been stated [**25**] 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. Examples of such features used in recognition schemes are the intrinsic properties of parametric models such as algebraic surfaces [**54**], or superquadrics [**7, 76**]. Here, the intrinsic properties used are the parameters of the models themselves. The strategy described in this chapter is applicable to any such representation. The strategies used for matching based on these parameter sets will be discussed in the next section.

With the problem context defined, the question of how to match objects based on parametric descriptors is raised. The next section will show how applying the Bayesian recognition framework to this recognition context is relatively straight-forward.

2. 3D Part Recognition

The goal of the part recognition system is to compute the likelihood of the unknown part matching each of the K object parts in the database. In this context, the 3D part models, represented by appropriate parametric shape descriptors, can now be used in a part identification system.

The bottom-up system described above is used to gather a sequence of data sets from an object's surface. The entire process, from range measurement to model fitting, is referred to as a *measurement* of an object. In this chapter, sequential recognition is passive and the choice of which navigation strategy to apply, whether it is autonomous exploration or random navigation, is left as an operational issue.

The system works as follows. From a particular viewpoint, range data are gathered and segmented into constituent parts. Let the data space, D, be defined as $\Re^{n_d}$ in this context, where n_d refers the number of measurements for each data set. Let the segmented range data set for an observed part be denoted by $\mathbf{d} \in \Re^{n_d}$, and let S be a discrete random variable measuring the parts in the database which, in this case, takes on a finite set of values, $\{S_i\}|_{i=1...K}$. The goal of the recognition strategy is to compute a discrete (conditional) probability density function $p(S_i|\mathbf{d})|_{i=1...K}$, describing the likelihood of each of the K parts in the database, $\{S_i\}|_{i=1...K}$, given the range data set, $\mathbf{d}$. Solving the recognition problem within this context involves representing each source of information defined by the inverse theory (as described in Section 1.2) by its appropriate probability density functions. Below, each of these are described in terms of the context of 3D parametric recognition.

2.1. Information from Training. The first goal is to build an appropriate distribution to represent what is known about the physical theory that predicts estimates of the observed parameters given an object in the scene. As no such formal theory exists, an estimate is built empirically during a process called *training* or learning. Here, Monte Carlo experiments are run on measures of a known part exactly as in traditional statistical pattern classification methods. In the current context, this implies gathering many measurements (in the defined sense) from different viewpoints about the object of interest. Let the model space, M, be denoted $\Re^{n_m}$, such that $n_m < n_d$. Through a fitting process, each measurement leads to a parametric part descriptor, $\mathbf{m} \in \Re^{n_m}$. The parameters of the N measurements gathered about each part, $\{\mathbf{m}_j\}|_{j=1...N}$, are used to build a sample mean, μ_i, and a sample covariance matrix, $\mathbf{C}_i$, for each part class, $S_i|_{i=1...K}$. This is accomplished using traditional methods:

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} \mathbf{m}_j, \tag{3.1}$$

$$\mathbf{C}_i = \frac{1}{N-1}\sum_{j=1}^{N}(\mathbf{m}_j - \mu_i)(\mathbf{m}_j - \mu_i)^T. \tag{3.2}$$

The most concise and informative way to represent the physical theory information is through a probability density function so that all sources of uncertainty become apparent, and encompassing this information with a Bayesian strategy becomes feasible. A representative distribution for each class is built by making the following assumption:

ASSUMPTION 2. *Multivariate Normal Distribution. Each object in the database can be represented by a single multivariate normal distribution in the n_m-dimensional parameter space.*

This is a common assumption within the Bayesian community, due to its ability to simplify calculations. Its validity is indirectly verified empirically through experimentation with the entire system in Chapter 6. Should the multivariate normal assumption prove invalid, however, the distribution's shape can be modified without modifying the underlying theory. The conditional probability density function representing this information, $p(\mathbf{m}|S_i)$, is calculated for each part in the database. This is formally represented as:

$$p(\mathbf{m}|S_i) = N(\mu_i - \mathbf{m}, \mathbf{C}_i), \quad 1 \ldots K, \tag{3.3}$$

where $N(\mathbf{x}, \mathbf{C})$ denotes the multivariate normal distribution with a covariance matrix, $\mathbf{C}$, such that:

$$N(\mathbf{x}, \mathbf{C}) = \frac{1}{(2\pi)^{\frac{n_m}{2}} |\mathbf{C}|^{\frac{1}{2}}} \exp\left[-\frac{\mathbf{x}'\,\mathbf{C}^{-1}\,\mathbf{x}}{2}\right]. \tag{3.4}$$

The process of fitting the appropriate data set to a multivariate normal distribution is then repeated for each part in the database.

2.2. Information from Measurements. Much of the knowledge we have about a problem comes in the form of experimental measurements. In the current context, a fitting procedure takes an observed range measurement, $\mathbf{d}$, and produces an estimate of the observed model parameters $\mathbf{m}$, and also an estimate of their uncertainty in the covariance operator $\mathbf{C}_d$. The probability density function representing the measurement information, $p(\mathbf{m}|\mathbf{d})$, measures the error in the model fitting process. This information must take into account both the errors in the approximation process as well as in the sensor noise. A direct solution to this problem is difficult to obtain due to the complexity of the physical process as well as a shortage of a statistically sufficient number of samples[**102**]. In fact, the problem is, in general, ill-posed.

In [**102**], the solution was shown to be locally approximated by a Gaussian distribution, whose mean, $\hat{\mathbf{m}}$, is the maximum likelihood estimate attained using an iterative least squares minimization on the range data gathered. In this case, Bayesian techniques were not employed. However, the result can be interpreted as a likelihood distribution, which leads to an estimate of the posterior distribution by assuming a uniform prior on model parameters, $p(\mathbf{m})$. The result is a multivariate normal distribution representing this information, denoted $N(\hat{\mathbf{m}} - \mathbf{m}, \mathbf{C}_d)$. This information is represented by the conditional probability density function $p(\mathbf{m}|\mathbf{d})$, such that:

$$p(\mathbf{m}|\mathbf{d}) = N(\hat{\mathbf{m}} - \mathbf{m}, \mathbf{C}_d). \tag{3.5}$$

2.3. *A Priori* Information on Models. In the current context, there are a discrete number of part hypotheses, $\{S_i\}|_{i=1\ldots K}$. The probability density function used to convey this knowledge is:

$$p(S) = \sum_{i=1}^{K} p(S_i)\ \delta(S, S_i), \tag{3.6}$$

where $p(S_i)$ is the subjective *a priori* probability that the i^{th} part occurs, and

$$\delta(A, B) = \begin{cases} 1 & \text{for A=B,} \\ 0 & \text{otherwise.} \end{cases}$$

2.4. Bayesian Solution to the 3D Part Recognition Problem. With each of these sources of information explicitly expressed with the appropriate probability density function for the current recognition context, computing the Bayesian solution for each conditional probability, $p(S_i|\mathbf{d})$, becomes a straightforward application of Equation (2.13).

$$\begin{aligned} p(S_i|\mathbf{d}) &= p(S_i) \int_M \frac{p(\mathbf{m}|S_i)\ p(\mathbf{m}|\mathbf{d})}{p(\mathbf{m})}\ d\mathbf{m}, \quad i = 1 \ldots K, \\ &= \frac{p(S_i)}{C} \int_M p(\mathbf{m}|S_i)\ p(\mathbf{m}|\mathbf{d})\ d\mathbf{m}, \quad i = 1 \ldots K, \end{aligned} \tag{3.7}$$

as $p(\mathbf{m})$ is the prior distribution on model parameters, previously assumed to be uniformly distributed, and therefore constant, C, over all i.

Substituting the sources of information from Equations (3.3), (3.5), (3.6) into this solution gives:

$$p(S_i|\mathbf{d}) \propto \; p(S_i) \int_M N(\mu_i - \mathbf{m}, \mathbf{C}_i) \; N(\hat{\mathbf{m}} - \mathbf{m}, \mathbf{C}_d) \; d\mathbf{m}, \quad i = 1 \ldots K, \tag{3.8}$$

$$\propto \; p(S_i) \; N(\hat{\mathbf{m}} - \mu_i, \mathbf{C}_{D_i}), \quad i = 1 \ldots K, \tag{3.9}$$

where $\mathbf{C}_{D_i} = \mathbf{C}_d + \mathbf{C}_i$. This result is derived from the fact that the convolution of two normal distributions is a normal distribution, whose mean is the difference of the two means, and whose covariance is the sum of the two covariances. This result was proved in [**1**]. In this case, the convolution of the two normal distribution expressed in (3.3) and (3.5) result in the normal distribution, $N(\hat{\mathbf{m}} - \mu_i, \mathbf{C}_{D_i})$. The term $p(S_i) \; N(\hat{\mathbf{m}} - \mu_i, \mathbf{C}_{D_i})$ will be denoted the *belief* for each part hypothesis S_i.

The constant of proportionality, C, is computed by summing the beliefs in each part:

$$C = \sum_{j=1}^{K} p(S_j) \; N(\hat{\mathbf{m}} - \mu_j, \mathbf{C}_{D_i}). \tag{3.10}$$

This leads to the following discrete conditional probability density function:

$$p(S_i|\mathbf{d}) \propto p(S_i) \; N(\hat{\mathbf{m}} - \mu_i, \mathbf{C}_{D_i}), \quad i = 1 \ldots K, \tag{3.11}$$

Alternately, this can be expressed as:

$$p(S|\mathbf{d}) \propto \sum_{i=1}^{K} p(S_i) \; N(\hat{\mathbf{m}} - \mu_i, \mathbf{C}_{D_i}) \; \delta(S, S_i).$$

Intuitively, this result consists of a series of delta functions, one for each part in the database. where each delta function is weighted by the belief in each part. The final distribution represents the "state of knowledge" of the parameters of each part.

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.

In computing the recognition result in this format, the strategy takes all the available sources of uncertainty into account. This leads to a more accurate assertion of the likelihoods of the various models. Not all matching strategies, however, take the uncertainties into consideration. Pentland and Sclaroff [**76**] introduce a method for the recovery of physically-based deformable superellipsoid models. Recognition is based on the modes of their finite element model. Using their scheme, proximity is measured by evaluating the normalized dot product between the mode values of the model of the unknown object and each of the database 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. Taking the dot product alone (or alternately, taking the Euclidean distance), without consideration of the uncertainties involved, leads to a biased interpretation of proximity. In the case described in this chapter, the uncertainties described by the covariances warp the space such that distances are no longer linear. (Directions with high certainty have stricter distance requirements than those with low certainty.) A simple dot product would not take this warping into account.

A strategy that takes uncertainties into account can be found in [**91**]. Here, recognition based on complex parametric models, i.e. fourth order polynomials [**54**], measures 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 [**91**], Appendix, p.39). Recognition is achieved by choosing the model that results in the smallest Mahalanobis distance.[1] Other methods that include uncertainties in the features can be found in [**49, 60**].

Superellipsoids are powerful models, as they are able to represent a wide variety of shapes in relatively few, object-centered, parameters. In fact, the five parameters that I used for recognition have the advantage of being directly linked to physical object characteristics, such as object size and shape, permitting an intuitive understanding of the successes and failures of both the modeling and recognition modules. Yet, superellipsoid parameters have not been widely used for the purposes of recognition. This is due to the perceived difficulty in overcoming the non-uniqueness of parameters. However, in previous versions of this work [**2**], it was shown that there are a finite number of degeneracies in their representation, and that these can be dealt with relatively easily. For now, these models are primarily used in modeling (as in CAD design), and in recognition of simple objects [**20**]. In [**79**], the authors fit objects to superquadric models as well. However, the goal of their strategy is not to recognize the object, but rather to find an appropriate representation for it. The range of possibilities is limited to one of twelve predetermined categories of 3D shapes (*geons*).

[1] Differences between the strategy proposed in this thesis and their approach were discussed in the previous chapter.

Classification is based on low level features derived from the superquadric model, such as bent or straight axis, and straight or curved edges.

3. Accumulating Evidence Over Time

The recognition problem described above is difficult due to the noise and uncertainties associated with each stage in the bottom-up system. As a result, recognition from individual viewpoints can lead to ambiguous results in terms of similar levels of belief in more than one model. For this reason, a dynamic approach to recognition is taken whereby data are gathered from several viewpoints and evidence, in the form of the probability distributions, is accumulated using the strategy outlined in Section 2. Once again, the assumption of independent data sets is made here. Note that navigation through the scene does not alter the properties of the data set, such as their independence.

Following the outlined strategy leads to the representation of the posterior probabilities from the previous time interval t as the priors for the current interval. Application of this strategy to the recognition problem leads to following updating functions for $p(S|\mathbf{d}_{t+1})$:

$$p(S|\mathbf{d}_{t+1}) \propto \sum_{i}^{K} p(S_i|\mathbf{d}_t)\ N(\hat{\mathbf{m}}_{t+1} - \mu_i, \mathbf{C}_{D_i})\ \delta(S, S_i). \tag{3.12}$$

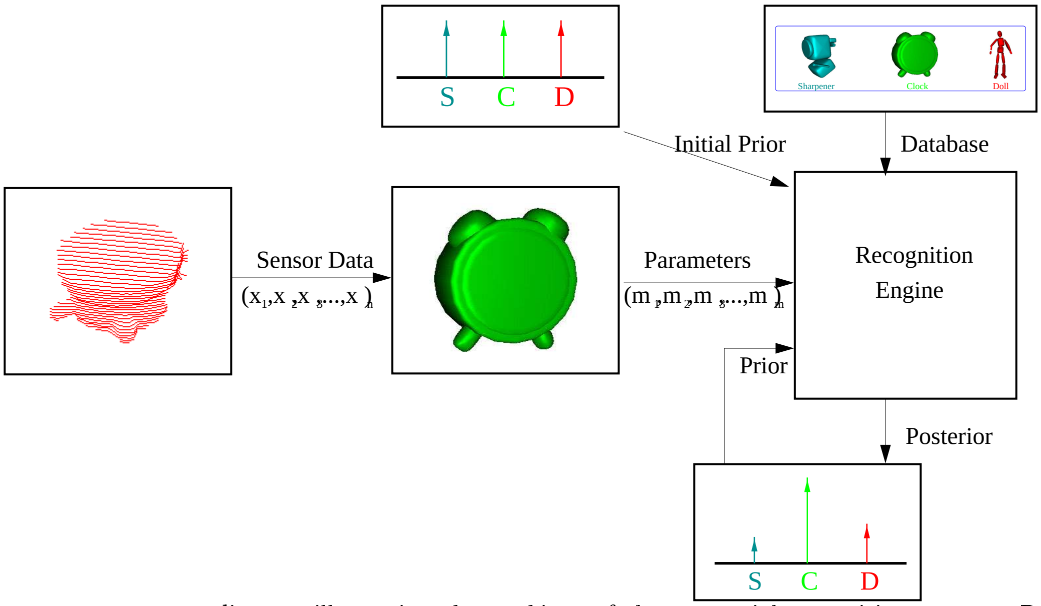
The idea is that ambiguities should be resolved as the data sets are gathered, leading to convergence to the correct hypothesis in a short number of iterations.

A flowchart illustrating the various stages in the sequential recognition process [2] in the context of 3D model recognition can be seen in Figure 3.2. It shows how data is represented by parametric models prior to being sent to the recognition engine. This information is combined with prior and database information (i.e. information from physical theories derived during training) to obtain the recognition solution, represented here by a discrete probability density function. This posterior density function is then fed in as a prior for the next iteration.

4. Extension to Multi-part Object Detection

With the part recognition strategy in place, extensions of the work can become simple matters of applying standard probability rules to the resulting probability density functions. In this section, the difficult problem of identifying multi-part articulated objects is

[2] The application of this Bayesian chaining strategy for the purposes of accumulating evidence can be referred to as *probabilistic feedback*, as the posterior is fed back as the prior for the next iteration.



Above one can see a diagram illustrating the workings of the sequential recognition system. Raw data measurements are parametrized prior to entering the recognition computations. This is combined with prior information and information from the database to obtain the posterior density function, represented as a discrete distribution showing the probabilities assigned to the different objects in the database. This becomes the prior for the next iteration. Notice that the belief in the true model, the clock ("C") is growing over time.

FIGURE 3.2. Sequential recognition system.

addressed. The task of interest is to detect the presence of an object, or a part of an object, in an unknown scene. In a recognition-by-parts strategy, recognition of the object as a whole becomes the task of merging the information about the identities of its constituent parts.

4.1. Part Detection. Consider the case of a collection of object parts presented to the recognition module, where the parts result from the segmentation of one or more object data sets. The question that is asked is: *Can we detect the presence of a particular part in the scene?*

One motivating factor in addressing this issue is due to its interest to the community working on techniques for indexing through large databases of images. Problems in *Content-Based Image Retrieval*, for example, involve the following: Given a particular image of an object, find all images containing it. The context is quite different from the one considered thus far, where recognition implied a closed-world assumption: each of the parts in the scene have been seen *a priori* during training. In most detection problems, however, this assumption is violated. The interesting question is whether the strategy outlined in this thesis can be extended to address problems such as these.

Let O denote an object in a database consisting of the collection of parts $\{S_i\}$. The scene consists of a set of N measured parts resulting from the data sets $\{\mathbf{d}\}$. Previously, the task was to recognize each part in the scene as being one of a set of parts in a database. Here, the task of the system is reversed. The goal is to decide upon the likelihood of a particular database part, S, being among the measured scene. This is denoted by the conditional probability $p(S|\{\mathbf{d}\})$.

The solution to this problem is to perform recognition for each of the data sets in the scene, maintaining the closed-world assumption. In this fashion, each of the data sets in the scene results in a posterior probability, $p(S|\mathbf{d}_i)|_{i=1 \ldots N}$. The likelihood of a particular database part, S, being among the measured scene can be defined as follows:

$$p(S|\{\mathbf{d}\}) = 1 - \prod_{i=1}^{N}(1 - p(S|\mathbf{d}_i)). \tag{3.13}$$

Intuitively, this implies that the probability that the part is among the measured parts is equal to the negation of the probability that it is not among any of the parts. Due to the probabilistic nature of the part recognition solution, extension to this problem is not difficult. Implicitly, it permits the possibility of other unidentified parts to be present in the scene.

4.2. Object Detection. With this framework in place, the extension to detection of multi-part objects is not difficult. Let R be the number of parts for a particular object, O. The collection of parts for O is $\{S_i\}|_{i=1 \ldots R}$. The posterior probability that O is within the scene, given a collection of part measurements, is denoted $p(O|\{\mathbf{d}\})$. It can be defined as:

$$p(O|\{\mathbf{d}\}) = 1 - \prod_{i=1}^{R}(1 - p(S_i|\{\mathbf{d}\})). \tag{3.14}$$

This implies that the probability that O is in the scene is the negation of the probability that none of its parts are in the scene. Both of these examples indicate that the probabilistic framework permits extensions to a wide variety of difficult vision problems.

5. Summary

In this chapter, the Bayesian solution was applied to the problem of 3D recognition-by-parts. This involved specifying the form of the probability density functions representing the sources of information enumerated in the previous chapter, and substituting them into the final solution. The particular context discussed in this chapter was that of recognition of 3D part models generated from a bottom-up system that performed various stages of processing: from laser rangefinder data to segmentation to model fitting. The benefits of having this stable system available is that recognition based on the resulting model becomes feasible. However, this problem is inherently difficult, due to the compounding uncertainties at each stage of processing, propagated throughout the system. Using a sequential estimation process should overcome the recognition ambiguities, and generate a clear winner in a short number of views. Empirical results will indeed illustrate that this is true (see Chapter 6).

This brings up the point that an interesting extension of the work would be to compute the uncertainties in the inverse solutions at each stage of processing. For example, it would be worthwhile to compute a density function describing the likelihood of the results of segmentation. A Bayesian formulation for propagating such information through the system, such as the one developed here, could easily be formulated for the entire bottom-up system. Such is the versatility of the Bayesian approach. Further, were this information available to the recognition module (where it would easily be incorporated into the final solution), it would certainly improve the final result. Currently, erroneous recognition results cannot be attributed to ambiguities at a particular stage of processing. Should uncertainties be incorporated in each stage, an external agent could decide to stop the process when insufficient information is available for a particular task. In addition, high level feedback could be provided to the system in order to aid lower level tasks.

This recognition context presented in this chapter is well suited to industrial applications whereby objects are placed on a platform, and an robot is free to move around it, examining it for identification purposes. The scene is not complex, and the viewing conditions are fairly controlled. The next chapter focuses on a difficult recognition task, whereby model fitting is not possible, the data is noisier, the scene more complex. The versatility of the Bayesian formulation presented in the previous chapter will be tested through its application to this difficult problem.

CHAPTER 4

Recognizing Objects Based on Optical Flow Images

The problem addressed in this chapter is structured as follows: consider the case of a user (human or machine) waving an object in front of a black and white camera connected to a computer system. The system's task is to identify the moving object from a set of objects it had seen at an earlier stage. The difficulty of the task lies in the fact that the lighting conditions and background scene may have changed significantly since the learning stage. This difficulty is partially alleviated by focusing the problem as follows: As the object is moved in front of the camera, it invokes a motion pattern on the retina. The specific task of the system is to recognize the object from the signature encoded by this pattern. The pattern is commonly referred to as the *optical flow*.

The scope of this problem is more general than in the previous case, and has many more difficulties associated with it. The scene is more complex, data noisier, the requirements more general, and the relationship between data and model less obvious. In this case, no specific model (such as a superellipsoid) is assumed. Alternatively, object representations are *viewer-centered*, and the data/model relationship is made explicit though empirical evidence gathered during a training process. Specifically, an appearance-based object recognition strategy is devised that builds on the works of Nayar et al. [**71**] and Turk et al. [**95**], due to its fast indexing time and its ability to deal with complex scenes where model-building becomes infeasible. These methods build a lower dimensional subspace based on the entire set of raw images acquired, using Principal Components Analysis (PCA) [**95**]. The lower dimensional space in which the images are represented is referred to as an *appearance manifold*[1]. Recognition (or rather, indexing) is based on projecting images acquired on-line onto the manifold and finding the closest stored image.

The majority of the existing appearance-based strategies focus on applications such as face recognition [**53, 56, 11, 87, 95**], lip-reading [**63**], and others [**18, 71, 75**]. For most of these applications, using only the raw grey-scale images as input to an appearance strategy has worked quite well. The major drawback is that tight control over the image formation parameters, such as lighting and background, has to be enforced in order to ensure repeatable appearance. One way to avoid some of these difficulties is to make direct surface measurements, i.e. range measurements, of the object to be recognized. This is precisely the strategy taken by Campbell and Flynn [**24**].

In this thesis, difficulties regarding sensitivity to lighting and background are overcome through the development of an appearance-based object recognition strategy, whereby differential image properties are exploited. Here, inputs for recognition correspond to the optical flow images induced onto the image plane by the relative motion between camera and object. Optical flow images used in this appearance-based strategy will be referred to as *appearance flows*. Using flow images as the input offers some advantages over using the original grey-scale images. First, the resulting flow field is somewhat invariant to scene illumination (assuming that it is constant during acquisition). This lends itself to a more robust strategy whereby lighting conditions need not be identical to those used during training. Second, using flow provides a means of figure/ground separation in the case of a stationary observer. This is a major advantage in that appearance-based strategies that use raw grey scale images have to ensure the exact same background during recognition and training. Here, arbitrary stationary backgrounds are possible as the algorithm can focus on the moving object in the scene. Furthermore, once the moving object has been detected, the object can be centered within the image. This type of position normalization is not feasible using traditional inputs. Third, the flow field itself will be shown to provide coarse information regarding the object's 3D structure. A final motivation for the approach has its roots in biology. A great deal of "wetware" in the mammalian brain is devoted to the computation of motion. This motivates the exploitation of its capabilities at solving many visual tasks.

Several strategies exist in the literature for recognition based on optical flow images. Primarily, these use the flow images for recognizing *actions* [**18, 19, 28, 66**]. These actions

[1] A manifold is an n-dimensional surface[**73**].

can include facial expressions, gestures, etc. The work presented here is distinguished from these strategies in that the focus is on recognizing the *object* itself, based on the signatures obtained from the flow field. After careful examination of the existing body of literature, it seems that this is the first work that uses flow for the purposes of recognizing objects.

Despite the advantages, there are difficulties associated with recognition based on motion images. It is well understood that the shape of the object, the relative motion between camera and object, as well as camera geometry are confounded in the resulting flow pattern. The problem of "factoring out" the different components can be exceedingly difficult (if not impossible). However, my contention is that it can be dealt with where suitable *a priori* constraints are available [**10, 18, 96**]. I attempt to get around the difficulty of "factoring out" the structural information by imposing constraints (a) on the camera geometry, by making the assumption that the image projection is scaled orthographically, and (b) on the permissible relative motions. Although this imposes some restrictions on the data acquisition stage, the resulting flow images (i.e. noisy and relatively low resolution signatures of object shape) are perfect candidates for testing the limits of the sequential recognition strategy. The challenge of the theory is to converge to a correct solution, even with such problematic input data.

The fundamental question of this chapter is now stated explicitly: Can appearance flows possibly be used for recognition purposes? The answer rests on the following two hypotheses:

HYPOTHESIS 1. *Appearance Flow Manifold.* *The manifold of appearance flows generated by training on the expected motions of an object can be used to distinguish between different objects.*

HYPOTHESIS 2. *Accumulation of Evidence.* *Confounding information in the optical flow signatures can be resolved by accumulating support for different object hypotheses over a sequence of observations using the theory of Chapter 2.*

This chapter describes how both of these hypotheses are used in conjunction with the theory in order to solve the problem of recognition based on flow images. Section 1 begins with a description of the algorithm that generates the flow images. Section 2 describes the generation of an appearance flow manifold during training, that will be used as a basis for on-line recognition experiments. The application of the Bayesian sequential recognition strategy described in Chapter 2 to the current context is described in Sections 3 and 4.

It is important to emphasize that, in this chapter, flow images are used for recognition as they provide a good illustration of one application of the theory. The implication is not, however, that this strategy can necessarily solve the general recognition problem. For instance, flow can be successfully used to identify objects based on 3D structure. However, should the system need to discriminate objects based on greyscale texture patterns on its surface alone, then flow images are certainly not the best features for the task. To solve a more general recognition task, it would be instructive to use flow images in conjunction with other features extracted from the image. Examples of such features are: color, texture, etc. In this chapter, the goal is to illustrate the application of the theory to one type of recognition task. As the details of the system are introduced, it should become evident that the addition of other features sets is fairly easy. Extension to include other feature sets is therefore left as an operational task.

1. Generating Optical Flow Images

As the object moves with respect to the camera, optical flow images are computed from the sequence of images gathered. For the purposes of this thesis, any stable optical flow algorithm can be used as an input to the system. Deciding which one is most effective is largely an operational issue. It was found that the system developed by Benoit et al. in [**14**] generated stable optical flow images that were sufficient for these purposes. Their system is completely bottom-up, using pixel and region matching techniques coordinated during a two-phase gradient ascent algorithm: pixel matching error measures are locally minimized and flow field consistency constraints are applied to the low confidence neighbors. Convergence is usually attained after a single iteration for an image frame pair. Predictions for future flow fields are computed using temporal integration and Kalman filtering. The system is designed to be flexible: large displacements are tracked as easily as sub-pixel displacements while higher-level information feeds flow field predictions into the measurement process. An example of a flow field resulting from this algorithm can be found in Figure 4.1.

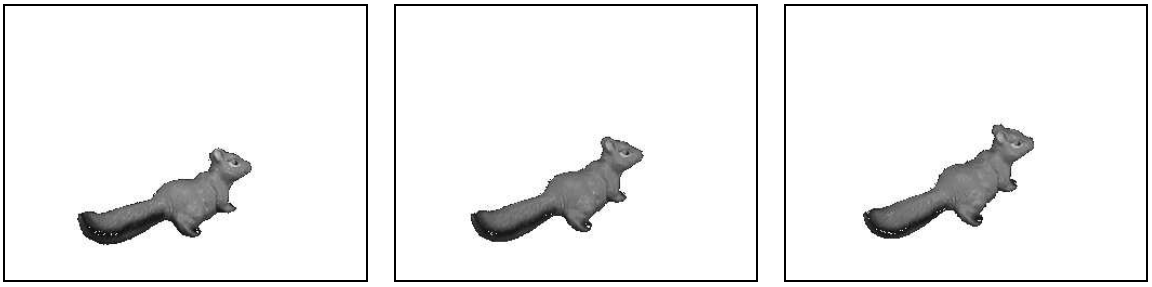
The prediction techniques are used to maintain a tracking window which follows the moving object through the scene. Once locked onto this moving target, the tracking window focuses entirely on the object of interest, computing the optical flow only within the window of interest. This implies two levels of figure/ground separation, virtually eliminating the effects of varying backgrounds: (a) removal of the static background image, and (b) removal of other moving distractors in other parts of the scene. The resulting image is padded with zeros everywhere outside the tracking window. Finally, the object is centered within the resulting image. This offers an advantage over approaches which: (a) require constraints to be enforced on the position of the object within the image or (b) require training on all possible object positions prior to recognition.

The resulting flow image contains information regarding the direction of motion as well as the magnitude of flow at each pixel location. Approaches that are interested in action recognition use the direction of motion as features. In this thesis, the interest is in recognizing objects based on 3D structure, therefore only the magnitude of the flow images is used. For a particular set of motions, the magnitude images can be shown to provide a rough kinetic depth map, giving some insight into the object's 3D shape. More will be said about this in the following section.

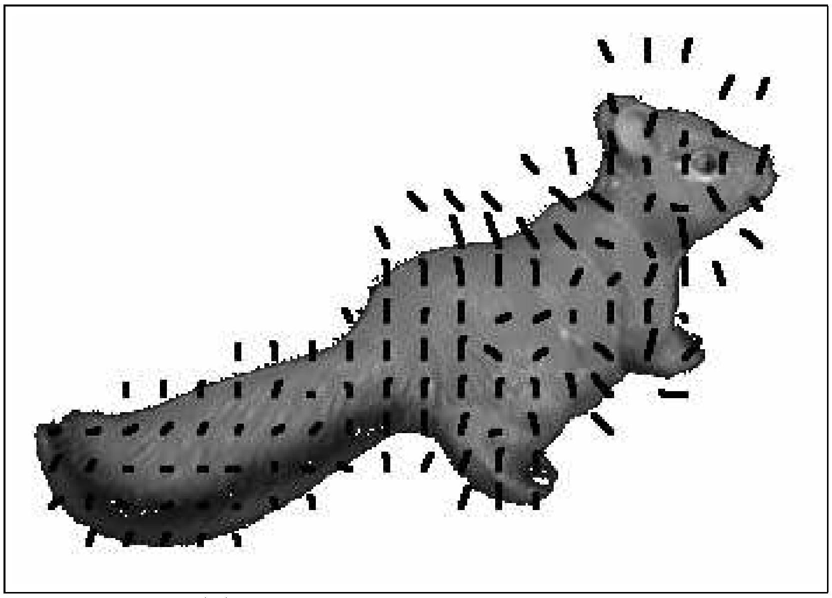
In the current context, the distance of the object from the camera is fixed to ensure the possibility of differentiation between objects with similar shape, but with varying size. Although no spatial scaling is performed, the system removes the effects of varying temporal scales by normalizing the magnitudes of the optical flow. This ensures that speed does not effect the appearance of the 3D structure, as well as adding more resolution to the values. An example illustrating the stages of processing for the flow input can be found in Figure 4.1. After processing, the flow images can be expressed as vectors in a high dimensional space. For the purposes of efficiency, a lower dimensional parametrization of the images is desired so as to concisely represent each input by a small set of features. The next section will indicate how to build the low dimensional basis for recognition, and how to represent the flow images in such a space, so as to best predict their appearance on-line. The off-line representation of the flow images is referred to as an *appearance flow manifold*.

2. Building an Appearance Flow Manifold

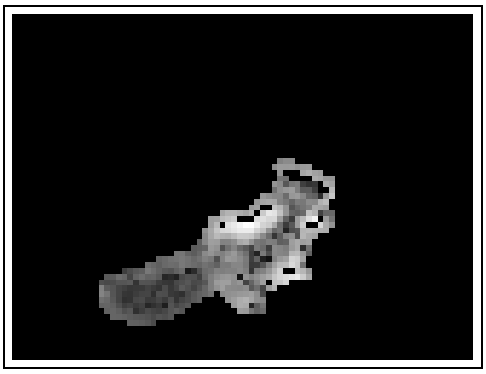
The goal of the system developed in this chapter is to be able to recognize objects from a wide variety of movements. In satisfying this requirement, the approach taken is to generate an appearance manifold for motion, that represents the set of possible motions for the object in question, and permits recognition from an even more general set. The motivation for this novel strategy originates from research in aspect graphs [**21, 34, 58, 81, 89**], where the common approach is to store a set of an object's characteristic views as a basis for recognizing it from arbitrary positions. The desire is to extend this notion and pose the



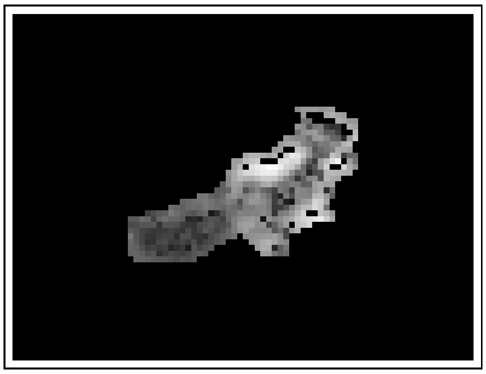
(a) Sequence of Three Images.



(b) Resulting Optical Flow Image.



(c) Magnitude Image.



(d) Centered Magnitude Image.

Above, one can see the stages of processing for a flow image. In (a), three greyscale images acquired in sequence are shown. In (b), one can see the computed flow vectors projected onto the object. In (c), the magnitude image is shown, shaded in accordance with the distance from the camera: from closest (in white) to furthest (in black). Notice that the background can easily be removed. In (d), the object is centered within the image.

FIGURE 4.1. Stages of processing of the flow image.

question: Is there a set of *characteristic motions*, analogous to characteristic views, which are sufficient to ensure successful recognition results from a wide set of motions?[2]

In order to determine the sufficient set of motions required for building the appearance flow manifold, the constraints imposed on the expected motions of a mobile observer in a stationary environment (and vice-versa) are described. These will subsequently be used to generate the appearance flow manifold for each object in a database.

ASSUMPTION 3. *Camera Constraints. Camera to object distances are bounded and scaled orthographic projection is assumed.*

ASSUMPTION 4. *Motion Constraints. The same motion model can be used to account for an object moving about a fixed observer provided that rotations are limited to axes that are approximately parallel to the image plane.*

ASSUMPTION 5. *Motion Decomposition. The trajectory of an observer moving through a stationary environment can be decomposed into a sequence of short, curvilinear segments. This motion model can be guaranteed in the case of an active vision system or mobile robot equipped with a suitable tracking system.*

Assumption 3 ensures that the object is placed within reasonable distance from the camera. As well, the assumption of orthographic projection (as opposed to perspective projection) indicates that the object is far enough from the camera so that within-object warping is not a factor. This assumption is central to the optical flow algorithm chosen, and facilitates the computation. A *scaled* orthographic projection encompasses information with respect to the distance from the camera. Assumption 4 restricts motions to curvilinear motions that are subsets of rotations around axes that are parallel to the camera plane. This is because application of the orthographic projection assumption leads to the fact that translations and rotations in the camera plane provide little information about the 3D structure of the object. The combination of the first two constraints, imposed in conjunction with the optical flow algorithm described in the previous section, results in the magnitude image representing a rough *kinetic depth map*, whereby larger motion values indicate surfaces that are closer to the camera. This gives us the structural information needed for recognition. The claim is that the above assumptions are not overly restrictive and can account for a reasonably wide range of viewing situations.

[2]Portions of this work will appear in [**6**].

Assumption 5 is central to the main hypothesis regarding building an appearance manifold for motion. The implication is that should a motion trajectory be simply a collection of short motion sequences, then training based on a small set of such motions can lead to the recognition of a wide range of such motion trajectories. The requirement is to find an interesting set of such characteristic motions so as to be able to identify the objects based on their combinations. The restriction to curvilinear motions originates from the previous two assumptions, and exists in order to ensure motion information that provides insight into the 3D structure of the object rather than about the its trajectory.

This brings us to the central hypothesis regarding building an appearance flow manifold:

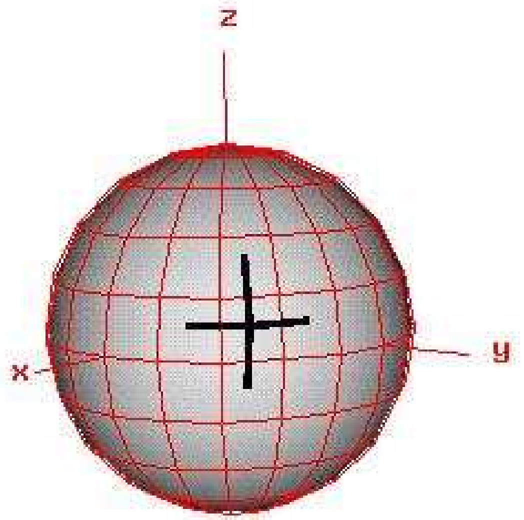
HYPOTHESIS 3. *Appearance Manifold for Motion.* *An appearance manifold for motion can be generated by the collection of a set of local motion descriptors taken uniformly about the object of interest.*

Following the definition in [**73**, Equation 8.4], a manifold M is a set that is covered by the images of a collection of one-to-one functions (and the composite of function pairs are differentiable). The hypothesis is that an appearance manifold for motion can be built from a collection of characteristic motions gathered at locations about the object of interest during training, using standard appearance-based techniques. The implication is that the resulting manifold can be useful for on-line recognition experiments based on a more general set of motions. The advantage of this approach is that it avoids the problem of having to predict and learn all possible on-line motions. Instead, it provides the flexibility that trajectories permissible for recognition can be comprised of combinations of the learned set. This hypothesis will be verified empirically in later chapters.

As a practical matter, the object of interest is placed at the origin of the viewsphere. The viewsphere is evenly tessellated, such that each region, associated with a particular camera location, assumes approximately the same area on the sphere. There are various strategies for attaining a uniform tessellation of a sphere. Deciding which to use is largely an operational issue[3]. In each segment, a local motion basis is computed by sweeping the camera along a set of short curvilinear arcs resulting in a motion sequence. These motions reflect segments of the expected on-line trajectories. An example of one type of tessellated viewsphere can be seen in Figure 4.2. Here, one can also see the arc-like motion

[3]One method to tessellate the viewsphere will be discussed in Chapter 6.

segments that form a basis for the training manifold, i.e. the elements in the manifold can be expressed as a combination of these segments.



Here one can see the tessellated viewsphere surrounding the object of interest. The perpendicular darker lines illustrate the arc-like trajectories comprising the motion basis for training.

FIGURE 4.2. Motion basis along viewsphere.

With the strategy for building the appearance manifold in place, an interesting issue arises: What should the density of sampling be in order to sufficiently characterize the motion manifold? Sampling the viewsphere, in this context, refers to both position and motion space sampling, i.e. where and how to gather motion images. Surely, the strategy has to be frugal in the number of images it gathers, even if sampling is off-line, as space and processing time are limited. Deciding, *a priori* which viewpoints and motion images are the most important for recognition is a difficult and open ended research issue. A strategy for empirically determining how densely to sample the viewsphere will be described in Chapter 6.

The specifics of the off-line generation of the appearance flow manifold are now described. As the object moves with respect to the camera, a sequence of grey-scale images is gathered. Three images in sequence are used to compute the optical flow images as described in Section 1. Let this data vector, i.e. a single flow image, be defined by the vector $\mathbf{d} \in \Re^{n_d}$. The dimension, n_d, can be very large for high resolution images. For instance, for a 640×480 image, the vector representation has dimension 307200. Representing each flow image in this high dimensional space is impractical, and unnecessary as many dimensions are clearly redundant for these purposes. Therefore, a lower dimensional parameterization in which to represent the flow images is sought.

There are many strategies for choosing an appropriate parameterization strategy. An ideal solution would be to use an appropriate object-centered model that clearly expresses a physical relationship between data and model parameters (should it be available as in the previous case). However, an appearance-based strategy was chosen due to the difficulty in finding physical representations in dynamic real world environments. There are many viewer-centered approaches available in the literature. Most focus on the problem of storing representative or characteristic views of the objects in the database, and building some object representation based on the appearance of these views that would permit its recognition based on instances of (or slight variations of) one of the stored views[**9, 25**]. Most strategies use the raw grey-scale images as inputs to the system, leading to a strong dependence on the the pixel configurations being identical to those trained on. Variations due to clutter and noise, as well as background and lighting variations lead to parameter representations that vary widely from those trained on. Using differential properties of the image, such as flow, helps to solve some of these problems.

The appearance-based strategy employed here is one that has become quite popular in recent years, Principal Components Analysis (PCA) [**15, 95**], a methodology derived from the Karhunen Loeve (K-L) transform. Its appeal stems from its speed and ease of use. The strategy takes data of high dimension, n_d, and builds a lower dimensional basis, referred to as an *eigenbasis*. The compression is based on correlations of the input data. In this case, the entire set of p optical flow magnitude images gathered during training is used to determine a lower dimensional basis that spans the representation space, $\Re^{n_m}$ where $n_m < n_d$,. An overview of the standard techniques involved in PCA can be found in Appendix A.

A representation for each object is then constructed by projecting its N corresponding flow image vectors $\{\mathbf{d}_j\}|_{j=1...N}$ onto this lower dimensional basis (see Figure 4.3). Let the projected flow vectors for a particular object in the training set be denoted $\{\mathbf{m}_j\}|_{j=1...N}$, where $\mathbf{m}_j = \mathbf{m}_j(\mathbf{d}_j) \in \Re^{n_m}$ refers to the projection.

Once the training images are represented as lower dimensional vectors, a concise representation for each object in the database is sought. In accordance with the inverse theory, this involves determining a physical theory, should one exist, that predicts estimates of the appearance flow parameters (i.e. positions in eigenspace) given an object in the scene. As

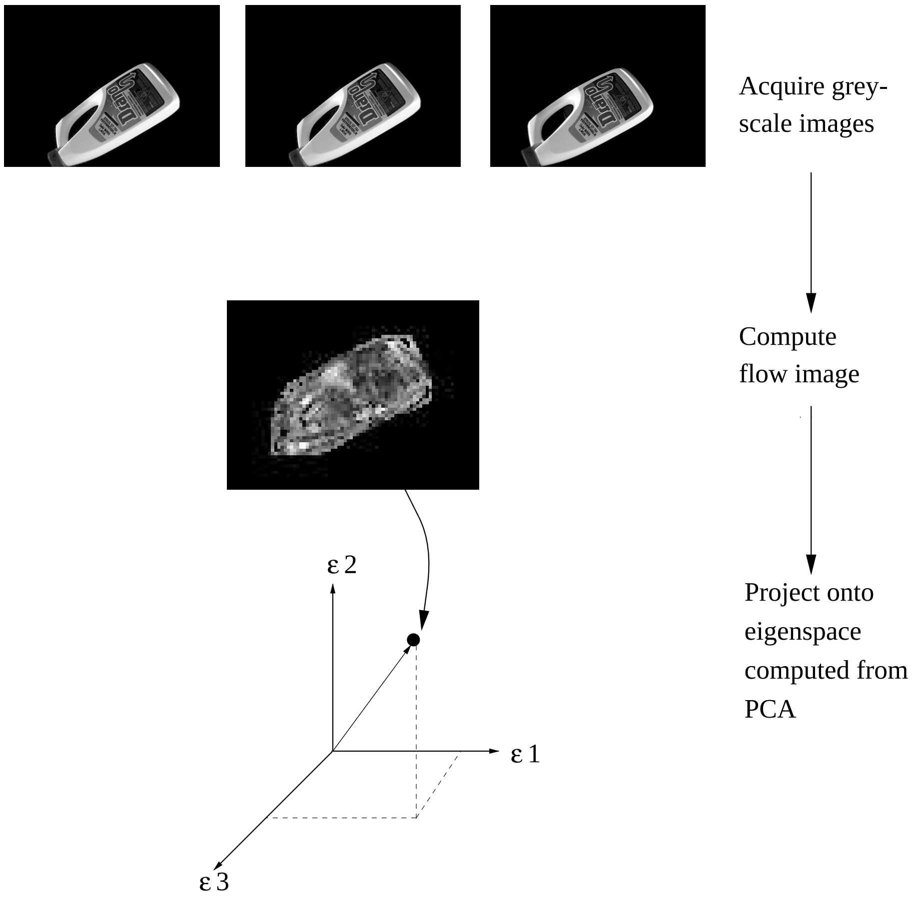


FIGURE 4.3. Bottom-up system for acquiring parametric flow descriptors.

no such formal theory is readily available here, one is built empirically through learning. This information can be concisely represented through a probability density function. In order to estimate the distribution, the following assumption is made once again:

ASSUMPTION 6. *Multivariate Normal Distribution. Each object in the database can be represented by a multivariate normal distribution in the lower dimension eigenspace.*

This is a common assumption, one that was also made in the previous chapter. The validity of the assumption can be verified empirically. Should it prove invalid, the normal distribution can easily be replaced with a different distribution without modification of the theory.

A multivariate normal distribution $N(\mu_i - \mathbf{m}, \mathbf{C}_i)$, is then estimated for each object, O_i, using the standard techniques described in Section 2. μ_i and $\mathbf{C}_i$ are the sample mean and sample covariances generated by the N projected flow vectors, $\{\mathbf{m}_j\}|_{j=1\ldots N}$. This information is represented by the conditional probability density function, $p(\mathbf{m}|O_i)$, which represents the physical theory predicting possible variations in parameters given each object hypothesis, O_i.

Having built the appearance flow manifold in statistical terms provides the means for the Bayesian recognition strategy outlined in the following section.

3. Recognition Based on Flow Images

This section describes a Bayesian recognition strategy that applies the techniques developed in Chapter 2 to the problem of recognition based on flow images. The goal is to represent the posterior beliefs over the entire set of K object hypotheses, $\{O_i\}|_{i=1...K}$, given flow image $\mathbf{d}$, by a discrete (conditional) probability density function $p(O_i|\mathbf{d})|_{i=1...K}$.

On-line, a sequence of flow images is gathered by either moving the camera with respect to the object (or by moving the object in front of the camera) via an arbitrary subset of the permissible set of movements. These include the camera moving about the object of interest through motions that consist of combinations of arc-like trajectories about the object's viewsphere. Alternatively, the object is moved about the camera through rotations (or partial rotations) along axes that are parallel to the camera plane in order to test the system's ability to recognize based on motions that differ from those trained on. Chapter 6 will provide an examination of the system's generalizability beyond the motions trained on.

Recall that $\mathbf{d}$ corresponds to a vector representation of a single optical flow image in the sequence. The first step in the recognition process is to compute the support for each object hypothesis O_i, given the single optical flow image. In what follows, each of the sources of information available, described in Section 1.2, is detailed in terms of the current problem leading to a working expression for $p(O_i|\mathbf{d})|_{i=1...K}$. This provides a formal structure for solving inverse problems such as this one.

3.1. Information from Training. Recall that each object in the database is represented by a multivariate normal distribution obtained through training:

$$p(\mathbf{m}|O_i) = N(\mu_i - \mathbf{m}, \mathbf{C}_i). \tag{4.1}$$

μ_i and C_{T_i} are the sample mean and covariance matrix of the normal distribution obtained through training using the methods described in the previous section.

3.2. Information from Measurements. As the object moves with respect to the camera, flow images are computed from the grey-scale images acquired. The resultant vector representation of the optical flow image, $\mathbf{d}$, is of high dimension. A more concise description is sought, so as to render computation easier and more efficient. Similar to the case of 3D recognition, an estimate of the model parameters $\mathbf{m}$, given the data $\mathbf{d}$ is desired. This case differs substantially from the case of recognizing 3D models in that, previously, this process was an ill-posed inverse problem, requiring an additional inference process prior to that of recognition. In this case, the model parameters, $\mathbf{m}$, are obtained by projecting the image onto the precomputed eigenspace. This process is not the solution to an inverse problem, but rather to a forward problem: given some observed data, determine its representation in a lower dimension space, through a well-known projection map. Each data set projects onto a single point, therefore the process is well-posed.

The theory calls for a description of $p(\mathbf{m}|\mathbf{d})$, the conditional density describing the uncertainty in the parameters given the particular flow image, $\mathbf{d}$. To date, we have not computed the uncertainty in the optical flow image, $\mathbf{d}$, and therefore the observational errors are assumed to be negligible [4]. Letting $\mathbf{d}_{obs}$ denote the observed optical flow image, the hypothesis of negligible observational errors can be described using Tarantola's notation as:[5]

$$\begin{aligned} p(\mathbf{d}|\mathbf{d}_{obs}) &= \delta(\mathbf{d}, \mathbf{d}_{obs}) \\ &= \begin{cases} 1 & \text{for } \mathbf{d} = \mathbf{d}_{obs}, \\ 0 & \text{for } \mathbf{d} \neq \mathbf{d}_{obs}. \end{cases} \end{aligned} \tag{4.2}$$

This brings us to the flow models themselves. In the particular case treated in this chapter, there is no uncertainty in the projection as $\mathbf{m} = \mathbf{m}(\mathbf{d})$. (i.e. The modelization errors are assumed to be negligible.) In this case, perfect flow measurements lead to an estimate of the model parameters $\mathbf{m}$. Given that the observed flow image is denoted $\mathbf{d}_{obs}$, the conditional probability density function is represented by:

$$\begin{aligned} p(\mathbf{m}|\mathbf{d}_{obs}) &= \delta(\mathbf{m}, \mathbf{m}_{obs}), \\ &= \delta(\mathbf{m}, \mathbf{m}(\mathbf{d}_{obs})). \end{aligned} \tag{4.3}$$

[4]An optical flow algorithm that produces measures of the uncertainties associated with the flow images is currently being developed in the Artificial Perception Lab.

[5]This representation is rather unconventional in Bayesian terms, where this is no notion of a "true" data set.

As is the case of many forward problems, different flow images could generate the same model parameters. Therefore, if the marginal probability of the model parameters are sought, the contributions of all the data sets to that model would have to be considered. This is represented by the marginal density function:

$$\begin{aligned} p(\mathbf{m}) &= \int_D p(\mathbf{d},\mathbf{m})\; d\mathbf{d}, \\ &= \int_D p(\mathbf{m}|\mathbf{d})\; p(\mathbf{d})\; d\mathbf{d}, \\ &= \int_D \delta(\mathbf{m},\mathbf{m}(\mathbf{d}))\; p(\mathbf{d})\; d\mathbf{d}. \end{aligned} \tag{4.4}$$

$p(\mathbf{d})$ represents the *a priori* data information. The result implies that to compute the marginal model probability, one must add the contribution of all the data sets that could generate the model parameters.

3.3. *A Priori* Information on Models. In the current context, there are a finite number of object hypotheses, $\{O_i\}|_{i=1...K}$. The probability density function used to convey this knowledge is:

$$p(O) = \sum_{i=1}^{K} p(O_i)\; \delta(O, O_i) \tag{4.5}$$

where $p(O_i)$ defines the subjective *a priori* probability that the i^{th} object occurs.

3.4. Bayesian Solution to the Flow-based Recognition Problem. Each of the sources of information available were expressed as probability density functions. The Bayesian solution, outlined in Chapter 2 can easily be applied to the flow-based recognition problem. The theory outlined a strategy to compute the posterior probability in each model hypothesis, given an observed data set, $p(O_i|\mathbf{d})$ (see Equation (2.13)). Applying this strategy to this context is quite straightforward. Recall the form of the inverse solution:

$$p(O_i|\mathbf{d}) = p(O_i) \int_M \frac{p(\mathbf{m}|O_i)\; p(\mathbf{m}|\mathbf{d})}{p(\mathbf{m})}\; d\mathbf{m}, \quad i = 1..K. \tag{4.6}$$

Substituting $p(\mathbf{m}|O_i)$ and $p(\mathbf{m}|\mathbf{d})$ with their contextual forms found in Equations (4.1), (4.4) gives:[6]

$$
\begin{aligned}
p(O_i|\mathbf{d}) &= p(O_i) \int_M \frac{N(\mu_i - \mathbf{m}, \mathbf{C}_i)\ \delta(\mathbf{m}, \mathbf{m}(\mathbf{d}))}{p(\mathbf{m})}\, d\mathbf{m}, \quad i = 1..K, && (4.7)\\
&= \frac{p(O_i)\ N(\mu_i - \mathbf{m}(\mathbf{d}), \mathbf{C}_i)}{p(\mathbf{m}(\mathbf{d}))}, \quad i = 1..K,\\
&\propto p(O_i)\ N(\mu_i - \mathbf{m}(\mathbf{d}), \mathbf{C}_i), \quad i = 1..K. && (4.8)
\end{aligned}
$$

Recall that $N(\mu_i - \mathbf{m}, \mathbf{C}_i)$ is the multivariate normal distribution representation for the object hypothesis, evaluated at the model parameters of the measurement. $p(O_i)$ is the prior probability of the object hypothesis. The constant of proportionality, in this case, is the sum of the numerator terms over all objects. This can be derived as follows:

$$
\begin{aligned}
p(\mathbf{m}) &= \sum_{j=1}^{K} p(O_j)\ p(\mathbf{m}|O_j), && (4.9)\\
&= \sum_{j=1}^{K} p(O_j)\ N(\mu_j - \mathbf{m}, \mathbf{C}_j). && (4.10)
\end{aligned}
$$

$p(\mathbf{m}(\mathbf{d}))$ is computed by evaluating $p(\mathbf{m})$ at $\mathbf{m} = \mathbf{m}(\mathbf{d})$ as follows:

$$
p(\mathbf{m}(\mathbf{d})) = \sum_{j=1}^{K} p(O_j)\ N(\mu_j - \mathbf{m}(\mathbf{d}), \mathbf{C}_j). \tag{4.11}
$$

This term is simply the sum of the $p(O_i)\ N(\mu_i - \mathbf{m}(\mathbf{d}), \mathbf{C}_i)$ term for each object O_i.

The final solution indicates that the probability of each hypothesis, given the flow image, is computed quite easily: the flow image is projected onto the eigenbasis, the multivariate normal representation of the object in the database is evaluated at the resulting parameters, and the result is multiplied by the prior probability in the model hypothesis. Normalization is simply the sum of all the terms for all O_i. The final recognition result is a discrete conditional probability density function describing the likelihood of each of the models in

[6] The function for $p(\mathbf{m})$ derived in Equation 4.5 was not replaced here for clarity. The formulation will be true for any form of $p(\mathbf{m})$.

the database, given the flow data:

$$p(O_i|\mathbf{d}) = p(O_i)\ N(\mu_i - \mathbf{m}(\mathbf{d}), \mathbf{C}_i), \quad i = 1 \ldots K, \tag{4.12}$$

which can alternately be expressed as:

$$\begin{aligned} p(O|\mathbf{d}) &= \sum_{i=1}^{K} p(O_i|\mathbf{d})\ \delta(O, O_i), \\ &\propto \sum_{i=1}^{K}\ p(O_i)\ N(\mu_i - \mathbf{m}(\mathbf{d}), \mathbf{C}_i)\ \delta(O, O_i). \end{aligned}$$

By using Bayesian techniques, one could obtain the same solution in several alternative ways. For instance, in the particular case treated in this chapter, $\mathbf{m} = \mathbf{m}(\mathbf{d})$ simplifies the computation. A standard Bayesian approach, in this case, might lead to a solution more quickly. Without loss of information, $p(O|\mathbf{d})$ can be expressed as:

$$p(O|\mathbf{d}) = p(O_i|\mathbf{d}, \mathbf{m}(\mathbf{d})) \quad i = 1..K, \tag{4.13}$$

$$\begin{aligned} &= p(O_i|\mathbf{m}(\mathbf{d})) \quad i = 1..K, \\ &= \frac{p(O_i)\ p(\mathbf{m}(\mathbf{d})|O_i)}{p(\mathbf{m}(\mathbf{d}))} \quad i = 1..K. \end{aligned} \tag{4.14}$$

However, as was discussed in Section 1.4, the inverse solution provides a general recipe for the solution to a wide set of inverse problems. In general, the difficulty of applying the Bayesian approach is that it often requires foresight into the terms of interest in the final solution.

4. Accumulating Evidence Over Time

In previous sections, various problems associated with recognition based on flow images were discussed. These stem primarily from the problems associated with factoring out structure from the confounding effects of camera geometry, direction of motions, object shape and texture. Furthermore, the resulting image is noisy, and of poor resolution. The effect of all these factors is that recognition from a single viewpoint alone will not necessarily be stable. The hypothesis is that by accumulating evidence over time, a more robust solution is possible. In a dynamic environment, temporally gathering evidence is easily accomplished as the object (or the sensor) is moving through the scene acquiring images of the object from different viewpoints. As each of the flow images is computed from a set of intensity images, evidence can be inexpensively and efficiently accumulated on the level of the probabilities over time, by using the Bayesian chaining strategy, introduced in Section 2, that assigns the posterior probability distribution at time t, $p(O|\mathbf{d}_t)$, as the prior at time $t+1$. Once again the assumption of statistical data independence is made. As each flow image in the sequence is presented to the system by an external agent (either by a robot navigating through a scene or through a human moving an object in front of a camera), the assumption is that sampling one data set does not give specific information about the others. Although somewhat strong, this assumption permits simplification of the equations, and empirical evidence will indicate that it is not seriously violated in practice.

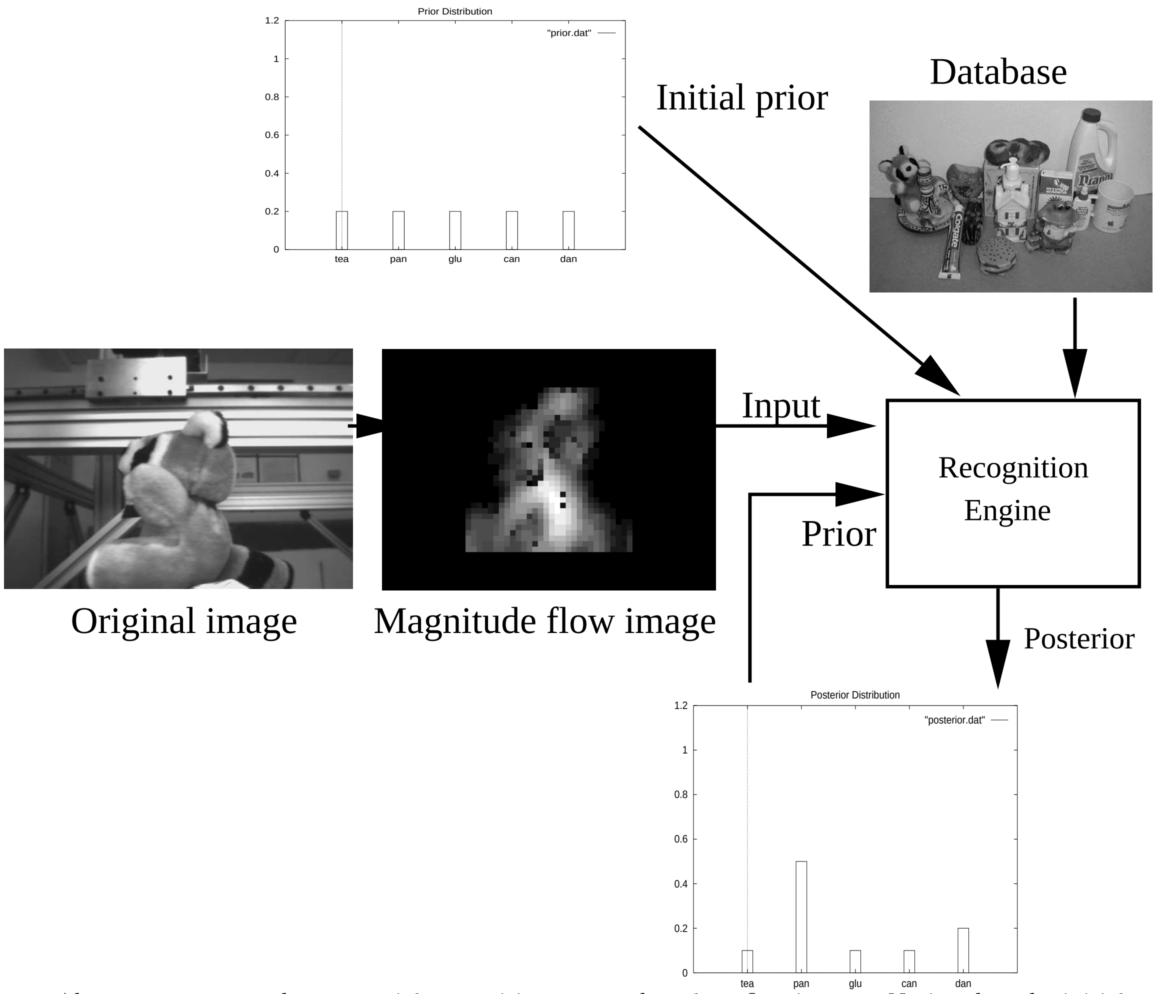
The accumulation process works as follows: as each flow image in the sequence is introduced to the system, probabilistic evidence is cascaded exactly as in equation (2.18) until a clear winner emerges. Substituting the posterior density function derived from one view (Equation(4.12)) as the prior for the next view leads to the following updating function for $p(O|\mathbf{d})$:

$$p(O|\mathbf{d}_{t+1}) \propto \sum_{i=1}^{K} p(O_i|\mathbf{d}_t)\; N(\mu_i - \mathbf{m}(\mathbf{d}_{t+1}), \mathbf{C}_i)\; \delta(O, O_i). \tag{4.15}$$

The idea is that conditioning inference on prior evidence should resolve ambiguities and lead to a winning solution in a short number of views. Figure 4.4 illustrates the flowchart of the appearance flow recognition system. Here, the object (a panda bear) is waved in front of a camera. The resulting flow images are sent to the recognition engine, along with the prior distribution and the database information. The result is a posterior distribution for the object. This example illustrates how the initial prior, a uniform distribution, is modified with the incoming data evidence (in the form of flow images). The resulting posterior illustrates a stronger belief in the panda than the rest of the objects. This posterior, in turn, becomes the prior for the next iteration.

5. Summary

In this chapter, the generalizability of the general Bayesian solution presented in Chapter 2 is illustrated through its application to the problem of appearance-based object recognition. This problem is quite different from the problem of 3D object recognition presented in the previous chapter, yet the same Bayesian framework could be applied to both. The



Above, one can see the sequential recognition system based on flow images. Notice that the initial priors are distributed uniformly. As the panda bear is moved in front of the camera, the belief in the panda (symbol: *pan*) increases. The new distribution showing a high degree of confidence in the panda is fed to the system as the new prior for the next iteration.

FIGURE 4.4. Sequential recognition system based on appearance flows.

particular problem tackled was that of recognizing objects, moving with respect to a camera, based on the signatures extracted from the resulting optical flow images. By using optical flow images, the strategy shows considerable improvement over traditional appearance-based methods, in its robustness in varying lighting conditions and backgrounds. Despite its benefits, no similar strategies have been found in the literature.

Rather than attempt to predict all possible motion trajectories *a priori*, the system generated a set of *characteristic motions* during training. This finite set was used to build an appearance manifold for motion, that should permit recognition based on a wider set of motions. Chapter 6 will illustrate that the system is indeed capable of such extrapolations, through recognition tests based on sets of motions outside the training set.

Both of the applications presented in this thesis are inherently difficult due to sensor noise. In the particular case of optical flow images, separating structural signatures from the confounding motion information is a particularly difficult problem, one that remains open. However, experimental results with both types of measurements will demonstrate that accumulating evidence for the different hypotheses over a sequence of views leads to a quick reduction of ambiguities, as the growing evidence in the correct model reduces the credibility of the other hypotheses. Choosing a winner after several viewpoints gives the correct result in most cases.

In the context presented thus far, at each iteration in the sequential estimation process, the recognition module passively receives a set of flow measurements acquired from an external agent. In this chapter, no discussion was presented regarding how the order and locations of the acquisition were decided upon. In fact, a random strategy was assumed. In the next chapter, an active recognition strategy is developed whereby the system navigates along an optimal trajectory for recognition. The hope is that the system will converge to the correct solution in a shorter number of steps and with a higher degree of accuracy.

CHAPTER 5

Entropy-Based Autonomous Navigation

The sequential recognition system introduced in previous chapters has been passive in the sense that the choice of locations for data acquisition has been left to the discretion of an external agent. In this chapter, additional constraints on the system lead to the development of an active recognition system that computes an optimal trajectory for data acquisition. The context is as follows: An active observer, moving through a scene with the task of identifying and localizing known objects, now has limited resources. It cannot spend a lot of time in one place so the computational overhead must be low. It needs to minimize the effort expended in gathering data so it must be economical in its movement. Finally, it must minimize its chances of error as false identifications are very costly. These constraints typify many applications of active vision, particularly in the context of mobile robotics. The strategy adopted in this chapter for overcoming these constraints is based on two observations: (i) that strong assertions can be made by accumulating evidence that might appear to be weak instantaneously, and (ii) knowing how to explore an environment (i.e. where to look) can be learned from local interactions with the objects that populate it. It would be most efficient to make maximal use of this learned information in planning gaze for recognition. This leads to the computational framework which is at the heart of this chapter.

In previous chapters, it was shown that evidence for different assertions can be expressed in the language of probability theory and the accumulation of evidence defined formally in terms of Bayes accumulation. Since the observer is free to interact with each of the objects prior to exploration (i.e. off-line), then object probabilities associated with image measures can be learned from training. This idea can be taken one step further. If the parameters associated with each measurement are also recorded then one can also build a model of how these parameters influence the confidence for the resulting assertions. Section 1 will illustrate that the Shannon entropy can be used to define a measure of ambiguity for the resulting posterior distributions. This leads to the introduction of the *entropy map*, used to relate ambiguity to camera (viewing) position. These maps serve as the basis for the active vision system described in Section 2. By choosing viewpoints that minimize ambiguity, the system seeks out locations off-line that are maximally informative. On-line navigation then uses these maps to guide the sensor to these optimal locations. Hence, fewer observations are required to arrive at a confident assertion.

The strategy can be seen by analogy to human behavior. Consider a person trying to identify an object she sees. She has an idea that it is a particular object that she has seen before, and she wishes to verify that it is indeed the object she is thinking of. The natural reaction is to move her head to a location that would confirm the hypothesis. This location would be one that exposes a feature that she learned is unique to the object. This information has been stored *a priori* during her introduction to the object.

This leads to a second important contribution resulting from the introduction of the entropy map: a strategy for enumerating the *informative views* of an object. Here, the extent to which a viewpoint is informative refers to the degree of discriminability of the object with respect to the others in the database. The problem of storing a set of views that are informative for recognition for each object in the database has been of interest to researchers for many years [**21**]. This has been particularly important for approaches that represent objects by a set of a viewer centered representations referred to as *aspect graphs* [**21, 34, 58, 81, 89**]. These graphs represent the possible appearances of an object (in terms of areas that contain a visible set of features) in the form of graphs, where nodes refer to characteristic views: stable viewpoints that capture sufficient object structure for accurate recognition. Arcs refer to visual events between the views. These views are often chosen by hand, leading to a subjective selection that is often biased and non-repeatable. In fact, the choice is often based on the number of visible features from the particular view, where the choice of feature is subjective. It will be shown that entropy maps can be built around the object of interest in a structured fashion, linking recognition ambiguity to each view. These maps can be used to *automatically* identify viewpoints that are informative for recognition for each object. This decision is based on information theoretic notions of ambiguity rather than subjective notions of "good" and "bad" views. Further, the system assigns a level of confidence to each assertion. This permits higher level processes to determine operationally which views (if any) are sufficiently informative for the task at hand. In this sense, an notion analogous to characteristic views can be attained for motion as well.

Many active vision approaches employ minimization techniques in order to navigate to optimal locations. The problem is that most use techniques such as on-line gradient descent which are local in nature. This risks leading the sensor towards local minima which, in the context of recognition, can be associated with high belief in the wrong object. In the strategy proposed in this chapter, the *global* entropy minimum can be precomputed off-line, and on-line navigation can be guided towards these locations. This offers an enormous advantage over most approaches, especially where obtaining a correct solution is critical. Furthermore, expensive on-line computations are off-loaded to the training process.

Although the approach is general and can be applied to many contexts, Section 4 illustrates one interesting application of the strategy presented in this chapter. Here, the focus returns to the difficult problem of recognizing objects based on flow images. For this application, the mobile observer consists of a monochrome television camera mounted on the end effector of a gantry robot (Figure 5.1). The camera is free to move about the workspace of the gantry in which the different test objects are placed (i.e. stationary environment). As the camera moves relative to an object, an optical flow pattern is induced on the image sensor. At each iteration, the recognition system described in Chapter 4 is implemented. The task that the system must perform is to generate an optimal trajectory (the shortest sequence) that will result in the correct assertion.

1. Entropy Maps

With the recognition strategy in place, an active agent can move around a scene, gathering evidence in the various object hypotheses until a satisfactory confidence level is attained. It is essential that the strategy minimize the chances of the system of arriving at an incorrect recognition result. Furthermore, as resources are limited, the system needs to converge to a solution in a short number of steps.

In addressing these requirements, a strategy is proposed that takes maximal advantage of *a priori* information available in order to attain a fast and reliable on-line solution. Specifically, *entropy maps* are built off-line during training to relate recognition ambiguity to viewing position. Once a map is built for each object in the database, the system can store the locations that are maximally informative in terms of disambiguating between the

FIGURE 5.1. (a) Navigation setup.

objects in the database. This information can then be made available to the agent on-line to aid in recognition. The hypothesis is that if the structure of the maps are such that: (i) entropy varies continuously almost everywhere, and (ii) "good" neighbourhoods exist (in terms of high belief in the right model), it becomes feasible to base on-line navigation experiments on them.

This leads to the question of how to build these maps in practice. The first thing that is required is a measure that predicts the likelihood of ambiguous recognition results as a function of viewing position. The recognition result is defined as a posterior distribution function over the set of K object hypotheses, $\{O_i\}_{i=1..K}$, given a particular measurement vector $\mathbf{d}$, denoted $P(O|\mathbf{d})$. Therefore, a suitable measure is defined in terms of the Shannon entropy [**26**] exactly as in Equation 2.26:

$$H(P(O|\mathbf{d})) = \sum_{i=1}^{K} p(O_i|\mathbf{d}) \; log \; \frac{1}{p(O_i|\mathbf{d})}, \tag{5.1}$$

which is a measure of the ambiguity of the posterior distribution produced by a recognition experiment. Higher entropies reflect greater ambiguity.

This result is useful on several levels. First, a measure of recognition ambiguity is useful for an external agent to be able to assess the results from a particular viewpoint. It can be used to determine whether further data are required, or whether sufficient convergence to a single hypothesis has been attained. Further, it can be used to assess whether a viewpoint is informative or not (as discussed in Chapter 2). These important results are possible due to the probabilistic nature of the recognition result, and would be much more difficult (and context-specific) had a deterministic strategy been developed. The generality of the result is apparent, as it can be used to obtain the ambiguity of any distribution, regardless of the strategy used to determine it.

For the problem at hand, the measure is useful in building an entropy map for each object in the database, off-line during training. For each object, the corresponding map is parameterized on a tessellated viewsphere with the object at origin. Each map is then built as follows:

1. During training, image measures, $\mathbf{d}$, are sampled at each coordinate of the viewsphere for each object in the database. Each image measurement, $\mathbf{d}$, is stored along with its coordinates of acquisition on the viewsphere (i.e. latitude and longitude of camera on viewsphere). The pose of the object during training defines its reference pose.
2. A Bayesian learning strategy (see Chapter 2) is applied to the cumulative set of measures as before.
3. Recognition is then performed on each training measurement, resulting in the association of the posterior distribution, $P(O|\mathbf{d})$, to each coordinate.
4. The entropy for each measurement, $H(P(O|\mathbf{d}))$ is computed ((2.26)) and stored at its associated coordinate of its respective viewsphere.

This recipe can best be seen through an example. Figure 5.3 in Section 4 illustrates the application of the recipe to the context of recognition based on flow images.

In practice, some additional conditioning is required before the entropy map can be used for gaze planning purposes, namely in the enforcement of particular smoothness constraints which are essential for stability. For example, the gaze planner, using the map to determine minimal entropy viewing positions, wishes to avoid locations corresponding to areas in the entropy field where the likelihood of making a wrong assertions is strong. Should these positions lie near the minimum entropy location, then slight errors in positioning (or equivalently in determining the pose of the entropy map relative to the data acquired) could result in sampling at precisely the *wrong* locations. These constraints are made explicit by applying the following non-linear smoothing operator,

$$H(P(O|\mathbf{d}_i)) = \frac{\sum_j cos(\theta_{ij}) \times H(P(O|\mathbf{d}_j))}{\sum_j cos(\theta_{ij})}, \tag{5.2}$$

where $\mathbf{d}_i$ is the data vector gathered at viewsphere location i of the operator, $\mathbf{d}_j$ are the points in the local neighbourhood indexed by j, and θ_{ij} the angle subtended from the origin of the sphere between $\mathbf{d}_i$ and $\mathbf{d}_j$. The resulting function can be seen as weighing the local entropy value with the support from its neighbours, i.e. neighbourhoods with strong belief in the wrong object provide strong negative support for a location. The minimal entropy location on this map will correspond to an optimal location which is stable with respect to localization errors. This implies that its low entropy belief in the correct object is maximally supported by its neighbouring entropy values. An example of an entropy map and its corresponding smoothed map, within the context of recognition based on optical flow images, can be found in Figure 5.4. A full description of the structure of the map will be given in Section 4.

The resulting entropy map can be very informative in the context of planning gaze for object recognition. It provides a *quantitative* prediction of the level of difficulty of recognizing each object in an on-line experiment. A secondary bi-product is that a set of informative views can be automatically generated off-line. These are based on mathematically defined notions of ambiguity, rather than on subjective interpretations of importance. In addition, a degree of ambiguity is associated with each tile in the viewsphere, permitting higher level processing based on the results. The next section will introduce a strategy for using these maps in practice, in on-line recognition experiments.

2. Using Entropy Maps to Plan Gaze

With the entropy maps built, gaze planning based on them becomes possible. Prior to planning gaze, two problems must be solved: 1) a particular map must be selected and 2) the pose of that map must be determined relative to the data acquired. Each of these problems will be addressed next.

As measurements are made on-line, the maximum a posteriori (MAP) solution corresponding to $p(O|\mathbf{d})$ is used to determine the most likely object hypothesis for the measured data $\mathbf{d}$. In order to develop a working solution, the following assumption is made: The entropy map of the MAP result is sufficient for planning the next best view. The MAP result communicates the object that the system has the most confidence in. By examining its entropy map, the system has access to its most informative views, and can use them in on-line experiments. Of course, the MAP result is not always accurate, and, in the worst case, the system might choose the wrong map on which to based its navigation. This case is no worse, on average, than taking a random step: The system will acquire data, and will gain new evidence in the various hypotheses.

Of course, cases may exist where little confidence is placed in the MAP solution. In this case, it might be more instructive to visit several entropy maps at each iteration (more will be said about this later). The strategy can easily be modified to accommodate such motions, however it is accompanied by an additional cost which might bring to question its gain in speed over a random navigation strategy. This would have to be examined in the particular context of the problem at hand.

With the most likely object hypothesis map chosen, what should the navigation strategy be? The approach taken is based on the following assumption:

Assumption 7. *Discriminant Views. For each object in the database, there exists a finite set of viewpoints which, taken together, permit discrimination of the object from the others in the database.*

This assumption implies that it should be possible to visit a finite set of viewpoints in order to accurately identify each object in the database. In the current context, a somewhat stronger version of this assumption will be taken in order to constrain navigation to a shorter number of steps. For each object in the database, the assumption is that there exists a *single* viewpoint from which discrimination from the others is possible. In cases where this assumption is too strong, extension of the approach to include a larger set of viewpoints is possible.

Should this assumption prove valid, one could store the location of the most discriminating view for each object in the database. The next view chosen is then the best view on the most likely object hypothesis map. Navigation would then entail moving towards this location during on-line recognition experiments. The most discriminating viewpoint in the current context refers to the one with the lowest entropy (i.e. lowest ambiguity) belief in the correct model.

In practice, the validity of the assumption can be tested through an operationally defined level of entropy acceptability. Should there not exist a sufficiently low entropy location for an object of interest, then that object cannot be recognized from any single viewpoint using the existing set of features. In some cases, it might then be desirable to remove the object from the database altogether (or to increase the sampling resolution of the entropy map). Alternately, the assumption can be weakened to permit several discriminant viewpoints to be visited. The ability to pre-examine the recognition potential of the objects database *a priori* in this fashion is an advantage of the approach.

Pose can be estimated at minimal expense by retaining the location information along with the image measures acquired during training. For example, appearance-based methods can be used to index these measures using the data acquired on-line [**71**]. This involves choosing the closest image on the appearance manifold to the one acquired on-line. In fact, the implementation described in Chapter 4 already uses appearance-based techniques in the process of determining the likelihoods for the different object hypotheses. As such, the computational overhead of determining pose is minimal in this case. Errors in pose estimation are accommodated, in part, by the smoothing applied to the entropy map, and a strategy that chooses the next best view at the minimum entropy location of the smoothed map. This avoids placement in the vicinity of singularities and discontinuities. Figure 5.2 (Steps 1–2) illustrates the step that matches the current camera pose to the corresponding pose on the entropy map.

Once the camera pose is established in the coordinates of the training viewsphere (i.e. the entropy map), it is straightforward to determine the relative transform taking the camera to the desired position within this frame (Figure 5.2, Steps 2–3)[1]. By applying this same transform to the current camera frame, the camera is positioned accordingly as shown in Figure 5.2 (Step 4). As the camera navigates through the scene, the sequential recognition strategy is applied to accumulate evidence in the various hypotheses. Over time, the expectation is that confidence in an incorrectly chosen hypothesis will decrease as further evidence is uncovered.

It is worth noting that in orthodox statistics, it is often the case that estimators and statistical tests depend on the strategy used for sampling. Here, it is emphasized that by choosing the location for data acquisition, the system is not biasing the recognition results systematically away from the true result [**68**]. The *likelihood principle* states that inferences should be based on the actual data acquired, not on some hypothetical data set that we might have gathered but did not. Bayesian techniques are consistent with this principle. In this case, the posterior distribution expresses the belief in each object *given* the actual data

[1]Strategies for computing the transform are left as an operational issue.

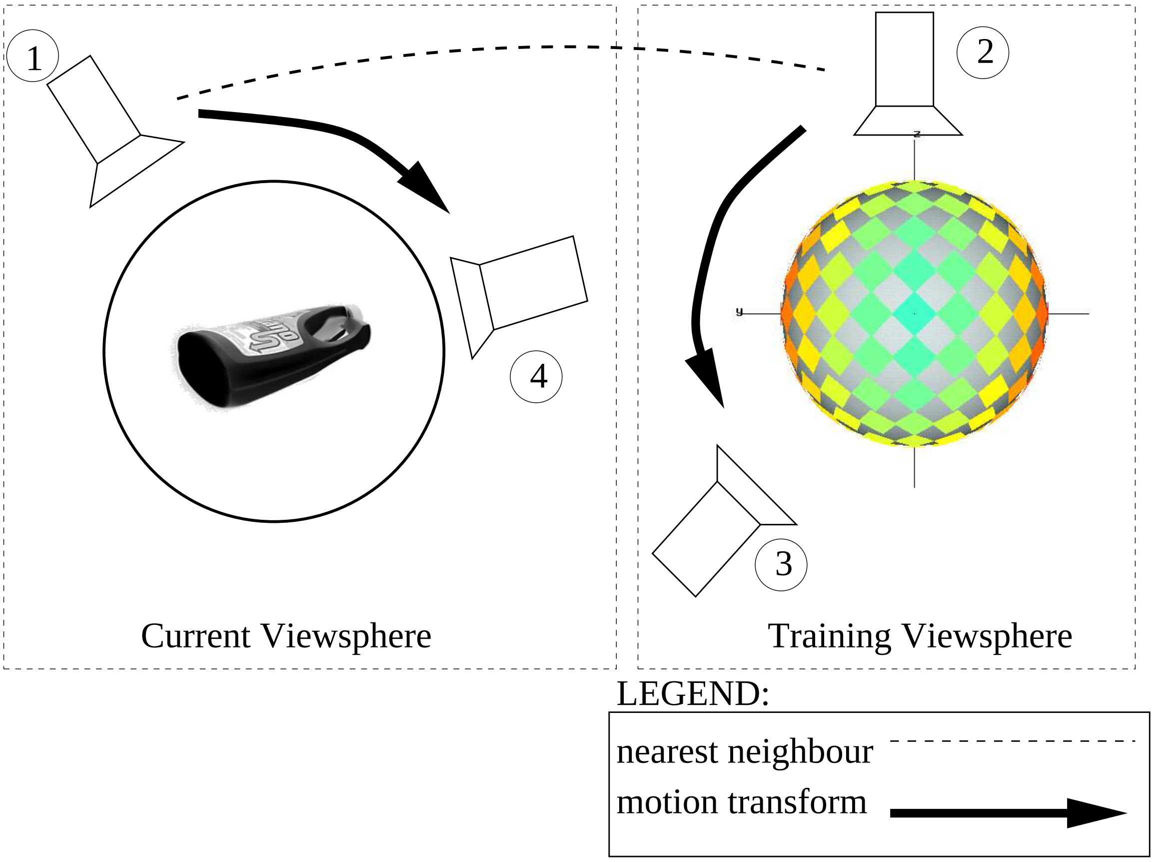


FIGURE 5.2. Navigation strategy.

set, **d**, gathered. The inference is communicated only in terms of this data set. The notion of a *true result* does not exist.

3. Related Work in Active Recognition

Although various active vision strategies exist, very few concentrate on the problem of moving a sensor to best disambiguate objects in a database (i.e. active recognition). One strategy that focuses on the problem of locating, and attending to an object in the scene is that of Rimey and Brown [**82**]. They use Bayesian networks to explicitly model the interaction between objects in a scene. This involves computing conditional probabilities relating the likelihood of the location of one object given another in a particular scene. The approach is focused more on scene context than on recognition of a particular object.

An approach based on functional verification procedures is that of Stark et al. [**90**]. The authors develop functional rule-based object descriptions based on relations between parts. The idea is to verify object identity through examination of these functional features.

The method focuses on recognition of CAD models of chairs. It is not clear how well it generalizes.

Wilkes and Tsotsos [**104**] present an active recognition strategy for the identification of polyhedral objects. Their strategy is to move the camera to locations that minimize the projected lengths of two non-parallel edges in the image.

Dickinson et al. [**30**] present an active recognition strategy that combines the use of an attention mechanism for focusing the search for a 3-D object in a 2-D image, with a viewpoint control strategy for disambiguating recovered object features. The strategy is based on an aspect representation of an object's parts, the aspect prediction graph, that represents the visual events that occur when the camera moves from one viewpoint to the next. The authors derive a function based on probabilities to rank the possible directions in which the camera should be moved in order to distinguish an object part.

Hutchinson and Kak [**48**] describe a method for choosing the next viewpoint in order to disambiguate between objects in a database. Based on a set of current hypotheses about the identity and position of an object, they evaluate candidate sensing operations with regard to their effectiveness in minimizing ambiguity. It does so by predicting, for each hypothesis, the set of features that would be observed given the candidate action, and the next hypothesis set that would form if those features were found. The ambiguity of the predicted set is computed. This is repeated for each hypothesis set. The maximum ambiguity that is associated with that motion is noted. The sensing operation that minimizes the maximum ambiguity is taken. The authors use the Dempster-Shafer methodology to perform reasoning, and define a notion of ambiguity which is similar to entropy. It is not a true measure of entropy as the Dempster-Shafer formalism produces *beliefs* rather than true probabilities (as discussed in Chapter 2).

Strategies for sensor planning that make use of prior information can be found in the mobile robotics literature. One such method was developed by Takeda et al. [**92**] for the problem of robot localization within a scene. Their strategy is to build "Sensory Uncertainty Fields" (SUF) *a priori*. These estimate the distribution of possible errors in a robot configuration. Path planning consists of navigating from one point to the next based on minimizing a functional based on the magnitude of the SUF.

Other strategies exist that use a similar notion of entropy during *on-line* computations, in order to maximize information gain during navigation. Callari and Ferrie [**23**] proposed such a strategy for active recognition, and a similar approach was taken by Burgard at al.

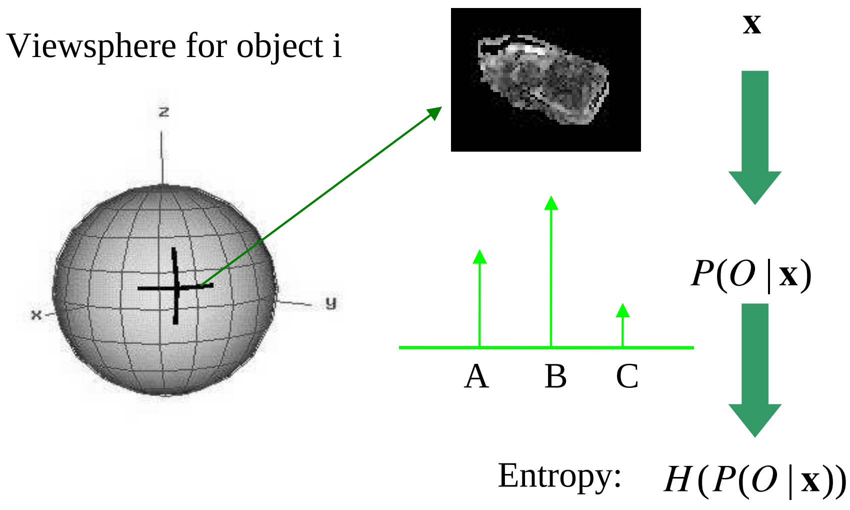
[**22, 83**] in the context of the mobile robotics problem of localizing an object in a scene. The key difference in the work presented in this thesis is that here the idea is to maximize the *a priori* information available, by building the entropy maps *off-line* and using them to guide the on-line navigation. The advantages of such a strategy were discussed above.

The most relevant active recognition work is that of Schiele and Crowley [**85**], who performed statistical recognition based on multi-dimensional receptive field histograms built from local Gaussian derivative operators in a 2D image. This work is similar in philosophy to the strategy presented in this thesis in that recognition ambiguity, linked to measurements during training, is used during data acquisition. The strategy is different in a number of important details: (1) The nature and form of the recognition problem and solution were different. (2) Their strategy did not involve constructing a 3D entropy map about each object in the database. (3) No accumulation of hypothesis evidence was performed (each viewpoint was considered independently). (4) The goal was simply maximum likelihood identity verification, which was completed after two consistent interpretations. (5) Most importantly, no navigation experiments were performed. The algorithm was only tested on the set of 2D Columbia University database images acquired by rotating each set of objects in front of a camera along one dimension.

4. Case Study: Navigation Using Flow Images

The preceding framework is now applied to the problem of planning the gaze of an active observer moving through a stationary environment with the task of identifying objects within it. Motion induces a sequence of optical flow images on the camera retina which will serve as the basic input to the recognition system. In essence, the strategy outlined in the previous section can be directly applied to the context of recognition based on flow images. The main difference in using flow images is that, at each position on the viewsphere, the system stores a series of entropy values, each associated with a different movement of the camera relative to the object at that location. The system then computes the average entropy at that location and proceeds with the non-linear smoothing as in Equation (5.2). The best location to go to is chosen as before, and the movement to induce is the one that generates the lowest entropy among the set of possible movements at that location.

Figure 5.3 illustrates a diagram of the steps involved in building an entropy map in the context of recognition based on flow images.



Above, one can see the steps involved in computing the entropy map for an object in the database off-line, during training. At the center of each patch in the tessellated viewsphere about the object, the camera is swept along several arc-like motions. The result is a flow image corresponding to each movement. Recognition is performed on each flow and the posterior distribution is computed. The entropy of the distribution is stored along with the coordinates associated with each motion.

FIGURE 5.3. Computing entropy maps.

The structure of the resulting entropy maps is best seen by example. Figure 5.4 shows two viewpoints of the entropy map resulting from gathering optical flow images densely about a viewsphere surrounding a particular database object. Each tile corresponds to a camera view of the object at the origin (superimposed onto a shaded sphere). The crosses about the maps depict the axes at the origin of the sphere, indicating camera position relative to the viewsphere. For illustration purposes, the lowest entropy values (among all the motions) are displayed at each location. In cases where some directions have correct and some incorrect recognition results, the lowest entropy value in the correct model is illustrated. The tiles are colored in accordance with their raw entropy value in Figure 5.4(a), and to their smoothed value in Figure 5.4(b). The colors range from low entropy (blue) to high entropy (red). The greyscale tiles indicate an incorrect MAP solution, with entropy ranging from black (low) to white (high).

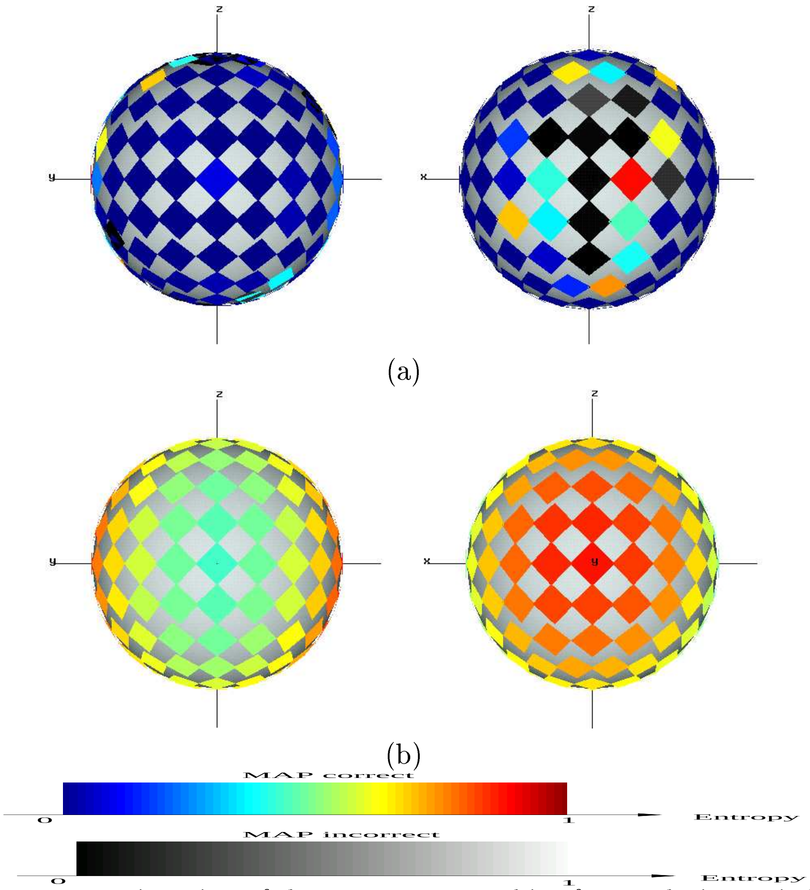
Examining the resulting maps, one can see that their structure is such that areas that result in inter-class confusion are concentrated into relatively isolated patches on the view-sphere. In addition, the ambiguity increases in the areas surrounding the "worst" locations. The structure of the map lends itself to our navigation strategy as generally large patches are comprised of optimal viewpoints, in terms of high confidence in the correct model. The smoothed versions of the entropy map at the same location, as seen in Figure 5.4(b), illustrate both the best (the left map) and worst (the right map) locations for discrimination, in terms of location and movement. Notice that the best location is maximally far from the worst areas of high confusion or strong belief in the wrong object. This leads to the hypothesis that moving towards this location will lead to a correct solution with relative safety.

The navigation based on these entropy maps works exactly as described in Section 2, with the system now using the location and direction of movement to build a camera frame at its current position. The difference here is that the sensor is directed to the optimal location *and* optimal direction of motion, where new data are gathered. Over time, the expectation is that confidence in an incorrect hypothesis will decrease as further evidence is uncovered. Figures 5.5 and 5.6 illustrate the steps involved in gaze planning based on flow images,

Preliminary experimental results based on the strategy proposed in this chapter were presented in [**4, 5**]. Chapter 7 will illustrate the results of a set of more comprehensive navigation experiments. The input to the system will be optical flow images. The system will be compared to a navigation approach whereby the next viewpoint is chosen randomly. It will be shown that the proposed strategy converges to a solution faster and more reliably than the random approach.

5. Summary

In this chapter, an active recognition strategy was introduced whereby *entropy maps*, encoding object discriminability as a function of viewing position, are computed off-line. During on-line recognition experiments, these maps serve to guide a sensor towards areas that minimize the inter-class confusion between competing object hypotheses. The strategy was applied to the problem of recognizing objects based on signatures in their optical



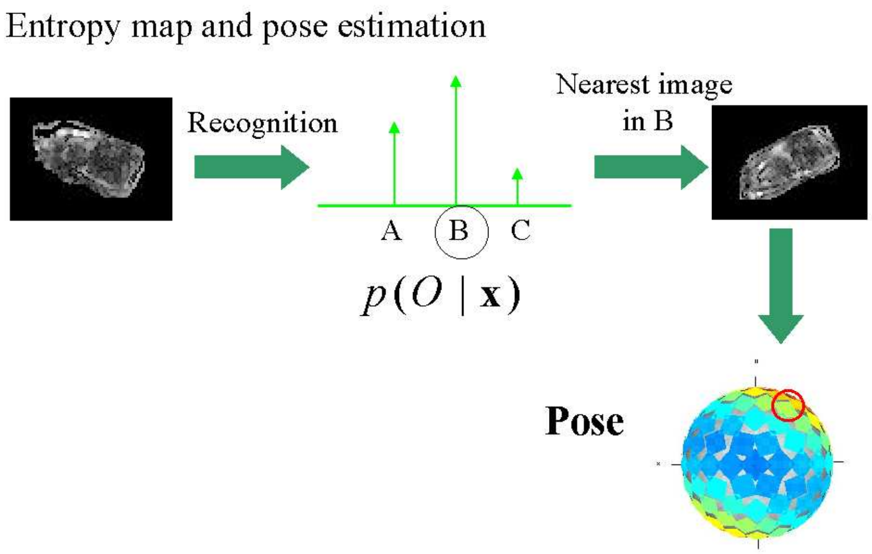
Above, one can see two viewpoints of the entropy map resulting from gathering optical flow images densely about a viewsphere surrounding a particular database object. The tiles are colored in accordance with their raw entropy value in (a), and to their smoothed value in (b). The colors range from low entropy (blue) to high entropy (red). The greyscale tiles indicate an incorrect MAP solution, with entropy ranging from black (low) to white (high)

FIGURE 5.4. (a) Two views of an entropy map, (b) corresponding smoothed maps.

flow images. The Bayesian recognition strategy introduced in previous chapters was easily incorporated into the navigation system. The advantage of the approach is that the combination of the feedback afforded by the gaze-planning strategy and the use of temporal filtering should reduce recognition ambiguity, and lead to a more robust solution. In the particular case of optical flow images, an interesting question (and a direction for future research) concerns the degree to which the basis obtained through training generalizes to a broader range of motions. The hypothesis here is that the combination of the two strategies should permit generalization such that recognition based on a fairly wide range of motions is possible. Later chapters will verify that the strategy works in practice.

The strategy presented here can be considered as an *object verification* strategy: the system estimates the object identity and moves to the best location to confirm it. The

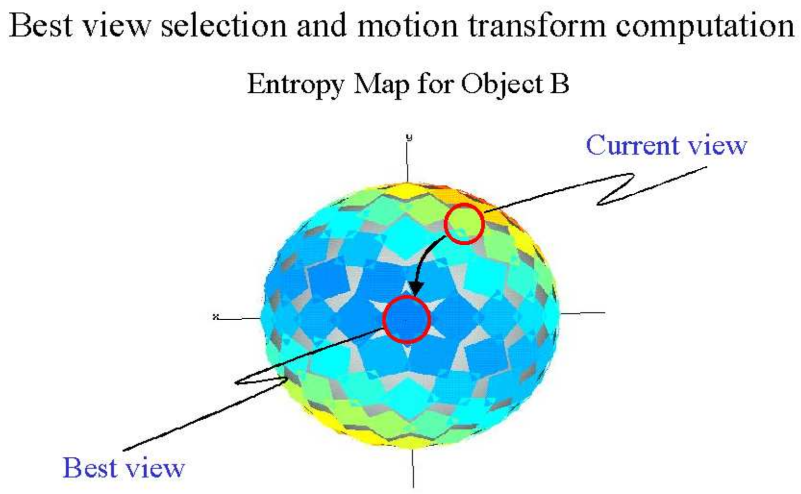
(a)



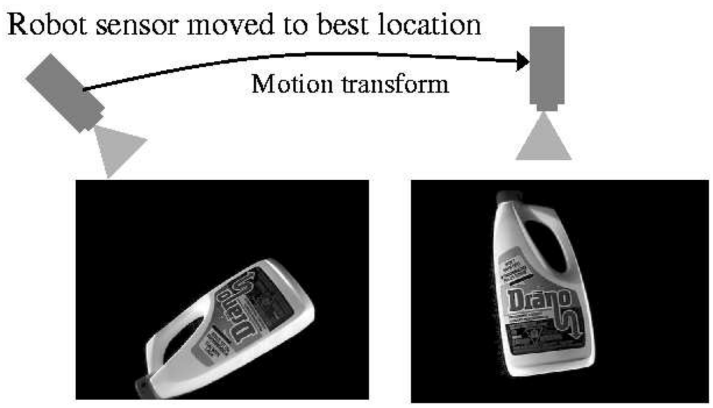
(b)

Above one can see the first two steps of the navigation process. First, flow images are computed in the usual manner. Next, pose is computed as follows: Recognition is performed on the flow image. The MAP result is computed (i.e. object B in this case), and the closest image in B to the input is computed. From here, the estimate of the position on B's entropy map is found.

FIGURE 5.5. Gaze planning with flow images. (a) Flow image acquisition, (b) pose estimation.



(c)



(d)

In the third stage of the navigation process, the best view on the entropy map of the MAP result (i.e. B) is located. The motion transform taking the sensor from the current position to the next one is computed. Finally, the motion transform is applied to the sensor, and it is moved to the next location. From here, new flow images will be acquired.

FIGURE 5.6. Gaze planning with flow images. (c) Best view selection and motion transform computation, (d) sensor moved to best location.

strategy may not always be optimal, however the goal is to best guide the sensor based on all *a priori* information available. The posterior probabilities keep track of system's current understanding regarding object identities.

The strategy is quite general. It should also be adaptable to any other sensory modalities which encode the structure of the local environment. One interesting application would be to apply the strategy to the problem of multi-part object recognition. The system could learn *a priori* which parts of the object render it unique based on the objects in the database. On-line, the system can be guided towards these parts of the most likely object hypothesis. In this case, pose can be determined based on part neighbourhood relationships.

One interesting question that was posed in the previous chapter was: Is there a set of *characteristic motions*, analogous to characteristic views, that is sufficient to ensure successful recognition results from a wide set of motions? This question was partially answered in the previous chapter, through the introduction of a strategy that built an appearance manifold for motion based on a small set of motion segments. Recognition on a larger set became possible. The entropy maps developed in this chapter can be used to extend the notion of *characteristic motions* to one that is analogous to the traditional notion of characteristic views, i.e. the most informative and stable position/motion combinations for recognition can be extracted directly from the maps. An interesting extension of the work would be to automate the extraction of a sufficient set of such viewpoints for recognition. To date, this problem has not been explored in the literature.

Various improvements to the navigation strategy can be employed. First, although the nearest neighbour strategy for estimating pose is quite effective, a more precise pose estimate would ensure even faster and more reliable convergence to the correct solution. However, estimating object pose is a difficult open problem, and therefore left as a topic for future research.

Second, the postulate that a single entropy map can be used for on-line navigation needs to be re-examined. Due to possible errors in the entropy map selection, a better approach would be to base navigation on the entropy maps of *several* competing hypotheses. This should ensure greater stability during traversal. However, finding an adequate solution to this problem is difficult. The obvious possibility is to visit each of the best views for the top competing solutions. Of course, if high ambiguity exists, this would involve visiting many locations. This would add undue expense to the strategy in cases where the ambiguity could be reduced significantly after only one view. Another approach would be to store, *a* *priori*, all the combinations of entropy maps for the set of pairs, triples, etc. of objects in the database. This quickly leads to a combinatoric explosion for any non-trivial database. Finally, an average entropy map could be built, one that leads to locations that are, on average, useful for discriminating between *all* the objects in the database. Of course, should the best locations for each object be spread out all over the viewsphere, the best locations on average could be quite poor for distinguishing any of the objects.

Finally, the on-line entropy results were not used except to establish convergence. A strategy for future consideration would be to use the on-line entropy as a measure of ambiguity in the recognition assertions made at the current location. This permits the possibility of assessing the validity of using the MAP assertion in navigation.

CHAPTER 6

Recognition Experiments and Results

This thesis introduces a general framework for object recognition, based on mathematical foundations that invoke fundamental laws in statistics. However, the robustness of the theory can only be verified through experimentation with a real vision system, where unforeseen problems can occur, rendering assumptions invalid and requiring additional constraints to be placed. This chapter attempts to validate the main hypotheses in this thesis through a series of real experiments. In particular, the system's robustness at sequentially recognizing objects will be examined through its application to two very different contexts.

The first set of experiments tests the ability of the system to perform recognition based on 3D parametric models. Section 1 applies the theory to the case of recognition of superellipsoid models, fit to laser ranger-finder data gathered during an exploration sequence. In Section 2, experiments are performed on parametric descriptors of optical flow images of moving objects, acquired through appearance-based techniques. The expectation is that, in both cases, the system is capable of converging to the correct solution after a short number of views.

1. Case I: 3D Part Recognition

The focus of the first set of experiments is "recognition by parts", where objects are seen as collections of independent parts, and recognition consists comparing each part of a segmented object to the parts of objects in the database. Here, topological relationships are not yet considered. Recognizing parts of articulated models is challenging due to problems of self-occlusion and segmentation.

Two articulated models were used in the recognition experiments: (1) a potato-head toy consisting of two ears, two eyes, a nose and a head, and (2) an alarm clock with two

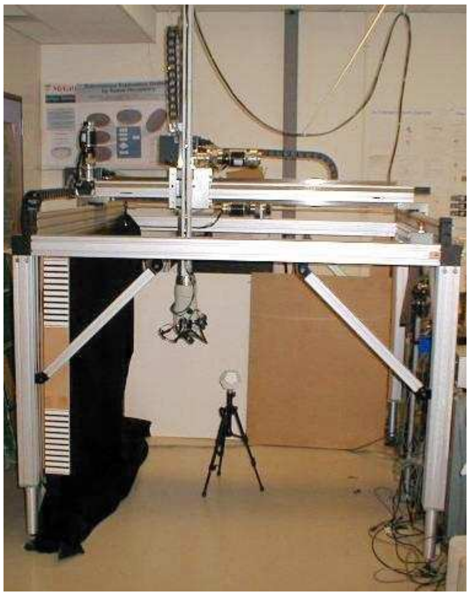
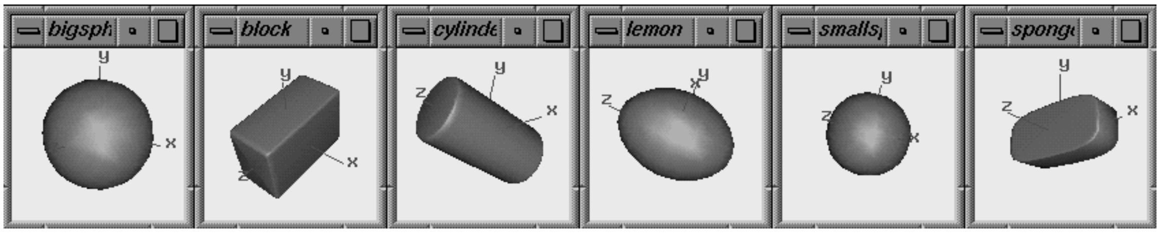


FIGURE 6.1. The CRS gantry robot used for experiments. The mobile laser range-finding system that was used to construct object models was mounted onto the end-effector of the robot arm.

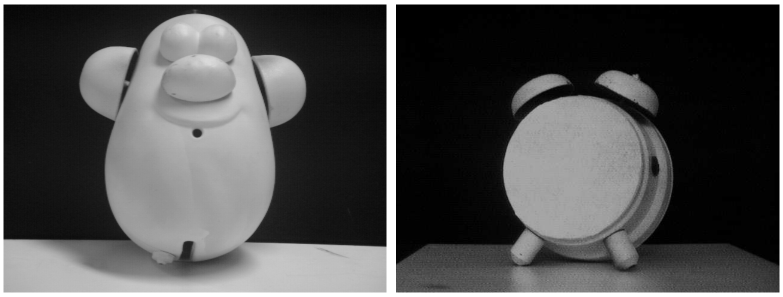
bells, two legs, a cylindrical face and a back. In addition, six single-part "distractors" were placed in the database in order to render the recognition task more difficult. These objects consisted of: two spheres, a block, a cylinder, a lemon, and a block with rounded edges. These objects were chosen for the experiments because they consisted of parts that generally conformed well to non-deformable superellipsoids. In addition, the parts varied in size and shape, so as not to be clustered together too tightly in five-dimensional feature space. However, sufficient overlap of their distributions in several dimensions ensured that the recognition procedure was challenged in its discrimination task.

1.1. Experimental Setup. During training, models are constructed through a process of *autonomous exploration* [**98, 102, 103**] in which a part-oriented, articulated description of an object is inferred through successive probes with a laser range-finding system (see Chapter 3). Figure 6.1 shows the set-up used to perform experiments — a plane-of-light laser range-finding system mounted on the end-effector of a CRS gantry robot arm. Although superellipsoid models were chosen to test the system, the strategy applies to the recognition of any parametric primitive. Training was difficult due to the compounding effects of noise and uncertainty, propagated from each stage of the bottom-up process. Figure 6.2 illustrates the actual potato-head and alarm clock objects used in recognition experiments, and the representative models of each object that result from training.

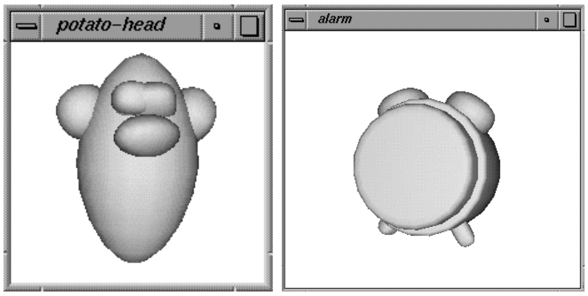


BS B C L SS RB

1emDisplayed above are reference objects that result from training acting as "distractors" for the recognition procedure: a big sphere (BS), a block (B), a cylinder (C), a lemon (L), a smaller sphere (SS), and a rounded block (RB).



a) Original potato-head and alarm clock.



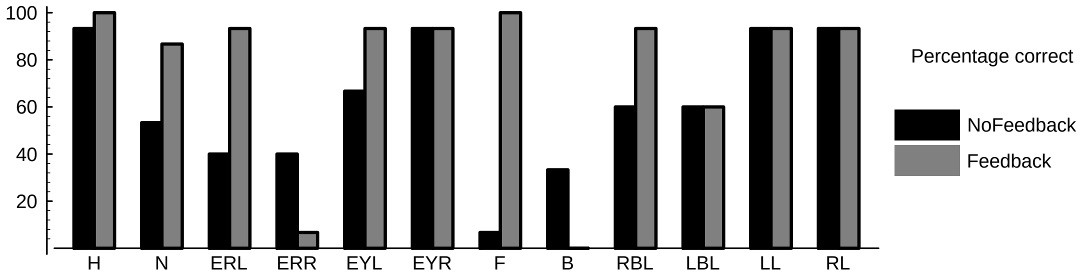
b) Reference potato-head and alarm clock models created by training.

FIGURE 6.2. The reference parts resulting from training.

During each iteration of on-line sequential recognition experiments, models of the parts of the potato-head toy and the alarm clock were computed using only the partial surface information from that view (i.e. range data from previous views were not merged on the surface level). These models were computed using the same bottom-up strategy that was used during the training procedure. Once again, at each stage of the process, noise and uncertainty were propagated to the next level. The result was that the recognition engine, receiving information at the top of the system, had to make assessments as to the object's identity using information that was, at times, highly uncertain and even erroneous. This compounded the difficulty of the recognition task.

Recognition was performed using two different platforms: with and without probabilistic feedback. In the first case, evidence in the form of probabilities in each of the models in the database was accumulated over the viewpoints visited thus far by cascading the beliefs from one stage as priors for the next stage, exactly as described in Section 3. In the second case, the results were computed with no embedded prior knowledge, i.e. the case with a uniform prior at each iteration.

1.2. Experimental Results. The results of both types of recognition experiments for the various parts of each object can be found in figure 6.3. Here, the maximum a posteriori (MAP) solution (i.e. the object hypothesis with the highest posterior belief) was chosen at each iteration and the percentage of correct identifications was tabulated at convergence, i.e. excluding the results of the first few intervals. This ensured that the system had sufficiently stabilized prior to examination. For these experiments, the results of 15 recognition iterations were examined after an initial 5 intervals. The size of the initial interval was arbitrary, simply chosen for the purposes of illustrating the general trends.



1emAbove are the results of recognizing each of the models in the database (each abbreviated by a few letters) from new viewpoints. The bar chart indicates the results of a sequential recognition strategy, whereby the MAP solution is chosen at each of the last 15 iterations. The chart compares the results with and without feedback.

FIGURE 6.3. Recognition results for the potato-head and the alarm.

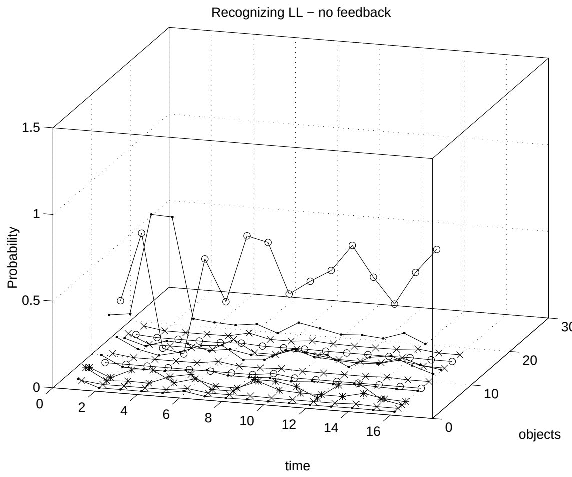
The results give indication as to the difficulty in recognizing instances of articulated parts of a complex object with only partial information available. In fact, with the compounded effects of self-occlusion, most parts had very little exposed surface from individual viewpoints. In fact, even when data were gathered from all around the object surface, many parts were significantly embedded within others, causing the fitting process to attempt to estimate a model with very little data to constrain it. In addition, many parts in the database were very close in parameter space (e.g. the eyes of the potato-head resembled the smaller sphere, the nose and the ears.) As a result, their distributions overlapped significantly, making it difficult to distinguish between them using the current modeling scheme. The difficulty of the problem leads to somewhat weak results in the MAP solution at individual viewpoints when no feedback was applied.

The more interesting result to examine is the effect of including feedback at each stage of the incremental recognition strategy. The results in Figure 6.3 indicate that by propagating evidence in the form of probabilities, the system is able to converge to the correct hypothesis in all but a few cases. This result illustrates the strength of the method, even with the compounded uncertainties at each stage of the bottom-up process, and with the sheer difficulty of this experiment due to self-occlusion and similarity of parts.

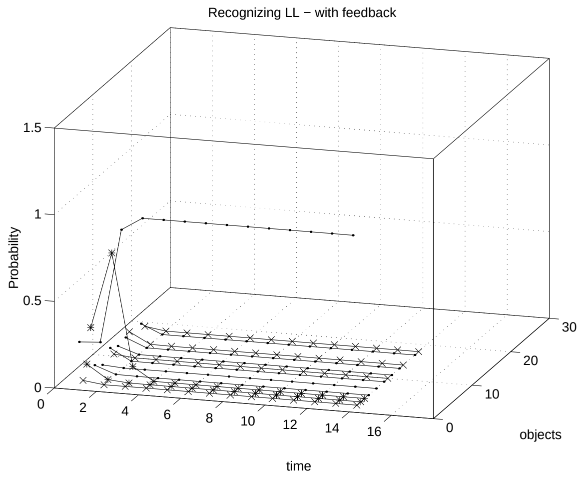
However, displaying the MAP results only tells a small part of the story. It indicates who the winner is, yet nothing regarding the competing hypotheses. The robustness of the approach can be conveyed more informatively by examining the probability distributions for different objects over time. In Figure 6.4, one can see the probabilistic recognition results in the cases of the left leg (LL) of the alarm clock, and the left ear (ERL) of the potato-head with and without the application of feedback at each iteration. By examination of the resulting plots, one can see the dramatic stabilizing effect of applying probabilistic feedback – without it, the beliefs in the various hypotheses fluctuate wildly. However, by cascading the posteriors as priors for each viewpoint, a clear winner emerges after very few iterations. All other hypotheses vanish quickly. One can see that using this method, an active agent can gather evidence in this fashion until a composite belief associated with a particular hypothesis exceeds a predefined level, or when only one hypothesis is left. Using the described strategy, the system would stop after very few iterations and choose the correct winner.

1.2.1. *Comparison with Previous Approaches.* In order to illustrate the benefits of propagating information using the method presented in this thesis, the strategy was compared to an accumulation strategy introduced in previous versions of this work [**3**]. Here, a voting strategy was employed whereby the unnormalized beliefs were binarized and histogrammed using empirically determined thresholds. The application was to recognize the same set of 3D parametric models (i.e. thresholds of 0.00001 and 0.0001 for the potato-head and alarm clocks respectively were used). In this section, a series of experiments are described whereby the two approaches are compared.

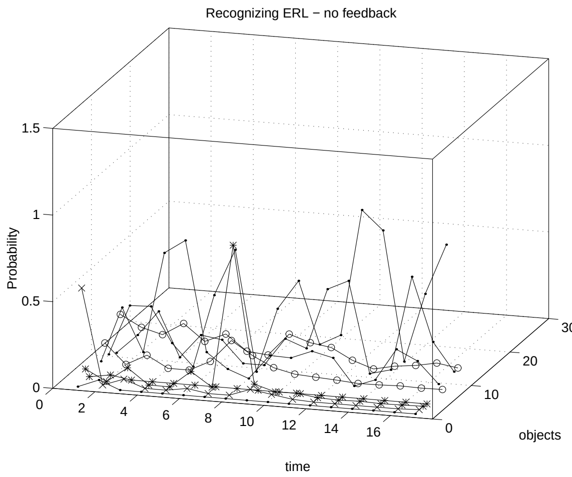
For both recognition strategies, training proceeded as usual. Here, 10 on-line recognition experiments on the potato-head and alarm clock were performed using the two different approaches. Both systems were permitted to iterate over a fixed, long interval to ensure



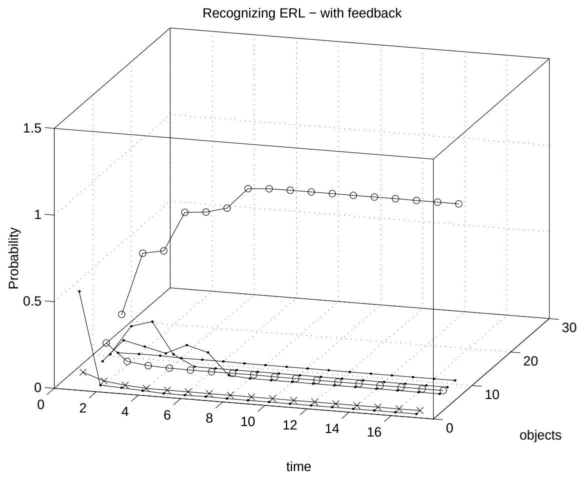
(a) Recognizing LL: without feedback



(b) Recognizing LL: with feedback



(c) Recognizing ERL: without feedback



(d) Recognizing ERL: with feedback

FIGURE 6.4. Results for the left leg (LL) of the alarm clock (a–b) and the left ear (ERL) of the potato-head (c–d) over time.

convergence in both cases in terms of repetition of a denoted winner. In the case of the strategy presented in this thesis, the winner is chosen based on the MAP solution. In the voting strategy, the winner refers to the object with the highest number of votes.

Table 6.1 illustrates the results of both experiments in terms of percentage and speed of convergence to the correct solution. The results indicate that the current strategy arrives at the correct solution quicker and more often than the previous approach, which does not perform well in accordance with these metrics. Furthermore, the previous strategy is somewhat sensitive to the threshold chosen. Table 6.2 illustrates that these results can vary depending on the different thresholds chosen for each object. The current approach does not depend on any such thresholds.

Comparison Metric:	Bayesian Chaining	Binning and Voting
% Correct convergence	60	38
Avg number steps to convergence	1.87	3.34

1emAbove one can see the results of running 10 recognition experiments on the potato-head and alarm clock, using both the Bayesian chaining and voting schemes to gather evidence in the various models. The thresholds used are those found in the paper [**3**], i.e. 0.00001 for the potato-head and 0.0001 for the alarm clock. The table compares the two approaches using two important metrics: (1) percentage of times the system converges to the correct solution (in terms of a MAP result), and (2) average number of steps to achieve correct convergence. Notice that the current strategy is much more powerful than the previous one.

TABLE 6.1. Comparison of bayesian chaining and voting strategies.

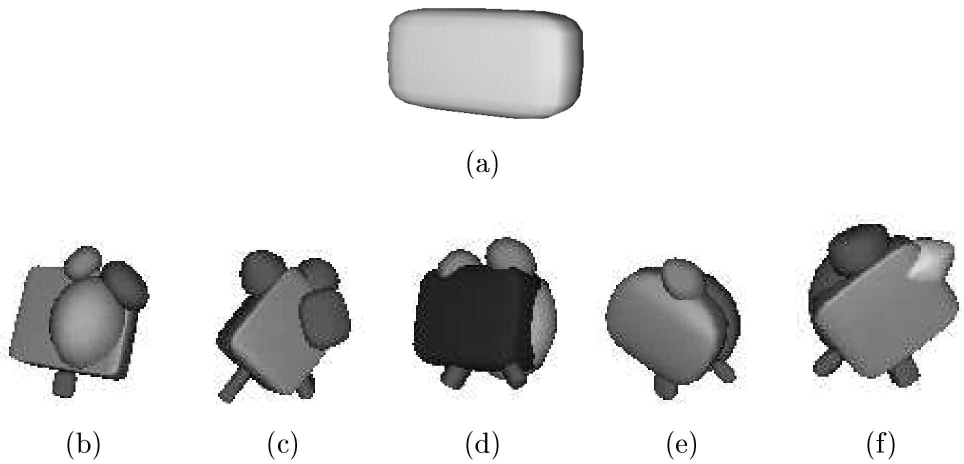
Thresholds:	pot:0.00001 alarm:0.0001	pot:0.00001 alarm:0.00001	pot:0.0001 alarm:0.0001
% Correct convergence	38	45	30
Avg number steps to convergence	3.34	3.68	3.85

TABLE 6.2. Voting strategy results using different thresholds.

Other metrics for determining convergence are possible for the strategy presented in this thesis. As statistical methods are employed, information theoretic techniques become available. For example, one possible stopping criterion can be when sufficiently low on-line entropy (in terms of a threshold) is attained. This is an accurate indication that the system has reached a conclusion with relative certainty. The voting strategy does not make use of statistical techniques and, therefore, cannot make use of such convergence metrics. As a result, one is forced to use strategies such as the one presented above, where convergence is measured as the repetition of a winner after many iterations of recognition. In this case, there is no formal way to determine the extent of the system's ambiguity in its assertions, i.e. an analogous notion to entropy, and therefore when true convergence is attained. Another strategy that uses a voting strategy to accumulate evidence can be found in [**84**].

1.2.2. *System Limitations.* Although the system works well in most cases, it would be interesting to investigate some of the system's limitations. Figure 6.3 illustrates two occasions whereby sequential recognition using feedback converged to an incorrect hypothesis. Here, it would seem as if *not* propagating evidence would be the better method of choice as, on occasion, it arrived at the correct MAP hypothesis. We will examine each of these cases in turn. In the first case, the system identified the right ear of the potato-head (ERR) as the alarm bell. Here, the extreme closeness in the part parameters in the database leads to difficulty in discrimination with only partial information available (both models can be seen in Figure 6.2(b)). This problem is quite common to any recognition system where objects are extremely close in parameter space, especially with the compounded effects of noise and uncertainty.

In the second case, the system matched the back of the alarm clock (B) with the block. This case illustrates the difficulty of the recognition task when problems occur at various stages of the bottom-up system. The first problem occurred at segmentation stage. As the back model was heavily embedded within the face model, very little surface area was exposed at any individual viewpoint. This lead to difficulties in establishing proper correspondence between range data and part patches. Next, a breakdown in the modeling assumption, namely that a model of the type chosen could be sufficiently constrained to fit the data, leads to models of the clock back that varied substantially from view to view. To see this, Figure 6.5(b)–(f) illustrates various examples of the models of the alarm clock as produced from varying viewpoints. For the most part, the model chosen for the back was square in shape causing it to resemble the block. It is important to recognize that this failure is not a fundamental flaw in the recognition strategy, but rather a breakdown in various lower level processing stages in the system for this part. Despite the difficulties in the alarm clock recognition task, the sequential strategy converged to the correct solution in all but one case, that of the back of the clock.



1emAbove, one can see the (a) reference model for the block and (b)–(f) alarm clock models as generated from the bottom-up system from different viewing positions. By examining the resulting models, one can see how the system converged on the hypothesis that the back part of the alarm clock was a block.

FIGURE 6.5. (a) Reference block model, (b)–(f) sample alarm models.

2. Case II: Recognition Based on Flow Images

The generalizability of the recognition framework is tested through its application to the problem of recognizing objects based on the optical flow images generated as the object moves with respect to a camera. For each of these experiments, the system was trained on a small set local motion descriptors uniformly sampled throughout the viewsphere. The first experiment tested the robustness of the system at identifying objects through motion sequences not trained on. Next, the system was tested on objects not trained on that were similar in shape to those in the database. This tested the system's ability to generalize based on shape. Finally, recognition was attempted based on human-induced movements. The general hypothesis for these experiments is that accumulating evidence on the level of probabilities should lead to a high belief in the correct model over a short number of views.

2.1. Experimental Setup. The first problem that needs to be addressed is how to gather the entire set of training images for each object in the database. In order to permit on-line recognition based on a wide set of motions, a complete set of flow images is required about each training object. Automation of the data acquisition process is a non-trivial task. Often strategies address this problem by rotating the object at fixed intervals by hand, grabbing images at each interval [**71**]. This becomes infeasible with large databases, especially if the entire viewsphere about each object is to be sampled. Furthermore, the system is bound to be uneven in is sampling, causing biases towards certain object viewpoints.

For these experiments, an *automated* control system was developed to permit precise gathering of training images from around the viewsphere of the object. The objects were placed on a rotary table at a fixed distance from a CCD (monochrome) camera affixed to the end-effector of a CRS gantry robot (i.e. the same robot that was used in the previous set of experiments). The distance was fixed in order to permit differentiation between objects with similar shape, but with varying size. The rotary table was used to access the longitudinal positions on the viewsphere about the object of interest. The robot arm was used to reach latitudinal positions, as well as to move along arc-like motion segments at each position. With this configuration in place, the top hemisphere of the viewsphere (i.e. latitudes $0°$ to $90°$) was reachable by the robot. In order to expose the underside of the object, the object was manually turned over in order to gather the images along the other half of the hemisphere. In this fashion, all positions along the latitude and longitude of the viewsphere were accurately reachable.

Figure 6.6 illustrates the experimental setup used. For the purposes of training, a black cloth was placed around the rotary table so as to focus the optical flow computations on the object of interest. Using this setup, the system was trained on a set of 25 household items such as cleaning products, fruit and toys (see Figure 6.7 for a group photo). Images of each of the objects can be found in Appendix C.

FIGURE 6.6. Experimental setup used for gathering flow images. Camera is anchored on the end of a gantry robot arm, and object placed on rotary table.

The next issue to be addressed is how to tessellate the viewsphere about the object. Any strategy that attained even (equal area) tessellation of the sphere would have been adequate. For the purposes of these experiments, the viewsphere was divided into coarse segments of equal area through a formulation that samples the sphere longitudinally at each latitude. Given a density of sampling at the equator, i.e. every 30 degrees in longitude, the question is how to sample each subsequent latitude to attain even discretization. The density of sampling needed to be gradually reduced from the equator to the poles. The following formulation was used determining the longitudinal angle to sample at each latitudinal level. Given a particular sampling angle at the equator, ang_{eq}, the longitudinal sampling angle at latitude, lat, denoted ang_{lat}, was computed as:

$$ang_{lat} = \frac{ang_{eq}}{cos(lat)}. \tag{6.1}$$

FIGURE 6.7. 25 database objects.

This was derived by enforcing the constraint that the sampling area remains constant at each sample. Using this sampling strategy, this resulted in 92 images gathered for each object. For a database of 25 objects, this amounted to 2300 images in total.

Once each sampling position on the viewsphere was computed, the next task was to compute a local motion basis at each position so as to permit recognition from a larger set. This involved choosing short, curvilinear, arc-like motion segments that simulated movement along a great arc about the object of interest. At each position on the viewsphere, the local basis was computed by moving the robot arm along an arc in the horizontal and vertical directions. The camera orientation was fixed to an upright position. Three images, grabbed along each arc path, were used to compute an optical flow image (of size 60×80). The hypothesis was that the representation method would successfully generalize the motion images that lie within this range. An illustration of a typical set of such arcs at a particular location on the viewsphere can be found in Figure 6.8. The illustration shows how the three images acquired along each arc-like trajectory were used to compute a flow image at that location. Two such motions lead to the storing of two flow magnitude images at each position on the viewsphere.

As was described in Section 1, the optical flow was instrumental in locating and centering the object of interest within the image (thus achieving position normalization). Temporal scaling was performed by normalization of the optical flow magnitudes, ensuring that

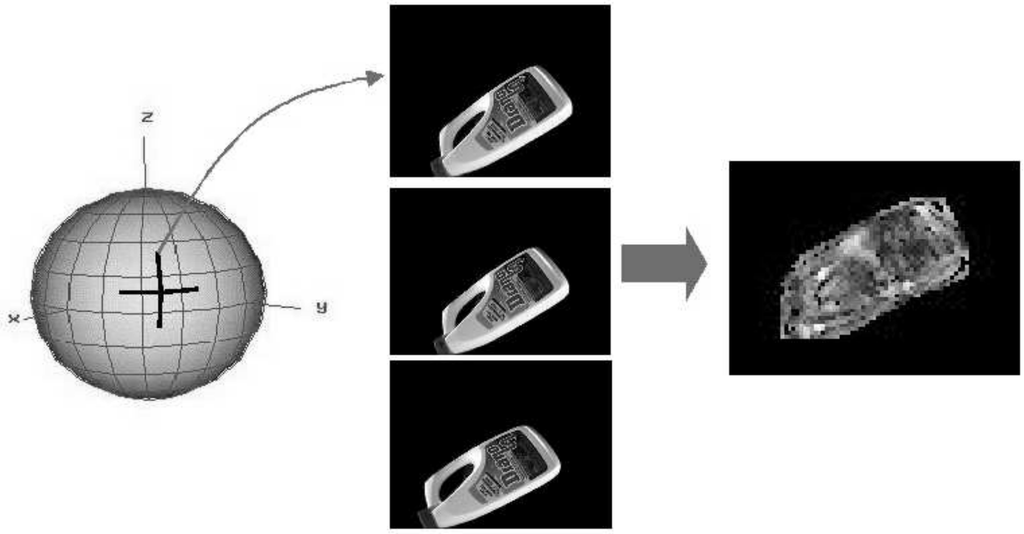


FIGURE 6.8. Flow image acquisition during training.

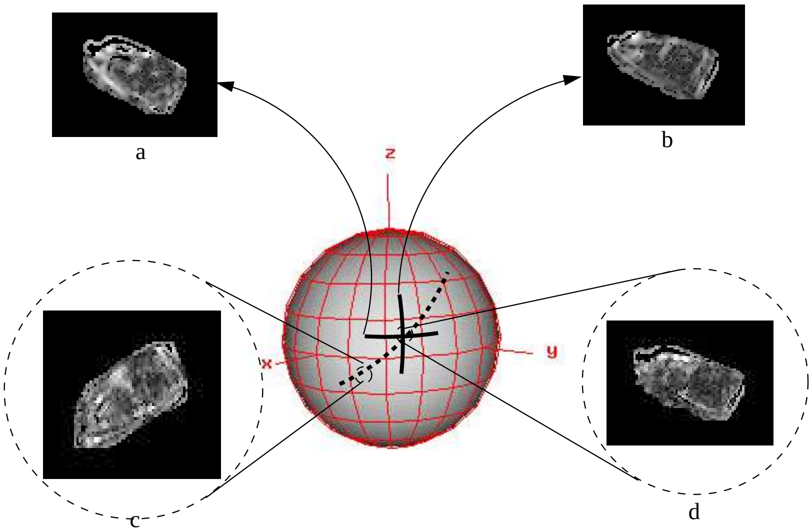
all values lay between 0 and 255. This ensured that speed did not affect the appearance of the 3D structure, as well as reducing quantization noise.

The entire set of optical flow magnitude images were parametrized using PCA techniques, where it was determined empirically that the top 20 eigenvectors were sufficient to represent the images. Low dimensional parametric flow descriptors were established by projecting each training flow image onto the resulting basis. Then, probability distributions were generated for each object in the database as described in Chapter 4.

With this system in place, a variety of experiments were performed. Each of the experiments and their results will be described next.

2.2. Experiments. Three separate tests were administered in order to test the generalizability of the strategy. The system was tested (1) on motion sequences outside the training set, (2) on objects not part of the training set, and (3) on movements generated by humans. The hypothesis was that the system would be able to identify the objects under these different conditions, and converge to the correct solution in most cases.

2.2.1. *Recognizing Objects from New Motion Sequences.* In the first set of experiments, the setup used to gather images was the same as the one used in the training procedure. For the purposes of testing the generalizability of the system, the images were gathered at locations on the viewsphere where no training images were taken, and along directions other than the ones used as a basis during learning. In order to generate flow field images, three images were gathered in sequence at fairly close distances from each other (at approximately 30 degree increments). Each interval, therefore, reflects a relatively small section of the viewsphere. An example illustrating a typical test sequence and the corresponding flow images of the liquid Drano bottle[1] can be seen in Figure 6.9(c)–(d). The figures illustrate how flow images computed along trajectories not trained on, i.e. in this case along a diagonal arc direction, should be close in appearance to one or both of the flow images that comprise the training basis at that location. This justified using a small orthogonal basis set for training.



1emHere one can see the tessellated viewsphere surrounding the object of interest (eg. a liquid Drano bottle). The perpendicular darker lines illustrates the arc-like trajectories comprising the motion basis for training at a particular location on the viewsphere. Examples of the resulting flow images at that location can be found in (a) and (b). The dotted lines indicate a typical test trajectory. The resulting flow images at 2 locations can be found in (c) and (d). Notice that the flow image in (d) resembles both (a) and (b) despite having been computed along a trajectory not trained on.

FIGURE 6.9. Motion basis and test trajectory along viewsphere. Liquid Drano bottle example: (a)–(b) Flow images along basis used for training. (c)–(d) Flow images along trajectory.

The results of applying the sequential recognition strategy to each of the objects in the database can be found in Figure 6.10. Here, we compare the results of applying probabilistic feedback at each iteration to the results obtained by independent measurements. In the first set of experiments, the system moved along 100 random trajectories[2] about each object of interest, accumulating evidence using probabilistic feedback, until a suitable level of convergence was attained. In this case, convergence was measured by an entropy level below

[1]This refers to a trademarked product, Drano^{TM}.

[2]Random trajectories refer to paths determined by a random-walk navigation strategy where tiles are visited only once, i.e. non-repeating.

0.01. The maximum a posteriori (MAP) solutions attained at convergence of each iteration were tabulated. For the second set of experiments, the system gathered 100 random single view data sets about each object. At each iteration, the system performed recognition and stored the MAP result. For both sets of experiments, the percentage of correct recognition identifications was tabulated. The barchart illustrates a comparison of the two strategies for all 25 objects in the database. The list of abbreviated names for the objects in the database can be found in Appendix B.

The results verify the hypothesis that the system is able to recognize objects based on the optical flow images generated by their motions relative to a camera. This indicates that the general recognition framework is applicable to a wide variety of contexts: from recognition based on 3D models to recognition based on flow images. In fact, recognition based on flow is particularly challenging due to noise and confounding information in the image. Furthermore, the results illustrate the plausibility of the hypothesis that by training on a small set of characteristic motions, one can recognize objects based on a wider set of movements. The system was able to identify each of the objects in the database based on previously unseen motion sequences. For the random set of images tested on, the MAP solutions resulting from experiments without feedback were quite good. The solution had occasional trouble where the appearance of the flow fields of different objects were similar.

It is interesting to note that the system succeeded in identifying the unknown object in cases where nearest neighbour search, a method commonly used in appearance-based recognition, failed. This was true even in cases where no feedback was applied. In fact, a comparison of the two approaches indicated that the strategy of choosing a MAP solution at single viewpoints using the probabilistic method without feedback had a win of 65% over the nearest neighbour approach. The reason for this is that by building a probability distribution for each object during training, the appearance manifold is warped in accordance with the class covariances, providing a more accurate representation of distances in eigenspace. An additional win of the probabilistic strategy is, of course, the added ability to use the uncertainty in the resulting distribution to determine the quality of the results from a particular viewpoint.

The more interesting result to examine is the effect of propagating the confidence in the various models from one iteration to another. The sequential recognition results validate the hypothesis that by accumulating evidence in the form of the beliefs, performance would improved considerably. In fact, the majority of the cases lead to convergence to perfect

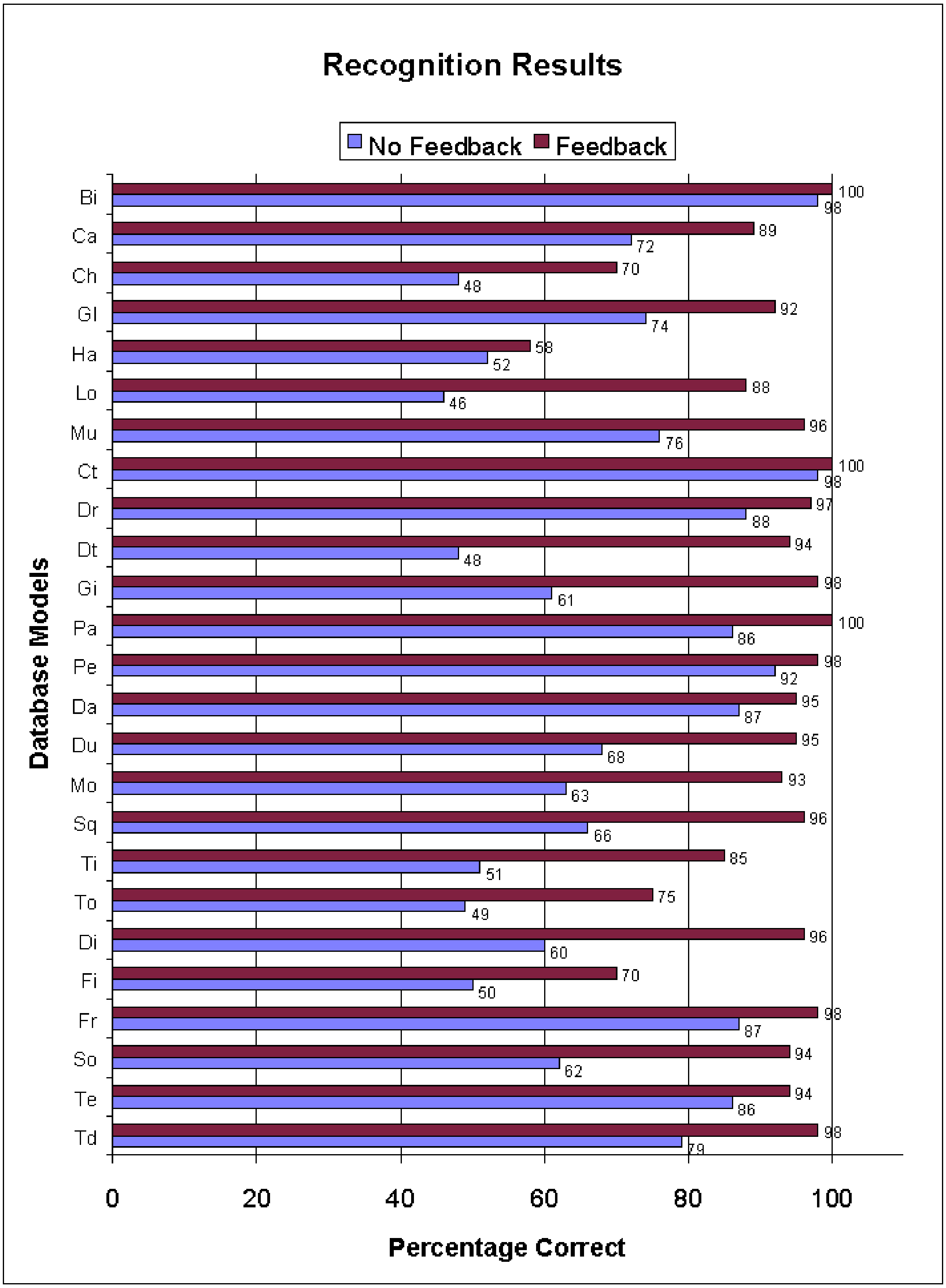


FIGURE 6.10. Recognition results for the appearance flows.

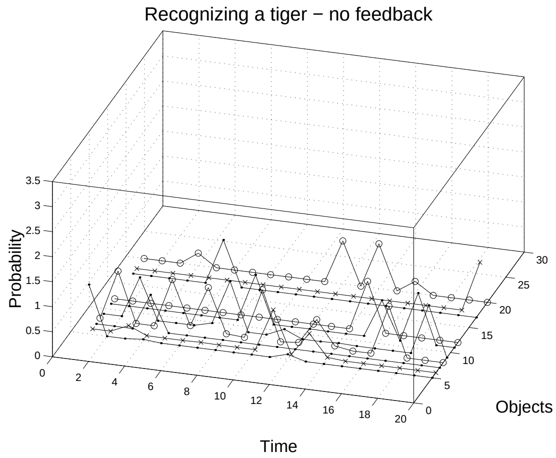
(or near perfect) results after a short number of views. This held true in cases where, without propagation, the MAP solution fluctuated wildly from one object to another. This shows that by inexpensively propagating information at the level of the probabilities, the recognition system is provided a powerful constraint which permits it to quickly converge to the correct solution.

The power of the sequential recognition strategy is best illustrated through example. Figure 6.11 shows the results of recognition experiments on two objects: a toy tiger and a toy fish. Here, one can see examples of results of experiments that either used or did not use probabilistic feedback from previous viewpoints. In both cases, the same random data sets for each object were used. The figure shows the probabilities in all the objects in the database plotted over time. The results show the system fluctuating wildly between the object hypotheses in the cases without feedback. Whereas when feedback was applied, the system converged to the correct solution, quickly eliminating most competing hypotheses. In fact, for all objects, the system converged to a single, correct solution in a very short number of views, in most cases. This can be seen in Figure 6.12, where the average number of steps (i.e. a step refers to a viewpoint visited) to convergence for all the objects in the database is plotted. Some cases were particularly difficult, as can be demonstrated in the case of the fish, in the example of Figure 6.11(c)). Here, the system took 9 iterations to converge to the correct identification. This is because in the previous iterations, strong evidence in the tiger model lead to a fluctuation between the two objects.

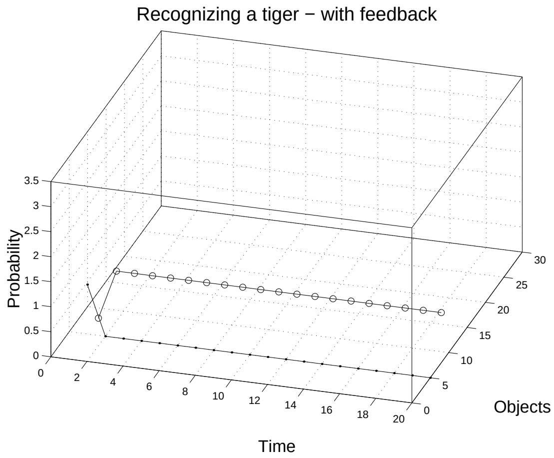
2.2.2. *System Limitations.* By examining Figure 6.10, one can see that the system had trouble recognizing the hamburger toy (Ha). The trouble is due to the statistical indistinguishability of the hamburger and the duck model in the database in many cases. This is because, the distribution for the hamburger model in the database lay entirely within the duck's distribution with a similar mean. In Figure 6.13, one can see the distribution of the data for both objects in the database projected onto the 3D eigenspace. An example of a particular hamburger flow image to be recognized is shown by the axes. This case illustrates a very difficult discrimination task for which the current scheme had limited success.

2.2.3. *Recognizing Previously Unseen Objects.* In order to further test the generality of the approach, a series of recognition experiments was performed on objects that were previously unseen by the system. These objects were similar in size and shape to objects in the database, however they had different texture patterns on their surfaces. The purpose of the experiment was to test whether the system would converge to objects in the database that were of the same class type. Positive results would be an indication that the system is robust to minor changes in size and shape, and that recognition based on flow signatures is not very sensitive to variations in texture patterns.

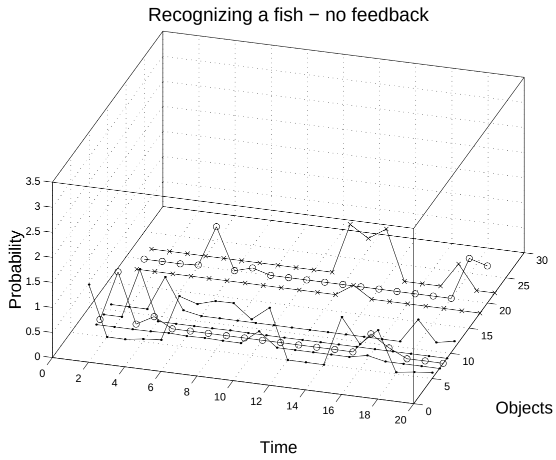
Specifically, the system was presented two objects: a tube of Crest toothpaste and a mug, different from the one in the training set. These items can be seen in Figure 6.14(a). Recognition experiments were performed whereby the camera was moved along random



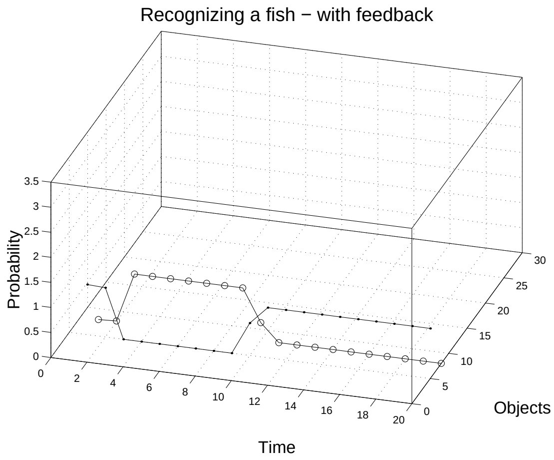
(a) Example 1: tiger - without feedback,



(b) Example 1: tiger - with feedback.



(c) Example 2: fish - without feedback,



(d) Example 2: fish - with feedback.

1emAbove, one can see the results of recognition experiments on two objects (toys): a tiger and a fish. (a) and (c) show the results without feedback, and (b) and (d) show the results with feedback. The y-axis shows the probabilities (ranging from 0 to 1 on an axis that reaches 3.5), the x-axis indicates the iteration in time, and the z-axis shows the corresponding objects numbered from $1 - 25$. Only those objects that have probabilities over 0.1 are plotted. Notice that the cases without feedback fluctuate wildly between the object hypotheses. In general, the cases with feedback converge to the correct solution fairly quickly. This is illustrated in the first example (although convergence occured in 9 iterations in the the second case).

FIGURE 6.11. Examples of recognition results with and without probabilistic feedback.

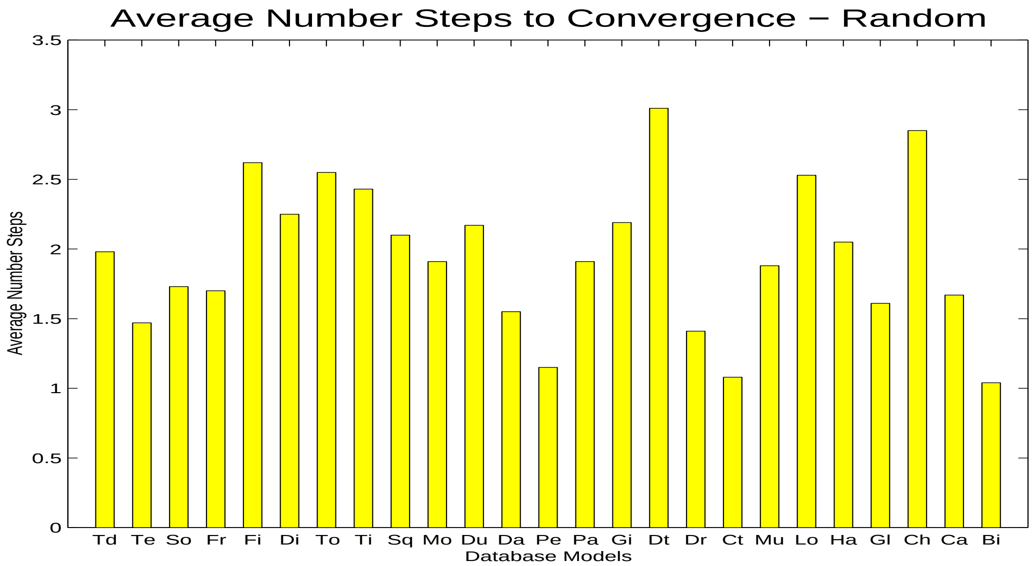
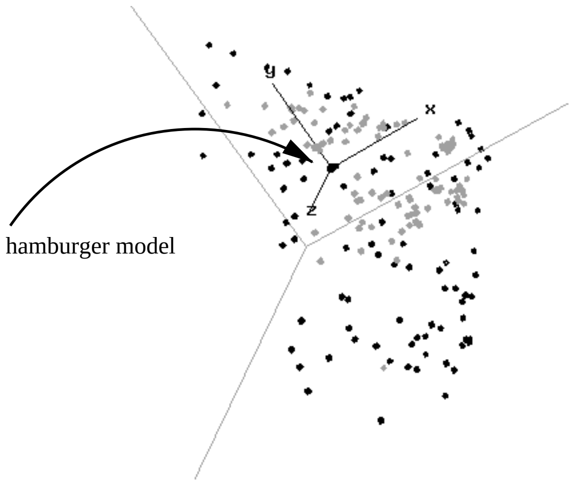


FIGURE 6.12. Average number of steps to convergence.



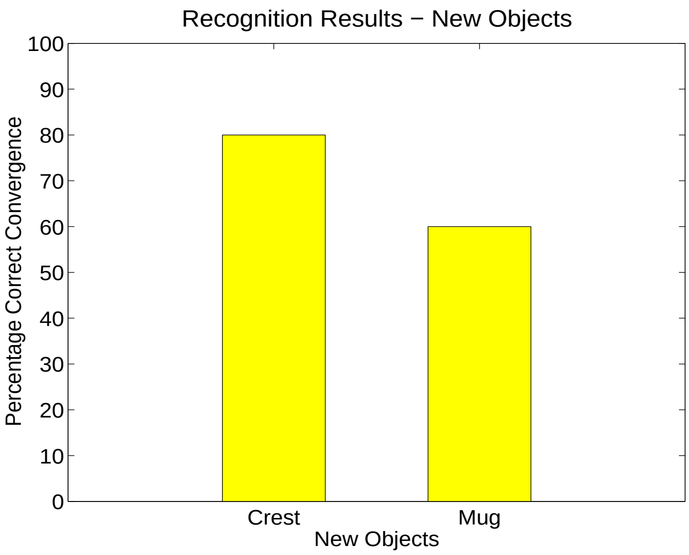
1emAbove are the data distributions for the hamburger (lighter color) and the duck models in the database projected onto 3D eigenspace. Notice that the two distributions overlap significantly, rendering the recognition task quite difficult. The sample hamburger can be seen as a point at the center of the xyz-frame.

FIGURE 6.13. The hamburger and duck distributions in the database, and a sample hamburger flow image projected onto 3D eigenspace.

trajectories, accumulating evidence in the objects until convergence. The goal was to for the system to converge to objects in the database similar to the test objects: the Colgate toothpaste tube and a mug (seen in Figure 6.14(b)). One hundred such tests were performed in order to observe the average performance of the system. The results can be seen in Figure 6.14(c), where the percentage of correct convergence results (in terms of MAP) are plotted for each test object.

(a) Test objects: Crest toothpaste tube and new mug.

(b) Identified database objects: Colgate toothpaste tube and mug.



(c) Recognition results over 100 trials.

Figure 6.14. Recognition results with previously unseen objects.

The results indicate that the system converged to the correct object type in the majority of the cases. In the case of the Crest tube, success occurred 80 percent of the time. The case of the mug proved more difficult, with a success rate of 60 percent. This is due to the differences in size between the test mug and the mug in the database. Appearance based techniques can only withstand a certain amount of variation in size before the projected components differ too substantially for any matching strategy to succeed. However, in terms of the generality of the approach, the results are quite promising. Further testing with a wider number of objects is currently being performed in the lab as the subject of a Master's project.

2.2.4. *Recognizing Objects from Human Motions.* The next level of experimentation tested the robustness of the approach in adapting to more realistic and more challenging environments. In this set of experiments, a human was asked to sweep an object in front of a camera with some arbitrary arc-like motion (within the permissible set), at roughly the same distance from the camera as in training. The added difficulty of this experiment lies in the motion consistency. The types of motion induced by humans vary substantially from the motion bases trained on using calibrated equipment. Further, the benefits of using flow to achieve figure/ground separation can be seen to the full extent, as both lighting and background varied substantially from training. In this experiment motion segmentation was also used to localize the flow to a specific window.

Initial results of attempting recognition by moving several objects by hand were quite promising. Two examples of a human moving the panda in front of a camera are illustrated in Figure 6.15. A greyscale image frame from the first motion sequence can be seen in Figure 6.15(a). Beneath this image, the recognition results in both experiments were plotted for the cases with and without probabilistic feedback. The probabilities were plotted over time for the top competing objects. Objects with negligible probabilities over the entire time series were omitted for clarity.

The results of attempts at identifying the panda without making use of prior information can be found in Figures 6.15(b) and 6.15(d). In these plots, the posterior probabilities in competing objects in the database can be seen changing over time. Notice that the system had a high degree of confidence alternately in each of the competing objects over time, with no clear winner at the end of the 20 iteration sequence. The second example was particularly challenging, with the system fluctuating between 9 different object hypotheses.

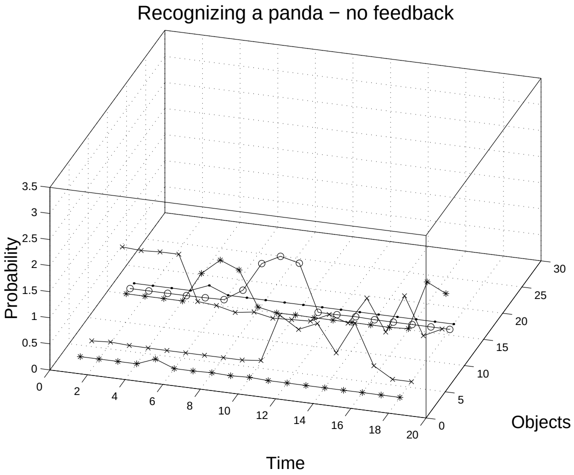
The effects of accumulating evidence over time using the same motion sequence can be seen in Figures 6.15(c) and 6.15(e). In Figure 6.15(c), the system converged to the correct solution in one iteration and remained there throughout the experiment (with a probability of close to 1.0). Figure 6.15(e) presents the results with the most complicated motion sequence. In the first 10 iterations, the system exhibited a high degree of confidence in two different objects other than the panda (represented by the first two curves). As the evidence in the panda grew, the system developed a high belief in the panda from iteration 10 onwards, and remained confident in the correct assertion throughout the experiment. All other objects were eliminated from its set of hypotheses. Results with several other motion sequences produced constant high beliefs in the panda after one or two iterations when feedback was included, with similar fluctuation patterns in the cases without feedback.

3. Identifying Informative Views

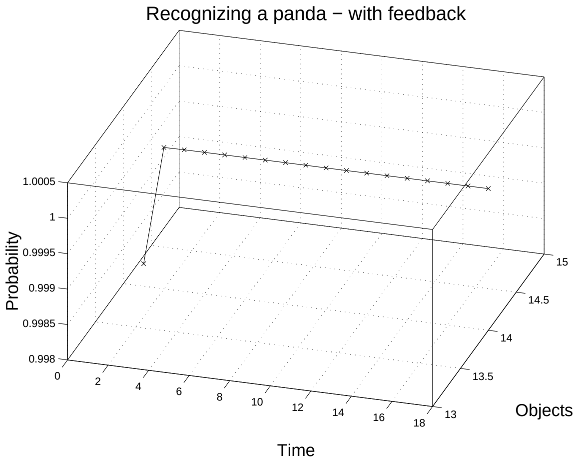
An important bi-product of the recognition framework is ability to define a measure of the degree to which a viewpoint is informative or not, for the purpose of identification of the unknown object. There are several strategies for accomplishing this goal. As the recognition results are in the form of probability density functions, information theoretic techniques become available. In fact, the entropy of the distribution was tested as a possible measure of recognition ambiguity, leading to a automatic strategy for determining the informativeness of the viewpoint.

Through a series of tests, it was confirmed that low entropy is a good indicator of an informative view, and high entropy, a good indicator of an uninformative view. Figure 6.16 shows examples of an informative and uninformative view of a dinosaur doll. Figures 6.16(a)–(b) illustrate the raw greyscale images from low and high entropy viewpoints respectively. Figures 6.16(c)–(d) show the corresponding posterior distributions resulting from the two recognition experiments. Notice that the low entropy viewpoint leads to a clear winner in (c), whereas it is difficult to denote a winner in the high entropy situation in (d). From this viewpoint, the system is confused as to whether the object is a dinosaur (Di) or a tiger (Ti). Figure 6.17 illustrates the cause for the confusion by showing an example of the tiger doll seen from the same camera view, relative to the reference pose, that was uninformative for the recognition of the dinosaur. Notice that in this case, it would seem that the dorsal ridge of the dinosaur is the feature that the system requires for discrimination.

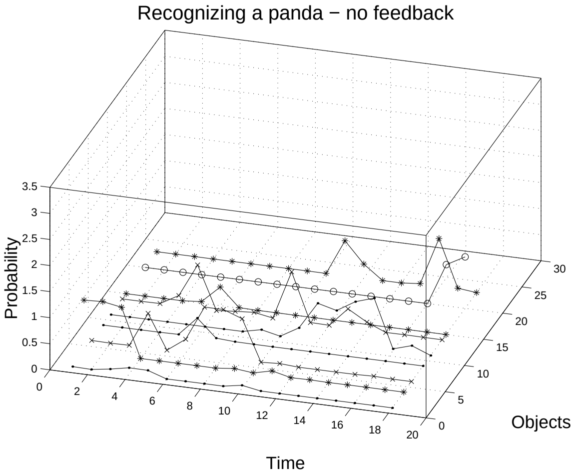
(a) Panda bear moved in front of the camera by a human.



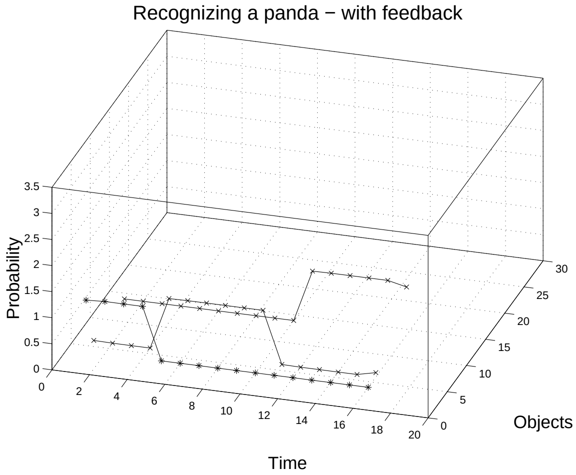
(b) Example 1: panda - without feedback



(c) Example 1: panda - with feedback



(d) Example 2: panda - without feedback



(e) Example 2: panda - with feedback

1emAbove, one can see the results of two sets of experiments whereby a panda bear toy was moved in front of a camera by a human. In (a), one can see an example image from the first motion sequence. Beneath this, the results of both experiments are plotted without the use of probabilistic feedback in (b) and (d), and with feedback in (c) and (e). Notice that the probabilities fluctuate wildly in the cases where no feedback was used. The system converged to the correct solution when evidence was sequentially accumulated. In the second and more difficult case, this took 10 iterations.

FIGURE 6.15. Results of recognizing a panda moved by hand.

These results lead to the possibility of using entropy to determine whether it is possible to choose a winner from that particular viewpoint at all. If several competing hypotheses have similar probabilities, a high entropy value would indicate that the MAP result, for example, would be a misleading indication of object identity. In addition, in a sequential recognition context, a high entropy results is an indication that further sampling is required.

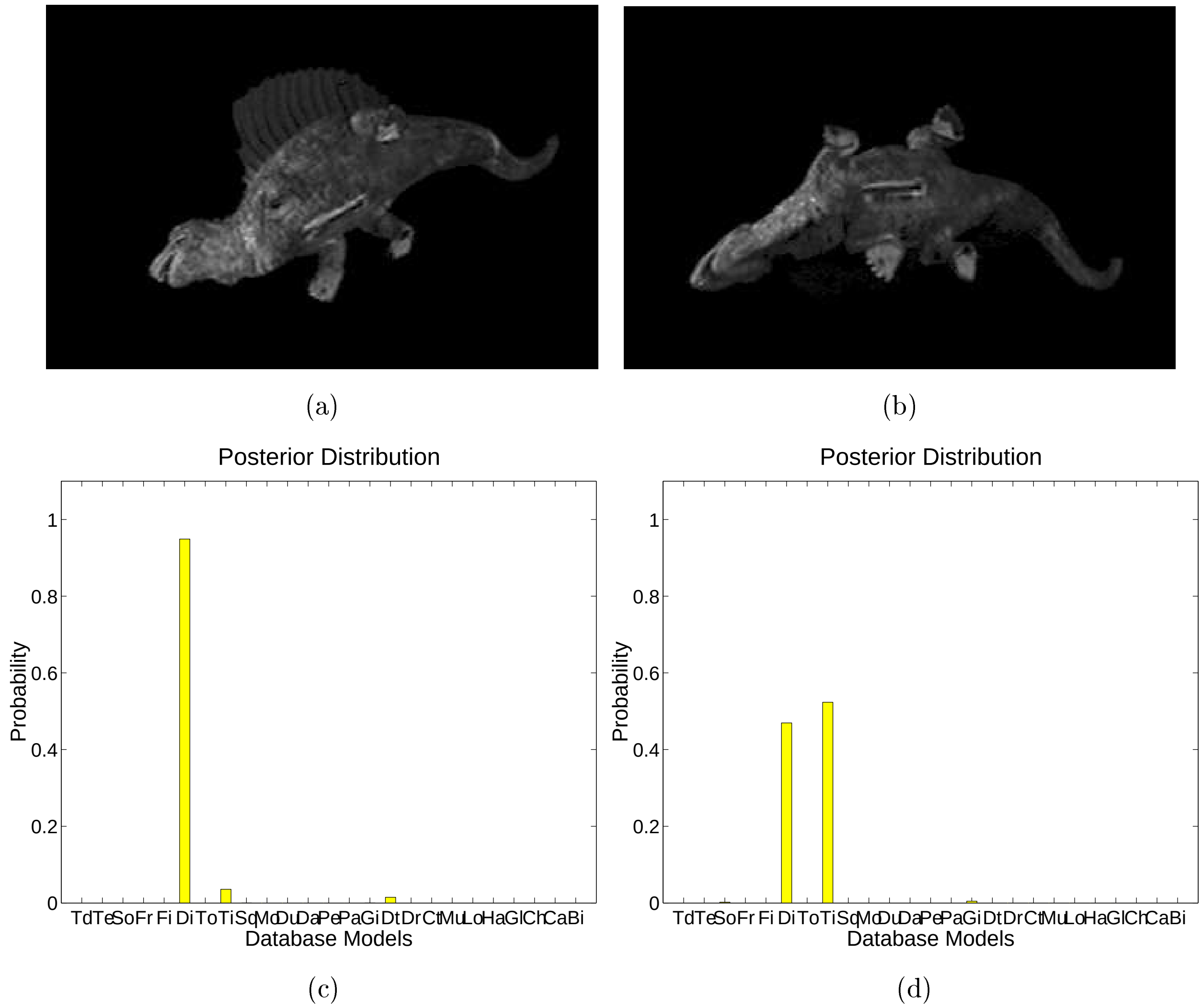
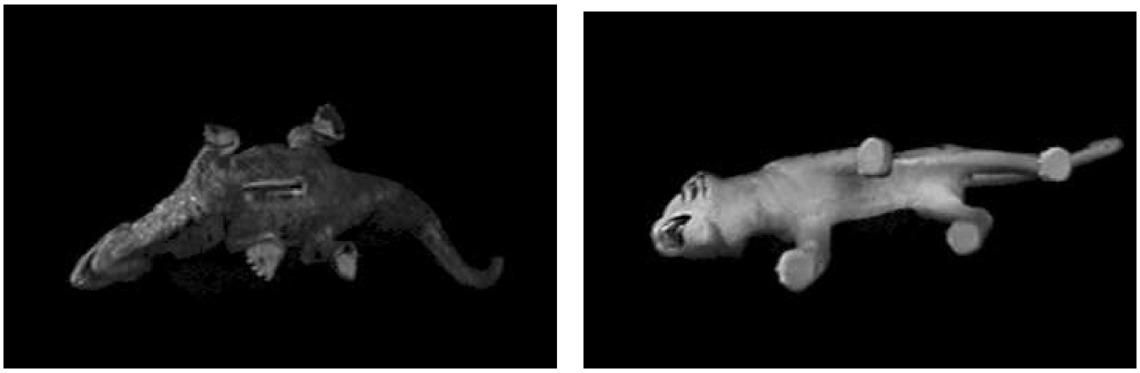


FIGURE 6.16. (a) Informative and (b) uninformative views of dinosaur (Di). Corresponding distributions with entropy values of (c) 0.24 and (d) 0.72 respectively.

The next chapter will confirm that every viewpoint in the entire viewsphere about each object can be labelled in accordance with its entropy *a priori* during training. In this



(a) (b)

FIGURE 6.17. (a) Dinosaur and (b) tiger dolls seen from the same viewpoint.

fashion, the most informative viewpoint for each object in the database will be located *a priori* and used during on-line navigation experiments.

4. Summary

In this chapter, an empirical justification of the theoretical framework for recognition described in earlier chapters was presented. The strategy proved to be versatile in its applicability to two very different object recognition tasks on either end of the spectrum: one based on 3D models generated from laser range-finder data and the other entirely appearance-based, with a novel twist in that motion images, rather than greyscale images, were used as inputs to the recognition engine. In both contexts, the compounded effect of noise and uncertainty lead to ambiguous and erroneous recognition results based on single views. The strategy of sequentially updating the beliefs in the various hypotheses was shown to resolve each of these ambiguous tasks, leading to a correct identification in most cases (in terms of maximum a posteriori results) after a relatively short number of views.

In the case of recognition based on flow images, the experimental results illustrate the plausibility of the strategy at successfully identifying unknown objects through a wide range of motions, generated from a different sources - either machine generated or human-induced, as well as under varying environmental conditions. This illustrated the robustness of the approach. Further, the system was shown to be capable of recognizing objects in the same shape class, but not identical to, objects that were trained on. Based on these results, further, more rigorous, testing of the extendibility of the approach becomes feasible.

One improvement to the current strategy would involve training based on different camera orientations in the image plane. In the current training system, the camera was fixed to an upright position while acquiring images. This leaves an additional degree of freedom in the rotation in the image plane (i.e. the optical axis). There are several possible strategies for overcoming this constraint. One strategy would be to train based on all possible stable resting positions for each object in the database. As most objects only have a small, finite set of stable positions, accounting for all positions should be feasible. Alternately, all possible orientations in the image plane could be artificially generated for each image acquired from single-pose training. Either strategy would result in a higher dimensional viewsphere. The composite set of acquired images would then be used to build the appearance manifold[3]. An alternate strategy would be to train based on a single camera orientation, but to artificially generate all possible image plane rotations of the image of the unknown object in the scene. One interesting solution provided by Campbell and Flynn [**24**] was to find a canonical transformation (i.e. planar rotation) that aligned the major and minor axes of each training image acquired on a viewsphere. This worked well with elongated objects.

It is worth emphasizing that, for the purposes of testing the recognition strategy, two completely different *real vision systems* were built. Each approach interacted with the physical world through sensors that acquired measurements of real objects. From these measurements, two different vision systems lead to object descriptors through interactions with hardware modules (cameras, robots, etc.) as well as software modules that processed the information at each level of inference. In each case, the unpredictability and noise of the environment, as well as equipment and software errors, rendered the task much more difficult than if simulations (a popular testing vehicle) were used. Furthermore, in the particular case of recognition based on flow, an entire control system was built from scratch for the purpose of testing the algorithm. Testing the strategy on two vision systems such as these is novel in that no other strategy in the literature goes to the same lengths to justify its approach. This adds to the contribution of the work.

Although sequential recognition performance based on random samples is quite good, further processing is required for cases where efficiency in terms of speed and reliability is imperative. In the next chapter, the strategy for optimally choosing the next gaze position is empirically tested for the case of recognition based on flow. The hypothesis is that the strategy will lead to faster solutions with an even higher success rates.

[3]It might be more constructive to build several manifolds in this case.

CHAPTER 7

Navigation Experiments and Results

Recognition results presented in the previous chapter indicate that the system is adept at sequentially recognizing objects when the camera (or alternately, the object) is moved along a random trajectory. However, this chapter will illustrate that random navigation can lead to incorrect assertions through the acquisition of data from locations that lead to poor recognition results. Furthermore, the system is not optimized to reach a solution in the shortest number of steps. For applications that require fast and reliable solutions, further optimization techniques are required.

In this chapter, the theory presented in Chapter 5 for planning the next best gaze position was tested through a set of experiments using a real vision system. For the purposes of illustration, the theory was applied to the context of recognition based on flow images. However, the strategy is general and can be applied to any other recognition context (recognition based on 3D parametric models, for example).

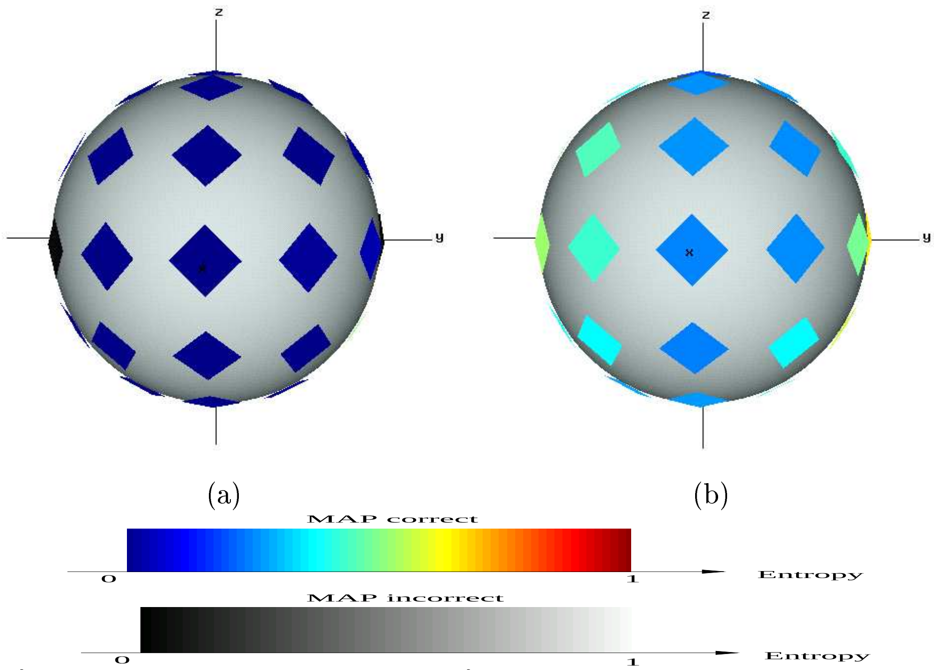
The goals of the experiments were: (1) to build an entropy map off-line for each of a series of objects in a database, and to examine its structure when using optical flow images as signatures for object shape, (2) to test the ability of using these maps to guide a sensor towards optimal locations in on-line recognition experiments. It will be shown that the structure of the entropy maps is conducive for navigation, and that active recognition experiments based on these maps indicate their superiority to random sequential recognition methods in terms of recognition accuracy and speed of convergence. Further, it will be shown that despite the difficulties in recognizing objects based on single flow images, the system is able to quickly converge to a correct solution by using Bayesian updating techniques.

1. Building Entropy Maps

In the previous chapter, the second set of experiments described an off-line training strategy whereby images of 25 household products were gathered at equally spaced locations around a coarsely tessellated viewsphere. For each object, two camera sweeps were performed at each position on the viewsphere. This lead to the acquisition of greyscale images, and then to the computation of two flow images at each location. With this framework in place, little additional effort was required in order to associate recognition ambiguity, in the form of entropy, to each viewsphere position as well. In this section, the structure of the resulting maps using real data will be investigated.

Entropy maps were built for each object in the database off-line during training using the strategy presented in Chapter 5, the hypothesis being that the structure of the maps lend themselves to on-line navigation based on them. Figure 7.1 shows an example of the actual (a) entropy map and (b) smoothed map generated for the glue bottle using the training data set. For illustration purposes, each tile in the entropy map in (a) indicates the lowest entropy value among both those generated (from horizontal and vertical motion sweeps) at that location. The smoothed map in (b) is generated by first averaging the entropy over all motion directions at each location. Then, the entropy is smoothed over the entire neighbourhood. Each tile is colored in accordance with the final averaged value in order to illustrate the best locations for recognition, on average. The crosses indicate the axes through the origin of the viewsphere in order to illustrate the position of the camera at each tile.

Notice that the viewsphere used for training was fairly coarsely tessellated, making it difficult to observe the structure of the map. In order to better observe the details in the variations in entropy values, entropy maps were computed for a more finely tessellated viewsphere. First, the viewpoint position was sampled more finely, i.e. at $15°$ intervals along the equator instead of $30°$ intervals. Using Equation 6.1, this resulted in in 186 segments for the viewsphere. In addition to finer position sampling, motion was sampled at higher density as well. At each location, motion sweeps were performed along 4 different curvilinear (locally, arc-like) directions: along $0°$ (horizontal), $45°$, $90°$ (vertical), and $135°$. In total, 744 flow images were acquired for each object in the database. Figure 7.2(a)–(d) illustrates the resulting entropy maps for each set of directions at the higher sampling resolution for the case of the glue bottle.

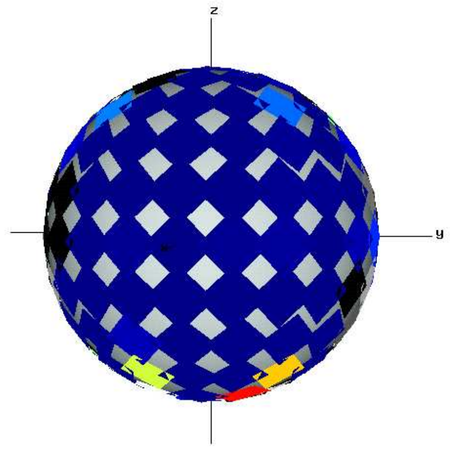


Images of the overall and smoothed entropy maps for the glue bottle. Recall that entropy ranges from low (blue) to high (red) if correct MAP results are attained. If recognition is incorrect, the entropy ranges from low (black) to high (white).

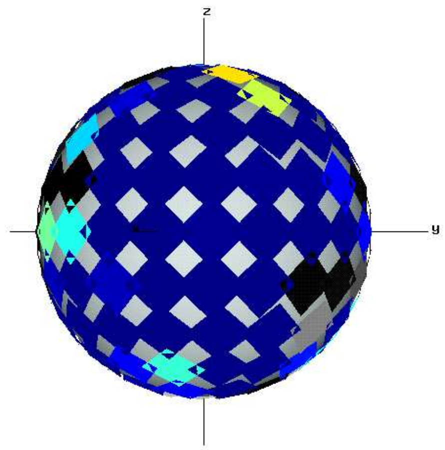
FIGURE 7.1. (a) Entropy map, (b) Smoothed entropy map using training data for glue bottle.

Figure 7.2(e) illustrates the overall entropy map seen from the same viewing position as the four others that were used to generate it. It illustrates the lowest entropy values (among all the motions) displayed at each location. In cases where recognition is correct in all directions, this is an indication of the *best* entropy value at that location. As well, in cases where some directions have correct and some incorrect recognition results, the lowest entropy value in the correct model is illustrated. However, when recognition is incorrect for all directions, the resulting tile illustrates the *worst* value at that location, i.e. the lowest entropy belief in the wrong model. The result is an slight exaggeration of the bad viewpoints for recognition in order to emphasize their locations.

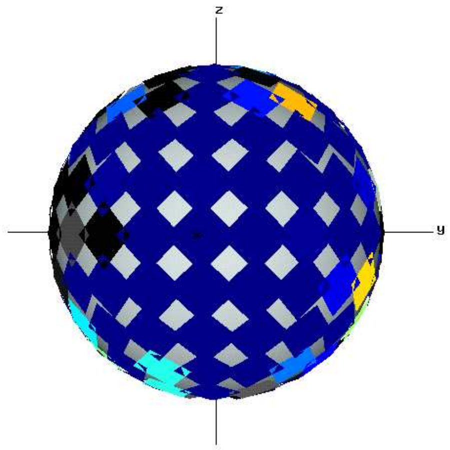
In order to examine the extent to which the structure of the map lends itself to navigation experiments, Figure 7.3 shows: (a) images of the glue bottle, (b) its corresponding overall entropy map (created as described above) and (c) its smoothed entropy map. Here, results from two different camera viewpoints are shown: one that the system identified as an informative viewpoint for recognition (in blue on the left smoothed map), and one that was determined as being the worst view (in red on the right map). The structure of the maps was such that the good locations were found in large patches, with the resulting map



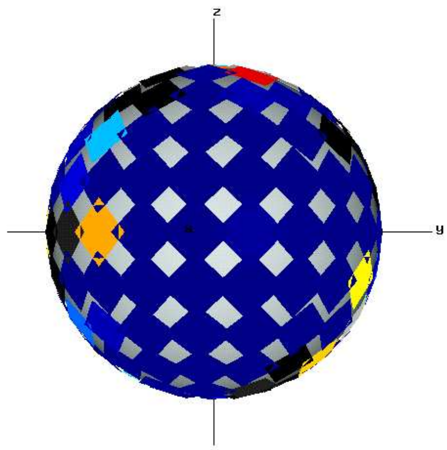
(a)



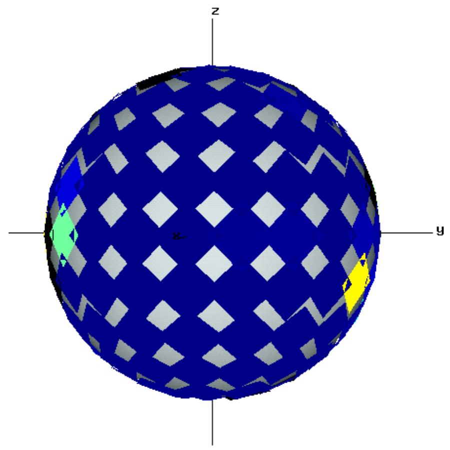
(b)



(c)



(d)



(e)

FIGURE 7.2. Glue bottle object entropy maps corresponding to four directions: (a) 0°, (b) 45°, (c) 90°, (d) 135°, (e) Final overall entropy map.

varying continuously. The optimal locations were chosen to be those at the center of these patches, so that navigation towards them should lead to correct recognition results, even if slight errors in positioning occur. Notice that the entropy maps match an intuitive notion of viewpoint ambiguity, as the system chose the viewpoint where the bottle is seen face-on as a relatively uninformative view. Here, very little object structure is visible. The side view, however, presents a much more distinctive viewpoint of the object shape.

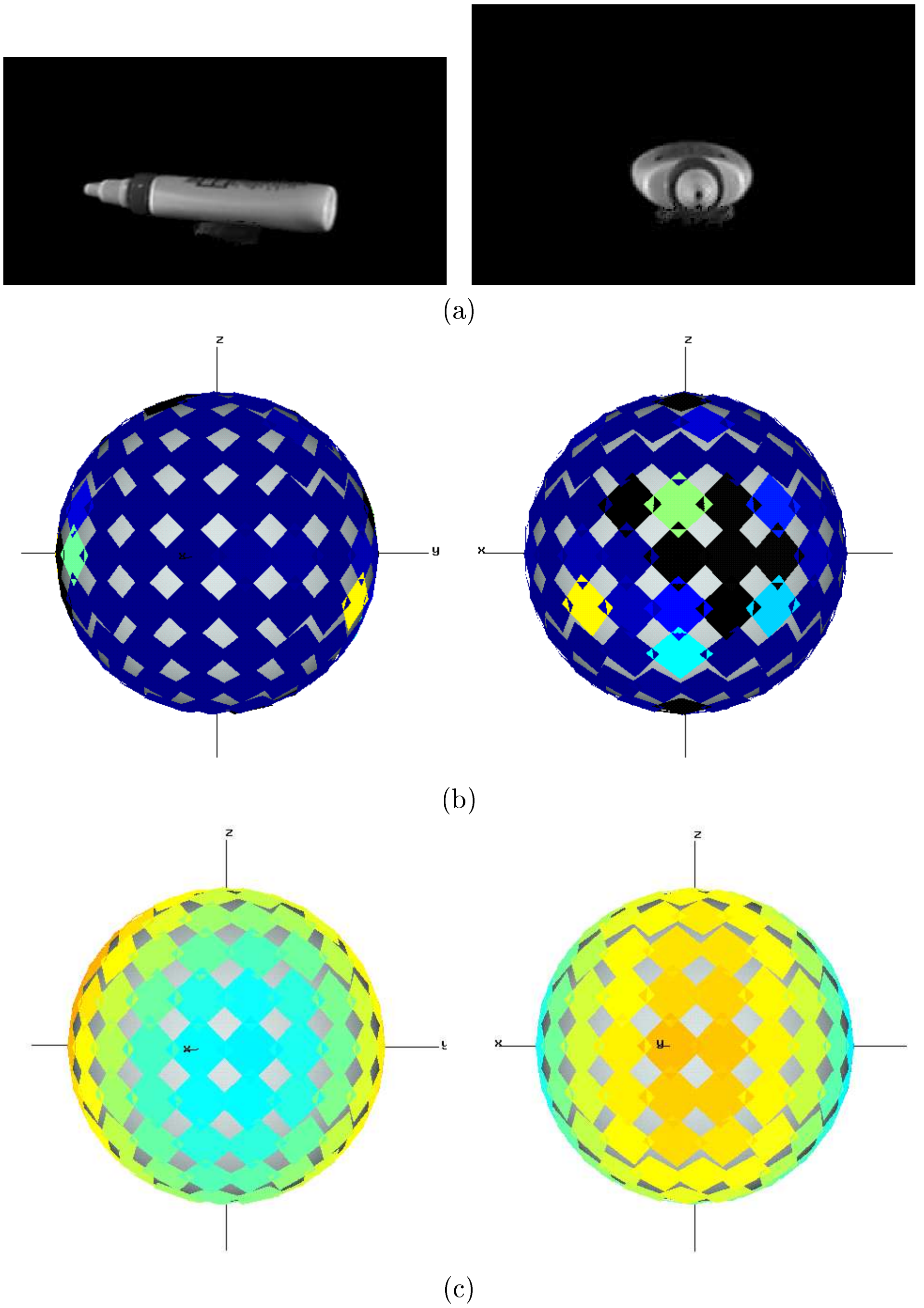
Figure 6.16 demonstrated an example of how entropy can be used to distinguish an informative view from an uninformative view in the case of the dinosaur toy. Figure 7.4 shows that these viewpoints can now be located *automatically* through the entropy maps. Here, the dinosaur toy and its corresponding smoothed entropy map are seen from the viewpoint that the system chose as the best (lowest entropy) on the left. The worst (highest entropy) view for recognition can be seen on the right. Previously, it was explained that the high entropy viewpoint corresponded to a viewpoint where the dinosaur closely resembled the tiger.

2. Navigation Experiments

Having computed the entropy maps off-line, a series of experiments were devised to test the on-line navigation strategy based on them. The hypothesis was that the maps would guide the sensor towards the optimal locations for recognition, and that convergence to the correct hypothesis would be more frequent than if one were to use a random sequential navigation strategy.

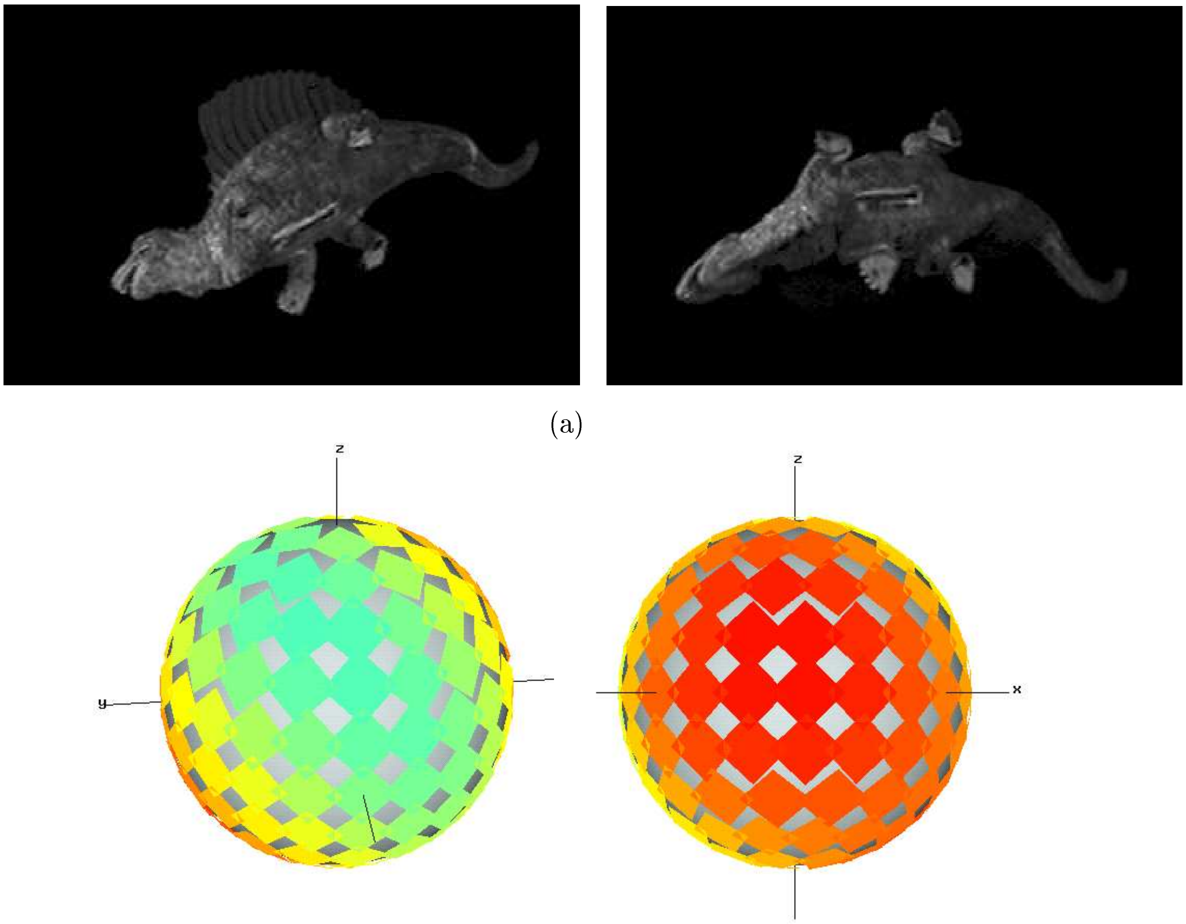
The first series of experiments employed the same physical set-up as in training for navigation. From an initial random viewpoint, the gantry arm was moved along a local, curvilinear trajectory on the viewsphere according to the proposed navigation strategy. At each coordinate sampled along this path, a local flow measurement was made by sweeping the arm along two short curvilinear arcs. Recognition was then performed using the corresponding optical flow generated by this local motion. The system iterated until the entropy reached an arbitrarily small convergence value (e.g. 0.01 was chosen).

The difficulty of performing robot navigation experiments on-line was due, in part, to issues of reachability. At any one time, only images on the top hemisphere of the viewsphere could be acquired. Moving the sensor to the lower hemisphere to acquire images caused the view of the object to be obstructed by the stage on which it rested. As a result, when the system requested to examine the unexposed part of the object, the user was prompted



Images of the glue bottle and the corresponding overall and smoothed entropy map are seen at two locations. The system chose the left view as the most informative (in blue, on smoothed map), and the right view as a relatively bad one (in red).

FIGURE 7.3. (a) Two views of the glue bottle, (b) Corresponding views of the entropy map, (c) Corresponding smoothed map.



(a)

(b)

Images of the dinosaur and the corresponding smoothed entropy map are seen at two locations. The system chose the left view as the most informative (in blue), and the right view as a relatively bad one (in red).

FIGURE 7.4. (a) Images of a dinosaur, (b) smoothed entropy maps at corresponding locations.

to "flip" the object, and navigation proceeded. This added further ambiguity to the pose estimation process, and slowed the system down.

In order to assure more accurate pose estimation, a set of object images that were not part of the training set (i.e. those attained through two additional motion sweeps at each training location on the viewsphere) were projected onto the appearance flow manifold. With very little extra work, this permitted a more densely sampled set of images to become available to the indexing routine, leading to a more accurate pose estimate.

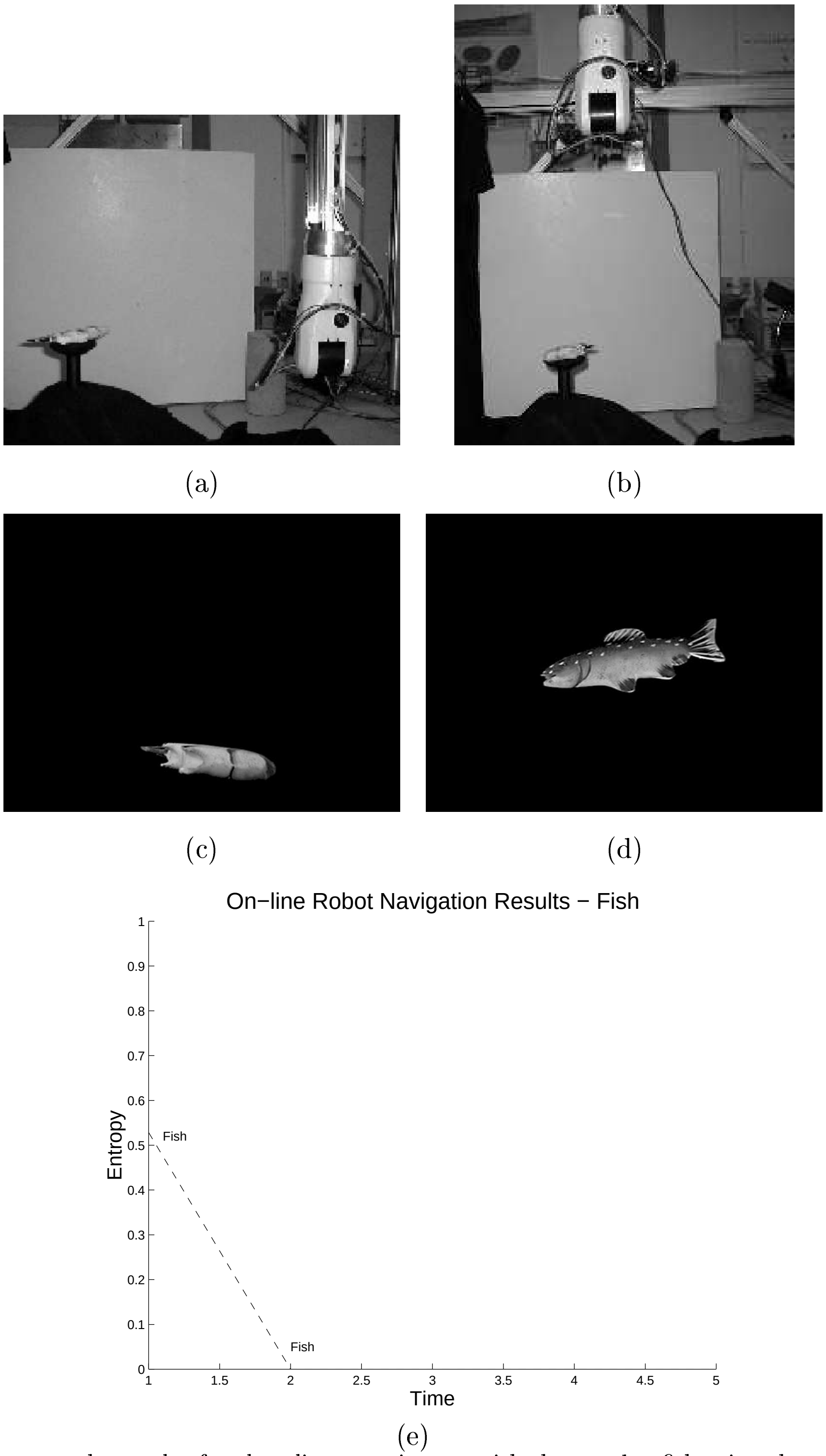
Using this set-up, empirical results indicate that the system converged to the correct solution in the majority of the cases, and in a relatively short number of iterations. Examples of results using on-line navigation can be found in Figures 7.5 and 7.6.

Figure 7.5 shows an example where the system navigated to the most discriminant position in a single iteration. In this case, the system began with a fairly ambiguous view of the wooden fish (i.e. entropy over 0.5). The initial gantry position and the image of the object acquired from that position can be seen in (a) and (c) respectively. The best viewpoint for the fish was at the south pole, where the entire shape of the object was visible. The system requested that the object be turned over, and moved the sensor to the exact location of the most discriminant view. The final gantry position as well at the image acquired from the viewpoint can be seen in (b) and (d) respectively. At the final location, the system converged to the correct iteration in one step. No errors in pose were reported despite the introduction of errors associated with human intervention. The convergence can be seen in (e) where on-line entropy is plotted over time. Several navigation results with the dinosaur can be found in Figure 7.6.

Although empirical evidence indicated that the system was working in on-line experiments, a large number of experiments needed to be performed on each object in order to quantify the system's overall average performance. Due to the slow speed of the particular optical flow algorithm chosen, as well as the relatively slow speed of the gantry robot movements (combined with the need to manually turn the object over), it was infeasible to run a large number of on-line tests on each object.

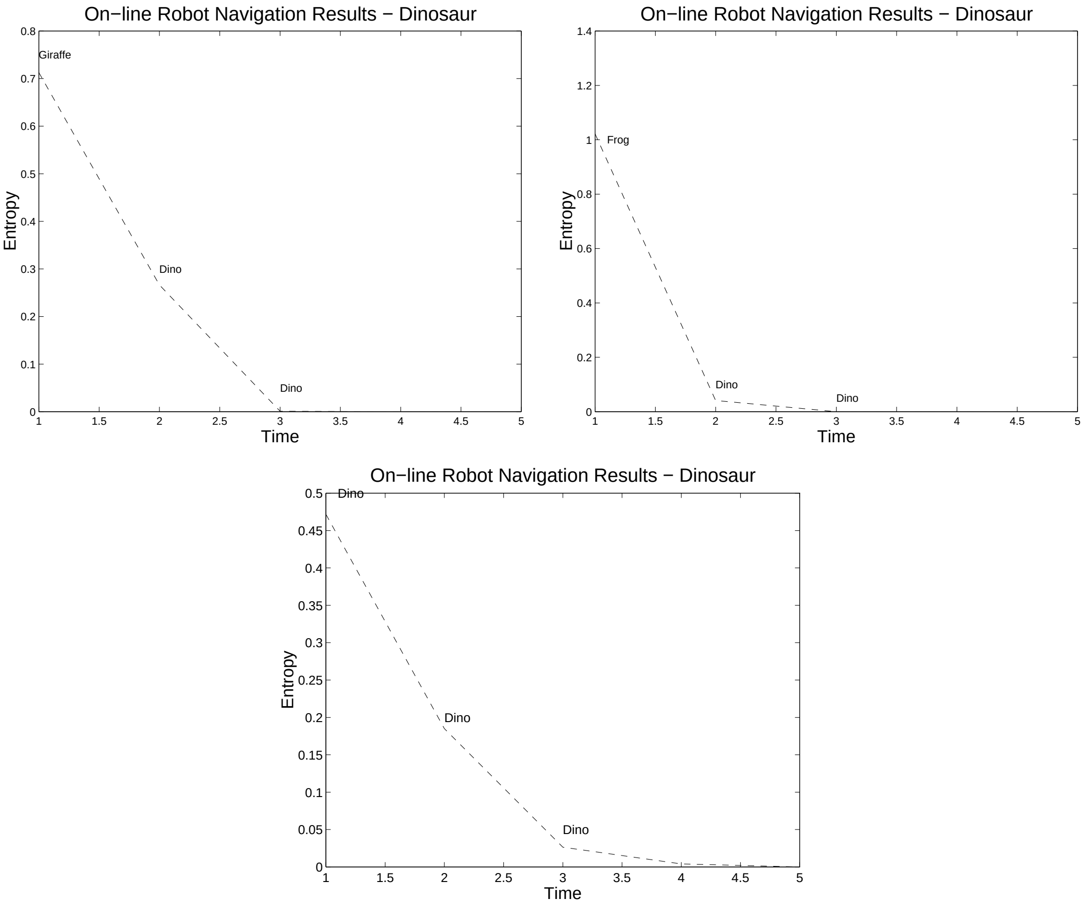
In order to run large scale experimentation, a series of on-line navigation experiments were performed using image data gathered by the gantry off-line. This required the system to first store a large set of flow images about each object in the database, which it would then be able to sample from during on-line navigation experiments. The advantage of this approach is that it permitted fast access to images throughout the entire viewsphere about each object. As was described in the previous section, a finely tessellated viewsphere about each object in the database was already gathered by the gantry robot for the purposes of illustrating the structure of the entropy maps. This set of 744 flow images for each object, along with the corresponding coordinates of acquisition, was therefore used as the sampling pool for on-line experiments.

Unfortunately, the sampling set drawn from for experiments included the images used for training, which is not usually desirable to use for recognition experiments. This set could not be excluded from the pool, in order to ensure that the system have free access to all locations on the sphere. As well, due to timing constraints, it was infeasible to gather a completely new data set for each object in the database. However, it must be re-emphasized



Above one can see the result of real on-line experiments with the wooden fish using the gantry robot. The first and final robot positions are seen in (a) and (b) respectively. The corresponding images acquired can be found in (c) and (d). The system takes the robot to the exact location of maximal discriminability in one step and converges to the correct solution (seen in (e)).

FIGURE 7.5. (a) First robot position, (b) Second robot position, (c) First camera view of fish, (d) Final camera view of fish, (e) Entropy over time.



1emAbove one can see the results of three on-line recognition experiments with the dinosaur toy. At each iteration, the curve is labelled with its MAP result. In the first two cases, the system began with a wrong hypothesis. In the third case, the system began from an ambiguous viewpoint. Notice that the system quickly converged to the correct solution in all three cases.

FIGURE 7.6. Results of several on-line navigation experiments with the dinosaur.

that attempting probabilistic recognition using training images as samples does not guarantee recognition success, as it does in the case of template matching and recognition based on look-up tables. This can be seen through examination of the entropy maps in the previous section, where, from several viewpoints, the system failed in successfully recognizing the object from sets of training images. This is due to the fact that comparisons are based on probability density functions of different objects, which may overlap significantly in some cases. As such, permitting the training set into the sampling pool should not bias the results.

With this sampling pool in place, the navigation strategy proceeded as follows in practice: During each iteration, the proposed navigation strategy performed recognition based on the current flow image, and then used indexing techniques to map its current location to the appropriate entropy map exactly as before. It should be emphasized that although the exact location information was available to the system, it was not used in order to simulate a real navigation experiment.

Next, the system extracted the best location and movement to perform next, and then computed the motion transform required to get it to that location. This transform was then applied in the current camera frame, leading to the next location and movement to perform. The appropriate flow image corresponding to that movement was then extracted from the stored set. If the flow image was not available, the closest image (in terms of proximity) was chosen. Although this limited navigation to a discrete set of locations, the fine tessellation of the viewsphere ensured that a relatively close flow image should be available. Using this strategy, repeated on-line experiments could be performed quickly.

2.1. Recognition Results Based on Perfect Pose Estimation. This thesis does not claim to provide a general solution for the problem of pose estimation, a difficult and open problem in the field of computer vision. Thus, in order to examine the workings of the proposed navigation strategy, the first set of experiments worked under the assumption that a perfect estimate of pose can be provided by an external pose estimation module. For this set of experiments, this module replaced the nearest neighbour approach described earlier. This permits fair assessment of the proposed navigation strategy without biases introduced by pose estimation errors.

In the first experiment, the system began from an arbitrarily chosen viewpoint, and was permitted to navigate autonomously, using the proposed navigation strategy, until convergence. Only those iterations that lead to convergence in more than one step (i.e. viewpoint) were considered, in order to fairly give the proposed navigation strategy a chance to be invoked. In order to examine average performance, this experiment was repeated until one hundred such experiments were completed for each object in the database. Figure 7.7 illustrates the percentage of cases where correct convergence occured. Notice that on average, the system converged to the correct solution in $80\% - 100\%$ of the cases.

A similar experiment to the one above was performed, this time using a slight variation of a random-walk navigation strategy. Here, a series of experiments were performed whereby the system was launched from the same starting viewpoints as in the previous experiment. The system was permitted to randomly navigate to previously unvisited locations, until convergence. A formal comparison of the random-walk and proposed strategy can be found

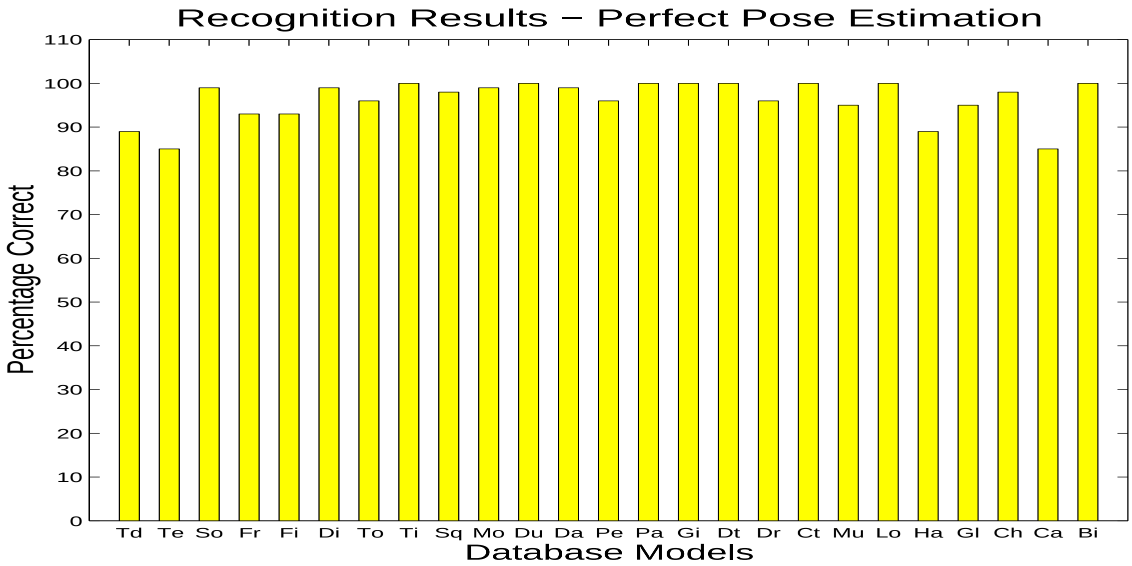


FIGURE 7.7. Percentage correct recognition results at convergence (Perfect pose estimation).

in Appendix D, Figure D.1, where the strategies were compared in terms of the percentage of iterations that lead to a correct convergence hypothesis. The results indicate that both strategies performed quite well, in terms of recognition results and quick convergence. This is mostly due to the strength of the Bayesian chaining algorithm at eliminating false hypotheses quickly. Still, the results indicate that the proposed approach out-performed the random case in most cases.

As any starting viewpoint was permissible, there were a few cases where the proposed algorithm started from a local minima (i.e. a low entropy belief in the wrong model), became stuck and converged to the wrong solution. In order to avoid such cases, one possible precautionary measure that could be considered would be to invoke an initialization procedure that would enable the system to recover when starting out from a local minimum in the entropy map. This would involve invoking the navigation strategy after moving to a *second* randomly chosen position. In previous versions of this work [**5**], it was found that this lead to an improvement in the results. However, this strategy added an extra step to the procedure. This is the tradeoff between recognition success and speed. Deciding which is more important is an operational consideration.

Empirically, it was found that navigating based on entropy maps outperformed the random approach in cases where the system started in high entropy (i.e entropy > 0.5) locations, with a MAP belief in the right model. Ambiguous views are caused by closeness in object appearance, leading to confusion between several hypotheses. The proposed approach would lead the system to the location that would reduce the ambiguity, i.e. lead to exposure of a unique feature, and lead to relatively quick convergence. A random navigation strategy would continue to acquire data randomly about the object, with the convergence solution determined by the luck of the sampling and by the percentage of the viewsphere that contains local minima.

A formal comparison of the two strategies when started from ambiguous views (with correct assertions) can be found in Appendix D. Here, the system was initiated from each high entropy, correct assertion, viewpoint on the densely sampled viewsphere. Figures D.3 and D.4 compare the results in terms of the percentage of iterations that lead to a correct assertion, and the average number of steps taken to achieve convergence. Using the proposed approach, the system converged to the correct solution in 100% of the cases, significantly outperforming the random case. Furthermore, the system converged in close to two steps (where a step refers to a recognition iteration at a single viewpoint) in all cases.

The one exception to the fast convergence results is in the case of the tiger (Ti), where the system converged to the correct solution in 6 steps. This case is an interesting example of the potential weakness of the particular smoothing operator chosen. In most cases, smoothing the entropy map lead to a good choice for the best viewpoint for navigation. However, the best viewpoint chosen for the tiger did not lead to a low entropy value, and was in fact higher in entropy than that of its neighbours. This is due to the fact that the smoothing operator placed a higher degree of weight on the neighbourhood support of the tile than in the local entropy value. As a result, the system lead the sensor to a relatively ambiguous viewpoint (but one that maximally supported by its neighbours), where the system remained stuck for many iterations until convergence to the correct solution was finally attained. In fact, later experiments will show that, in the case of the tiger, slight errors in pose will lead the sensor to neighbours of the best location, thus converging faster than in the perfect pose case. Determining the correct tradeoff between local entropy values and neighbourhood support can be tuned to the task at hand.

2.2. Recognition Results Based on Nearest Neighbour Pose Estimation. Having verified that the navigation system works correctly, the system was now tested using the nearest neighbour indexing techniques described earlier for estimating pose. The same conditions as in the first experiment were met. Once again, the system iterated until one

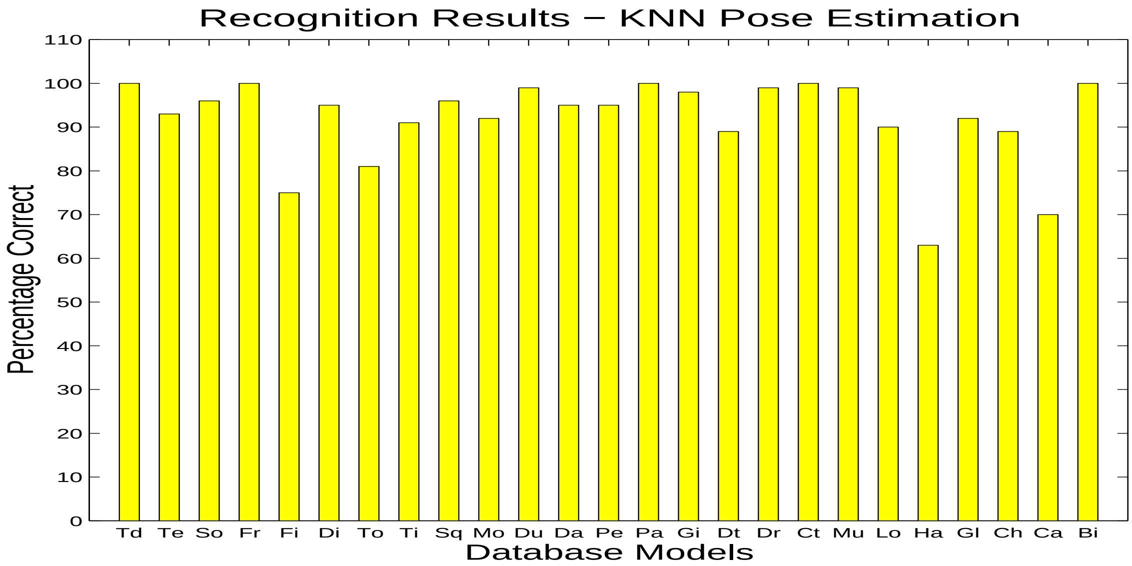


FIGURE 7.8. Percentage correct recognition results at convergence (KNN Pose Estimation).

hundred iterations of navigation (lasting a minimum of two steps) were performed. The percentage of correct MAP recognition results at convergence using the described navigation strategy can be found in Figure 7.8, where the results are plotted for each of the objects of the database. Here, one can see that the system converged to the correct solution in 70%- – 100% of the cases. It did so in less than 3 iterations on average. A slight degradation in performance can be seen as compared to the case where perfect pose was assumed. This is due to errors in pose estimation. In fact, the average pose error (in terms of the estimate of location as produced by the indexing routine versus the true location) over all the objects in the database was 49°. This accounted for significant error in cases where recognition success was sensitive to location. The complete breakdown of error in pose for each object can be seen in Appendix F. The results of comparing the average performance of each object in the database with the random walk approach can be found in Appendix D, Figure D.2.

The more interesting case to examine is when the system is initiated from viewpoints with high entropy (and correct MAP assertions). The same recognition experiment was performed as before, this time using the nearest neighbour approach to pose estimation. The approach was then compared to a random strategy.

Figure 7.9 shows an example comparing the two strategies when started from the same high entropy location (for the Old Dutch cleaner bottle). On the right of the figure, one can see the results of an entropy-based navigation sequence, on the left, the random strategy.

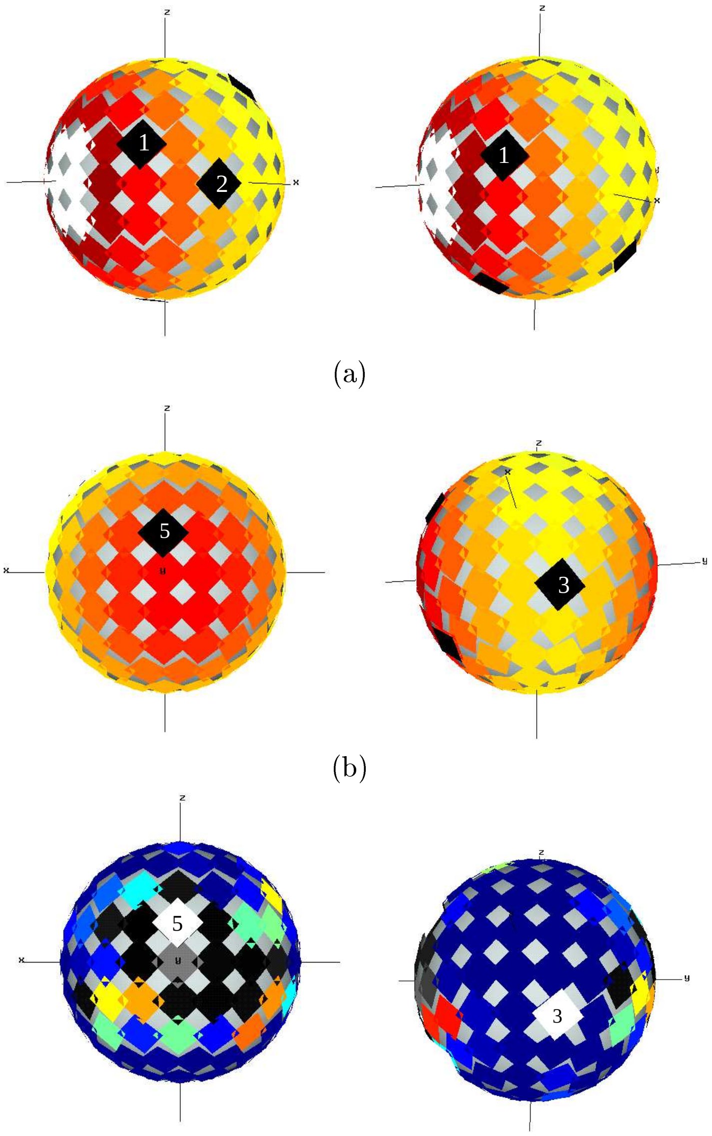
Figure 7.9(a) illustrates the first viewpoint, and Figure 7.9(b) the last viewpoint, in both navigation strategies (seen in black) superimposed onto the smoothed entropy map of the cleaner. Figure 7.9(c) illustrates the last viewpoint (seen in white) superimposed onto the overall entropy map. Using the proposed strategy, convergence to the correct solution occured in three iterations. Notice that the system leads the sensor near to the entropy map minimum of the correct hypothesis (yellow tiles on the smoothed map, blue tiles on the overall map). On the left of the figure, one can see the result of a random sequence, where convergence to the wrong model, was reached in five iterations. This resulted from the effect of images taken at locations where strong belief in the wrong model was present (red tiles on the smoothed map, black tiles on the overall map). This is the danger of the random strategy, where viewpoints can be randomly taken at relatively bad locations, causing convergence to the wrong object.

Figure 7.10 illustrates a comparison of navigation results (using both strategies) starting at high entropy locations in the cases of the Old Dutch cleaner bottle and the bread roll (chala). Here, entropy is plotted over time for each example. At each iteration, the MAP solution is shown above the curve. One can see that the proposed strategy converges quicker than the random strategy in both cases. In fact, in both cases, the random strategy caused the sensor to move to a "bad" local minimum (low entropy, wrong model case) causing convergence to the wrong model.

A formal comparison of the two strategies when starting from ambiguous viewpoints can be found in Appendix D, Figures D.3 and D.4. Once again, a degradation in performance as compared to the perfect pose case can be observed. This is due to errors in the pose estimation routine. However in the majority of the cases, the system still systematically out-performed the random case.

Examining the results in terms of the percentage of correct convergence iterations, the system out-performed the random strategy in 15 cases, had the same results in 6 cases and was marginally less effective in 3 cases. One interesting case is that of the hamburger (Ha) where the system only converged to the correct solution in 53% of the iterations. This relatively poor result can be attributed to the fact that the average error in pose for this object was over 75°. The significance of this high pose error for this object can be seen through a breakdown of its viewsphere into the following four components (as determined by its entropy map):

case1: non-ambiguous and correct identification,



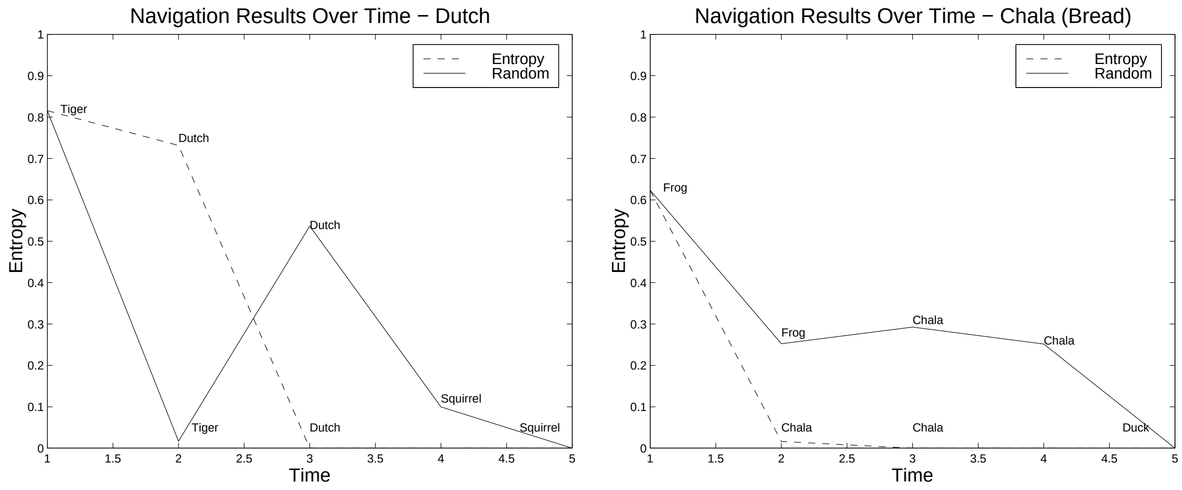
(a)

(b)

(c)

Above we can see the results of a random walk experiment (seen on the left), and one using our navigation strategy (on the right), both initialized from the same location. In (a) and (b), the first and last iteration (seen in black) are superimposed onto the smoothed entropy map of the Old Dutch cleaner bottle. In (c), the last iteration (seen in white) is superimposed onto the overall entropy map. Using the proposed approach, the system converged to the correct solution in the desired neighbourhood (yellow tiles in smoothed map, blue in overall map). Using a random strategy, the system converged to the wrong solution, as many random views were chosen in a "bad" neighbourhood (in red on smoothed map, black on overall map).

FIGURE 7.9. Comparison of entropy map and random navigation strategies: (a) First view on smoothed map, (b) last view on smoothed map, (c) last view on overall entropy map.



(a) (b)

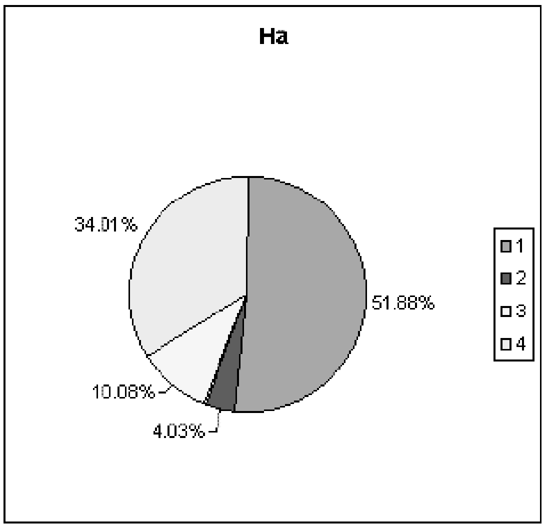
FIGURE 7.10. Navigation results over time for (a) Old Dutch bottle, (b) chala (bread) roll.

case2: ambiguous and correct identification,

case3: ambiguous and incorrect identification,

case4: non-ambiguous and incorrect identification.

Figure 7.11 illustrates a breakdown of the viewsphere about the hamburger into these four cases. Through examination of the hamburger's pie chart, one can see that half of its



Here, one can see the breakdown of the entropy map into categories: (1) non-ambiguous (entropy < 0.5) and correct identification, (2) ambiguous and correct identification, (3) ambiguous and incorrect identification, (4) non-ambiguous and incorrect identification. Notice that a large percentage of the sphere is made up of local minima (case 4). This leads to high possibility of false assertions.

FIGURE 7.11. Breakdown of entropy maps into components.

viewsphere was composed of either highly ambiguous views or viewpoints with high certainty in the incorrect model (i.e. This case makes up over 37% of the sphere). As a result, large deviations from the optimal viewpoint lead to many wrong assertions. The random strategy had similar results at a success rate of 60%. Other objects, whose viewsphere were not quite so sensitive to errors in pose, achieved a much greater level of success over the random case. Similar pie charts for all the objects in the database can be found in Appendix E.

The proposed strategy also out-performed or matched the random case in terms of the number of iterations to convergence. On average, the system took less than three iterations to converge. There were several interesting cases where the random case seemed to out-perform the proposed strategy. For the most part, these were due to failures in the pose estimation in cases where the location sensitivity was high. In the case of the glue bottle (Gl), for example. the system converged in 5.3 iterations on average. Here, the object converged in 2 steps for all but one case, where it took 95 iterations to converge! This was due to repeated errors in pose leading the system to being stuck at a high entropy location, thus taking many iterations to converge. The random strategy had an advantage here, as, never visiting the same location twice, it had no chance of getting stuck. For the frog (Fr), the system converged to the correct solution in 100% of the cases, but in 4.4 iterations on average. This is due to the repetition of a small error in pose, consistently leading the sensor to a sub-optimal location. This is an indication that not only the size of the average error in pose needs to be considered, but how often consistent pose error leads to convergence to the wrong location.

Notice that the tiger (Ti) showed marked improvement, in terms of the number of iterations to convergence, over the perfect pose case. This was due to its relatively small average pose error of 34° (see earlier discussion). Here, it converged to the correct solution in 94.3% of the cases, taking 4.1 steps on average. This success rate was remarkable given the difficulty of the case, where over 63° of the sphere lead to either ambiguous or false assertions. Still, a consistent error in pose lead to relatively slow convergence.

On a final note, empirical evidence indicates the benefits of using an off-line entropy minimization strategy, over on-line methods, in leading the system towards the *global* entropy minimum. In cases where the sensor began with a high confidence in the wrong model, the entropy may increase with each step as it leaves the local minimum, before converging to a global low entropy state. On-line entropy minimization strategies converge to a *local* entropy minimum, even if it belongs to a false assertion.

3. Summary

In this chapter, the active recognition strategy presented in Chapter 5 was tested through extensive experimentation with a real robotic vision system. For these tests, the strategy was applied to the problem of recognizing objects based on signatures in their optical flow images. The empirical evidence verifies the hypothesis that *entropy maps* computed off-line, encoding object discriminability as a function of viewing position, can serve to guide a sensor towards areas that minimize the inter-class confusion between competing object hypotheses during on-line recognition experiments. Results indicate clear superiority of the proposed navigation strategy over a random walk approach, both in terms of the accuracy and speed of convergence, provided an accurate pose estimation module is available. Using nearest neighbour techniques for estimating pose has been shown to be accurate for only some of the objects in the database. An inaccurate pose estimate, combined with a sensitivity to sensor location, lead to a degradation in performance. The rapid convergence rate, even in the case of the random-walk navigation strategy, is due to the strength of the Bayesian updating method.

Empirical evidence has indicated that, while it works in many cases, using indexing techniques for estimating object pose has its weaknesses. Large errors in pose lead to a reduction in the power of the navigation strategy. The results with a working pose estimation procedure were extremely encouraging in that convergence to the correct solution was rapid and consistent. An interesting extension of this work would involve the inclusion of a more robust pose estimation procedure in order to ensure accurate, rapid convergence.

One major strength of this work is the extent to which experimentation on the algorithm was implemented. The enormous technical difficulties encountered in carrying out extensive recognition experiments has often deterred researchers from attempting large-scale experiments. In this set of experiments, entropy maps were built through the acquisition of 744 images at evenly located positions about the viewsphere for 25 objects, for a total of 18,600 images. To avoid biases, this accuracy was ensured through a robotic control system. Furthermore, hundreds of experiments were performed for each object in the database under different conditions in order to examine average performances. The goal behind the considerable effort exerted in carrying out these experiments was be able to expose the intricacies of the workings of the algorithm, in order to permit future extensions of the work.

CHAPTER 8

Conclusions

The difficulty in solving the problem of object recognition lies in its under-determined nature: several models can give rise to identical measurements. As a result, recognition from single viewpoints can lead to ambiguous assertions where more than one solution is possible. Rather than constrain the solution space to attain a single object identity, this thesis introduced a novel solution to the ill-posed recognition problem: a *sequential object recognition strategy* in which evidence in the various hypotheses was accumulated over time until a sufficient degree of confidence in a single object was attained. The key contribution was the way in which prior information from the previous viewpoint *conditioned* the inference process at each iteration, by feeding forward the computed probabilities in the competing hypotheses. Robustness and accuracy were further improved by using prior information to condition the measurement process as well. This lead to a novel active recognition strategy where previously-computed entropy maps were used to guide the sensor to the most discriminant viewpoint for recognition in the shortest number of steps.

Key to the requirement of a sequential recognition system is the ability to *qualify* recognition assertions based on each data set acquired. In this thesis, a general inverse theory was applied to the problem of recognition, and the result was a probabilistic framework, applicable to a large class of inverse problems. This involved explicitly enumerating each source of uncertainty in the problem and representing each as a probability density function. The final recognition result was a probability density function describing the likelihood of the unknown object matching each of the objects in the database, given the acquired data set. As most recognition strategies are deterministic in nature and produce a single object label, this presents a contribution to the field. Furthermore, a novel sequential probabilistic framework was developed to inexpensively update the resulting probability density functions with evidence from previous viewpoints using Bayesian chaining rules.

The generality of the theoretical framework was demonstrated through its application to two different tasks. The first consisted of recognizing 3D parametric models from laser rangefinder data. Application to this problem entailed modifying the forms of the probability density functions. In addition, the strategy was shown to be easily extendable to part/object detection. A novel set of experiments was performed using a complete bottom-up vision system and, despite difficulties associated with compounded ambiguities at every stage processing, the system quickly overcame ambiguities, converging to the correct solution in most cases.

The second application presented an even more important contribution to the field. As a completely novel context, recognition was attempted based on the appearance of signatures in optical flow images resulting from objects moving with respect to a camera. This approach overcame difficulties with standard appearance-based approaches, as its differential nature rendered it somewhat invariant to changes in lighting, position, and background. Yet the difficulty in factoring out the structural signals from the confounding motion information, combined with the prevalent noise, made the problem of recognition based on flow images particularly challenging. Considerable experimentation of the proposed strategy with a real vision system demonstrated that these difficulties can be quickly resolved through the accumulation evidence for the different hypotheses over a sequence of views. Choosing a winner after several viewpoints gave the correct result in most cases.

Many applications, in mobile robotics for example, place constraints on the time allowed to attain a solution. In addition, a large cost may be associated with false assertions. This lead to another main contribution of the work: the introduction of an *active recognition strategy* that made use of *a priori* information in order to guide the sensor to locations of maximal discriminability. The advantage of this approach over the few existing on-line strategies was that (a) most of the computations were performed off-line, thus reducing the processing speed in on-line experiments, (b) the strategy led the sensor to a *global* solution, whereas minimization strategies such as gradient descent converge to local minima. Results of a large number of experiments indicated that the system converged faster and more accurately using this approach, especially with accurate estimation of pose.

Many future extensions of the active recognition strategy are possible. One important outcome of the thesis was the introduction of the notion of a set of *characteristic motions*.

Through a carefully built control system, precise motion samples were extracted about each set of objects in the database. Empirical evidence indicated that by training on this smaller set, recognition based on a larger set of motions becomes possible. Large-scale testing of this hypothesis is currently underway in the Artificial Perception Laboratory of McGill University. The entropy maps can be used to extend the notion of *characteristic motions* to one that is analogous to the traditional notion of characteristic views, by permitting extraction of the most informative and stable viewpoints for recognition. An interesting extension of this work would be to automatically determine a minimal subset of these position/motion views for accurate recognition.

Improvements to the navigation framework would include a more accurate pose estimation process. This would improve the results considerably. This, of course, is an open area of research. Inclusion of information from several competing entropy maps at each iteration of recognition, while maintaining reasonable computational cost, would be an interesting extension of the work. The addition of information from on-line entropy computations would help determine the validity of the MAP result at each iteration. Each of these would be interesting topics of future research.

The strategy is certainly applicable to many contexts. The system easily lends itself to the context of mobile robotics, for example, where an autonomous robot navigating through a scene can use the proposed strategy in order to recognize landmarks. This aids in self-localization, a difficult problem in the field. The versatility of the approach has already been seen through its application to the difficult problem of *Content Based Image Retrieval* [**15**], where the Bayesian framework was used to determine the likelihood of an image matching a set of images in a database.

On a final note, this thesis presents the first steps to a comprehensive active solution to the problem of recognition. The purpose in formulating a general framework for the solution to this difficult problem, and for presenting a large amount of experimental results, is to facilitate continuation of the work. If this path is continued, it can lead to a true autonomous visual system, capable of exploring and understanding its environment.

REFERENCES

[1] T. Arbel, *Recognizing volumetric objects in the presence of uncertainty*, Master's thesis, McGill University, Montréal, Québec, Canada, April 1995.

[2] T. Arbel and F. P. Ferrie, *Informative views and sequential recognition*, Tech. Report TR-CIM-95-10, Center for Intelligent Machines, McGill University, Montréal, Québec, Canada, nov 1995, Available via ftp at `ftp.cim.mcgill.ca` in `pub/techrep/1995/CIM-95-10.ps.Z`.

[3] ———, *Informative views and sequential recognition*, 4th European Conference on Computer Vision (Cambridge, UK) (B. Buxton and R. Cipolla, eds.), 1064, Springer–Verlag, Apr 14-18 1996,, pp. 469–481.

[4] ———, *Entropy-based gaze planning*, Proceedings of the Second IEEE Workshop on Perception for Mobile Agents (Fort Collins, Colorado), June 1999, In association with the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 87–94.

[5] ———, *Viewpoint selection by navigation through entropy maps*, Proceedings of the Seventh IEEE International Conference on Computer Vision (Kerkyra, Greece), Sept 20-25 1999, pp. 248–254.

[6] ———, *Recognizing objects by accumulating evidence over time*, Fourth Asian Conference on Computer Vision (Taipei, Taiwan), Jan 8-11 2000, p. to appear.

[7] T. Arbel, P. Whaite, and F. P. Ferrie, *Recognizing volumetric objects in the presence of uncertainty*, Proceedings 12th International Conference on Pattern Recognition (Jerusalem, Israel), IEEE Computer Society Press, Oct 9-13 1994, pp. 470–476.

[8] ———, *Parametric shape recognition using a probabilistic inverse theory*, Pattern Recognition Letters **17** (1996), no. 5, 491–501.

[9] F. Arman and J. K. Aggarwal, *Model-based object recognition in dense-range images - a review*, ACM Computing Surveys **25** (1993), no. 1, 5–43.

[10] G. Aviv, *Inherent ambiguities in recovering 3-d motion and structure from a noisy flow field*, IEEE Transactions on Pattern Analysis and Machine Intelligence **II** (1989), no. 5, 477–489.

[11] W. Wahid B. Moghaddam and A. Pentland, *Beyond eigenfaces: Probabilistic matching for face recognition*, Tech. Report 443, M.I.T. Media Laboratory Perceptual Computing Section, 1998, Appears in: The 3rd IEEE Intl. Conference on Automatic Face and Gesture Recogntion, Nara, Japan, April 1998.

[12] R. Bajcsy and F. Solina, *Three dimensional object recognition revisited*, Proceedings, 1ST International Conference on Computer Vision (London,U.K.), Computer Society of the IEEE, IEEE Computer Society Press, June 1987.

[13] A. H. Barr, *Superquadrics and angle preserving transformations*, IEEE Computer Graphics and Applications **1** (1981), no. 1, 11–23.

[14] S. M. Benoit and F. P. Ferrie, *Monocular optical flow for real-time vision systems*, Proceedings of the 13th International Conference on Pattern Recognition (Vienna, Austria), 25–30 August 1996, pp. 864–868.

[15] F. Beyrouti, *An improved appearance-based approach to image retrieval and classification*, Master's thesis, McGill University, Montréal, Québec, Canada, to be published 1998.

[16] M. J. Black and A. D. Jepson, *A probablistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions*, ECCV98 (Freiburg, Germany), vol. I, Springer-Verlag, June 1998, pp. 909–924.

[17] ———, *Recognizing temporal trajectories using the condensation algorithm*, Intl. Conf. on Automatic Face and Gesture Recognition (Nara, Japan), 1998.

[18] M. J. Black and Y. Yacoob, *Recognizing facial expressions in image sequences using local parametrized models of image motion*, Int. Journal of Computer Vision **25** (1997), no. 1, 23–48, Also found in Xerox PARC, Techinical Report SPL-95-020.

[19] A. F. Bobick and J. W. Davis, *An appearance-based representation of action*, Tech. Report 369, MIT Media Lab, February 1996, As submitted to ICPR 96.

[20] T. E. Boult and A. D. Gross, *On the recovery of superellipsoids*, Proc. of DARPA Image Understanding Workshop (Washington, D.C.), 1988, pp. 1052–1063.

[21] K. Bowyer and C. Dyer, *Aspect graphs: An introduction and survey of recent results*, Close Range Photogrammetry Meets Machine Vision, vol. 1395, SPIE, 1990, pp. 200–208.

[22] W. Burgard, D. Fox, and S. Thrun, *Active mobile robot localization*, Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (Nagoya, Japan), August 23-29 1997.

[23] F. G. Callari and F. P. Ferrie, *Active recognition: Using uncertainty to reduce ambiguity*, Proceedings of the 13th International Conference on Pattern Recognition (Vienna, Austria), International Association for Pattern Recognition, IEEE-CS, 25–30 August 1996, pp. 925–929.

[24] R. Campbell and P. Flynn, *Eigenshapes for 3d object recognition in range data*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Fort Collins, CO), vol. 2, IEEE Computer Society Press, June 23–25 1999, pp. 505–510.

[25] R. T. Chin and C. R. Dyer, *Model-based recognition in robot vision*, Computing Surveys **18** (1986), no. 1, 67–108.

[26] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley & Sons, 1991.

[27] T. Darrell, S. Sclaroff, and A. Pentland, *Segmentation by minimal description*, Proceedings, 3RD International Conference on Computer Vision (Osaka,Japan), IEEE Computer Society, IEEE Computer Society Press, December 1990, pp. 112–116.

[28] T. J. Darrell and A. P. Pentland, *Recognition of space-time gestures using a distributed representation*, Tech. Report 197, M.I.T. Media Laboratory Vision and Modelling Group, 1992.

[29] A. P. Dempster and G. Shafer, *A mathematical theory of evidence*, 1976.

[30] S. J. Dickinson, H. I. Christensen, J. K. Tsotsos, and G. Olofsson, *Active object recognition integrating attention and viewpoint control*, Computer Vision, Image Understanding **67** (1997), no. 3, 239–260.

[31] S. J. Dickinson, A. P. Pentland, and A. Rosenfeld, *Qualitative 3-D shape reconstruction using distributed aspect graph matching*, Proceedings, 3RD International Conference on Computer Vision (Osaka,Japan), IEEE Computer Society, IEEE Computer Society Press, December 1990, pp. 257–262.

[32] S. J. Dickinson, A.P. Pentland, and A. Rosenfeld, *3-D shape recovery using distributed aspect matching*, IEEE Transactions on Pattern Analysis and Machine Intelligence **14** (1992), no. 2, 174–198.

[33] H.F. Durrant-Whyte, *Sensor models and multisensor integration*, International Journal of Robotics Research **7** (1988), no. 6, 97–113.

[34] D. W. Eggert, K. W. Bowyer, C. R. Dyer, H. I. Christensen, and D. B. Goldgof, *The scale space aspect graph*, Proceedings, Conference on Computer Vision and Pattern Recognition (Champaign, Il.), IEEE, June 15-18 1992, pp. 335–340.

[35] T. J. Fan, G. Medioni, and R. Nevatia, *Segmented descriptions of 3-D surfaces*, IEEE Int. J. Robot. Automat. **3** (1987), no. 6, 527–538.

[36] ———, *Recognizing 3-D objects using surface descriptions*, IEEE Transactions on Pattern Analysis and Machine Intelligence **11** (1989), no. 11, 1140–1157.

[37] T.J. Fan, *Describing and recognizing 3d objects using surface properties*, Springer–Verlag, New York, NY, USA, 1990.

[38] F. P. Ferrie and J. Lagarde, *On computing stable surface descriptions from range images*, Proceedings 5th International Conference on Image Analysis (Positano, Italy), September 20-22 1989.

[39] F. P. Ferrie, J. Lagarde, and P. Whaite, *Darboux frames, snakes, and super-quadrics: Geometry from the bottom up*, IEEE Transactions on Pattern Analysis and Machine Intelligence **15** (1993), no. 8, 771–784.

[40] P. J. Flynn and A. K. Jain, *BONSAI: 3-D object recognition using constrained search*, IEEE Transactions on Pattern Analysis and Machine Intelligence **13** (1991), no. 10, 1066–1075.

[41] W. E. L. Grimson, *On the recognition of parametrized objects*, The 4th International Symposium on Robotics Research (Santa Cruz, California), August 1987.

[42] ———, *On the recognition of curved objects*, IEEE Transactions on Pattern Analysis and Machine Intelligence **11** (1989), no. 6, 632–642.

[43] W. E. L. Grimson and T. Lozano-Perez, *Localizing overlapping parts by searching the interpretation tree*, IEEE Transactions on Pattern Analysis and Machine Intelligence **9** (1987), no. 4, 469–482.

[44] J. Hadamard, *Sur les problèmes aux derivées partielles at lear signification physique*, Tech. report, Princeton University Bulletin, 1902.

[45] ______, *Lectures on the cauchy problem in linear partial differential equations*, New Haven, CT: Yale University Press, 1923.

[46] S. Herbin, *Recognizing 3d objects by generating random actions*, Proceedings, Conference on Computer Vision and Pattern Recognition (San Francisco), Computer Society of the IEEE, IEEE Computer Society Press, June 1996, pp. 35–40.

[47] R. A. Hummel and M. S. Landy, *A statistical viewpoint on the theory of evidence*, PAMI **10** (1988), no. 2, 235–247.

[48] S. Hutchinson and A. Kak, *Planning sensing strategies in a robot work cell with multi-sensor capabilities*, IEEE Transactions on Robotics and Automation **5** (1989), no. 6, 765–783.

[49] S. A. Hutchinson, R. L. Cromwell, and A. C. Kak, *Applying uncertainty reasoning to model based object recognition*, Proceedings, Conference on Computer Vision and Pattern Recognition (San Diego, Calif.), Computer Society of the IEEE, IEEE Computer Society Press, June 4-8 1989, pp. 541–548.

[50] D. P. Huttenlocher and S. Ullman, *Object recognition using alignment*, Proceedings, 1ST International Conference on Computer Vision (London,U.K.), Computer Society of the IEEE, IEEE Computer Society Press, June 1987, pp. 102–111.

[51] M. Isard and A. Blake, *Contour tracking by stohastic propagation of conditional density*, Computer Vision – ECCV 96 (Cambridge, UK) (B. Buxton and R. Cipolla, eds.), Springer–Verlag, Apr 14-18 1996, pp. 343–356.

[52] A. K. Jain and R. Hoffman, *Evidence-based recogntion of 3d objects*, IEEE Transactions on Pattern Analysis and Machine Intelligence **10** (1988), no. 6, 783–802.

[53] T. Jebara, K. Russell, and A. Pentland, *Mixtures of eigenfeatures for real-time structure from texture*, Proceedings, 6TH International Conference on Computer Vision (Bombay, India), Computer Society of the IEEE, IEEE Computer Society Press, Jan 1998.

[54] D. Keren, D. Cooper, and J. Subrahmonia, *Describing complicated objects by implicit polynomials*, Tech. Report 102, Brown University LEMS, Laboratory for Engineering Man/Macine Systems, Division of Engineering, Brown University, Providence RI 021912 USA, 1992.

[55] W. Y. Kim and A. C. Kak, *3-D object recognition using bipartite matching embedded in discrete relaxation*, IEEE Transactions on Pattern Analysis and Machine Intelligence **13** (1991), no. 3, 224–251.

[56] M. Kirby and L. Sirovich, *Application of the karhunen-loeve procedure for the characterization of human faces*, IEEE Transactions on Pattern Analysis and Machine Intelligence **12** (1990), no. 1.

[57] J. Kittler and E. R. Hancock, *Combining evidence in probabilistic relaxation*, PRAI **3** (1989), 29–51.

[58] D. J. Kriegman and J. Ponce, *Computing exact aspect graphs of curved objects: Solids of revolution*, PROC. of IEEE Workshop on the Interpretation of 3-D Scenes (Austin, Texas), IEEE, November 27-29 1989, pp. 116–122.

[59] ———, *On recognizing and positionind curved 3-D objects from image contours*, IEEE Transactions on Pattern Analysis and Machine Intelligence **12** (1990), no. 12, 1127–1137.

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

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

[62] A. Lejeune and F. P. Ferrie, *Partitioning range images using curvature and scale*, PROC. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (New York City, New York), June 15-17 1993, pp. 800–801.

[63] N. Li, S. Dettmer, and M. Shah, *Visually recognizing speech using eigen sequences*, MBR97, 1997, p. Chapter 15.

[64] D. G. Lowe (ed.), *Perceptual organization and visual recognition*, Kluwer, Boston, 1985.

[65] M. Izumi M. Onishi, N. Nishikawa and K. Fukunaga, *Object recognition using sequential images, and application to active vision*, 2nd International workshop on Statistical Techniques in Pattern Recognition (Sydney, Australia), 1998, Osaka Prefecture University, Japan.

[66] N. Takeda M. Watanabe and K. Onoguchi, *A moving object recognition method by optical flow analysis*, Proceedings of the 13th International Conference on Pattern Recognition (Vienna, Austria), A, vol. 1, International Association for Pattern Recognition, Aug 1996, pp. 528–533.

[67] D. J. C. MacKay, *Bayesian interpolation*, Neural Computation **4** (1991), no. 3, 415–447.

[68] ———, *Information-based objective functions for active data selection*, Neural Computation **4** (1991), no. 4, 589–603.

[69] D. Marr, *Vision*, W.H. Freeman & Co., San Francisco, 1982.

[70] D. Nair and J. K. Aggarwal, *Hierarchical, modular architectures for object recognition by parts*, ICPR96 (Vienna, Austria), Proc. of the 13th International Conference on Pattern Recognition, IEEE Computer Society Press, August 1996, pp. 601–606.

[71] S. K. Nayar, H. Murase, and S. A. Nene, *Parametric appearance representation in early visual learning*, ch. 6, Oxford University Press, February 1996.

[72] T. S. Newman, P. J. Flynn, and A. K. Jain, *Model-based classification of quadric surfaces*, Computer Vision, Graphics, and Image Processing:Image Understanding **57** (1993), no. 2, 235–249.

[73] B. O'Neill, *Elementary differential geometry*, Academic Press, New York, N.Y., 1966.

[74] A. Pentland, *Recognition by parts*, Proceedings, 1ST International Conference on Computer Vision (London,UK), IEEE Computer Society, IEEE Computer Society Press, June 1987, pp. 612–620.

[75] A. Pentland, R. W. Picard, and S. Sclaroff, *Photobook: Content-based manipulation of image databases*, International Journal of Computer Vision **18** (1996), no. 3, 233–254.

[76] A. Pentland and S. Sclaroff, *Closed form solutions for physically based shape modelling and recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence: Special Issue on Physical Modeling in Computer Vision (T. Kanade and K. Ikeuchi, eds.), vol. 13(7), July 1991, pp. 715–729.

[77] A. Pope and D. G. Lowe, *Learning probabilistic appearance models for object recognition*, ch. xx, pp. 67–98, Oxford University Press, 1996.

[78] S. Picard R. Mohr and C. Schmid, *Bayesian decision versus voting for image retrieval*, CAIP:97, 1997.

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

[80] B. S. Rao and H. Durrant-Whyte, *A decentralized bayesian algorithm for identification of tracked targets*, IEEE Transactions on Systems, Man, and Cybernetics **23** (1993), no. 6, 1685–1698.

[81] A. P. Reeves and R. W. Taylor, *Identification of three-dimensional objects using range information*, IEEE Transactions on Pattern Analysis and Machine Intelligence **11** (1989), no. 4, 403–410.

[82] R. Rimey and C. Brown, *Where to look next using bayes nets: Incorporating geometric relations*, Computer Vision – ECCV 92 (Santa margherita Ligure, Italy) (G. Sandini, ed.), vol. 588, Springer-Verlag, May 1992, pp. 542–550.

[83] N. Roy, W. Burgard, D. Fox, and S. Thrun, *Coastal navigation - mobile robot navigation with uncertainty in dynamic environments*, International Conference on Robotics and Automation, 1999, To Appear in ICRA '99.

[84] B. Schiele and J. L. Crowley, *Probabilistic object recognition using multidimensional receptive field histogram*, Proceedings of the 13th International Conference on Pattern Recognition (Vienna, Austria), International Association for Pattern Recognition, IEEE-CS, 25–30 August 1996.

[85] B. Schiele and J. L. Crowley, *Transinformation for active object recognition*, Proceedings, 6TH International Conference on Computer Vision (Bombay, India), IEEE Computer Society, IEEE Computer Society Press, January 1998, pp. 249–254.

[86] I. Shimshoni and J. Ponce, *Probabilistic 3D object recogition*, Proceedings, 5TH International Conference on Computer Vision (Cambridge, Massachusetts), Computer Society of the IEEE, IEEE Computer Society Press, June18-21 1995, pp. 488–493.

[87] L. Sirovich and M. Kirby, *Low-dimensional procedure for the characterization of human faces*, Journal of the Optical Society of America **4** (1987), no. 3, 519–524.

[88] G. Soucy, *View correspondence using curvature and motion consistency*, Master's thesis, Dept. of E.E., McGill Univ., 1992.

[89] T. Sripradisvarakul and R. Jain, *Generating aspect graph for curved objects*, PROC. of IEEE Workshop on the Interpretation of 3-D Scenes (Austin, Texas), IEEE, November 27-29 1989, pp. 109–115.

[90] L. Stark and K. Bowyer, *Achieving generalized object recognition through reasoning about association of function to structure*, IEEE Transactions on Pattern Analysis and Machine Intelligence **13** (1991), no. 10, 1097–1104.

[91] J. Subrahmonia, D. B. Cooper, and D. Keren, *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, 1992.

[92] H. Takeda, C. Faccinetti, and J. C. Latombe, *Planning the motions of a mobile robot in a sensory uncertainty field*, IEEE Transactions on Pattern Analysis and Machine Intelligence **16** (1994), no. 10, 1002–1017.

[93] A. Tarantola, *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., 1987.

[94] D. W. Thompson and J. L. Mundy, *Model-directed object recognition on the connection machine*, Proc. DARPA Image Understanding Workshop (Los Angeles, CA.), 1987, pp. 93–106.

[95] M. Turk and A. P. Pentland, *Eigenfaces for recognition*, CogNeuro **3** (1991), no. 1, 71–96.

[96] A. Verri and T. Poggio, *Motion field and optical flow: Qualitative properties*, IEEE Transactions on Pattern Analysis and Machine Intelligence **II** (1989), no. 5, 490–498.

[97] W.M. Wells III, *Statistical object recognition*, Ph. D., 1992.

[98] P. Whaite and F. P. Ferrie, *From uncertainty to visual exploration*, IEEE Transactions on Pattern Analysis and Machine Intelligence **13** (1991), no. 10, 1038–1049.

[99] ______, *Autonomous exploration: Driven by uncertainty*, Tech. Report TR-CIM-93-17, Center for Intelligent Machines, McGill University, Montréal, Québec, Canada, 1993, Available via ftp at `ftp.cim.mcgill.ca` in `/pub/3d/papers/tr-cim-93-17.ps.gz`.

[100] ______, *Model building and autonomous exploration*, SPIE - Intelligent Robots and Computer Vision XII: Active Vision and 3D Methods (Boston, Massachusetts), September 8–9 1993, pp. 73–85.

[101] ______, *Autonomous exploration: Driven by uncertainty*, Proceedings, Conference on Computer Vision and Pattern Recognition (Seattle, Washington), IEEE Computer Society, IEEE Computer Society Press, June 21-23 1994, See also [**99**], pp. 339–346.

[102] ______, *Autonomous exploration: Driven by uncertainty*, IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997), no. 3, 193–205.

[103] P. Whaite and F. P. Ferrie, *On the sequential determination of model misfit*, IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997), no. 8, 899–905.

[104] D. Wilkes and J. K. Tsotsos, *Active object recognition*, Proceedings, Conference on Computer Vision and Pattern Recognition (Champaign, Il.), Computer Society of the IEEE, IEEE Computer Society Press, June 15-18 1992, pp. 136–141.

[105] K. Wu and M. D. Levine, *Recovering parametric geons from multiview range data*, International Conference on Computer Vision (Seattle,WA.), IEEE Computer Society, IEEE Computer Society Press, June 1994, pp. 159–166.

APPENDIX A

Principal Component Analysis

Appearance-based strategies often employ a method referred to as Principal Component Analysis (PCA) for finding a lower dimensional subspace in which to represent large amounts of image data. The appeal of the technique is its speed, its reasonable simplicity, its accuracy in constrained environments, and its relative insensitivity to small perturbations in the image. The PCA methods used in this thesis follow the standard approaches in the literature [**95**], whereas the actual code implemented was that of Beyrouti et al. [**15**]. This appendix will provide a brief overview of the techniques involved.

The main idea in PCA (or Karhunen-Loeve expansion) is to find the optimal subspace in which to represent data of high dimension. If N is the dimension in which the data lies, the goal is to find a subspace of dimension M, where $M << N$ that minimizes the least-squares error.

In the case of appearance-based matching, each image of dimension $N \times N$ can be represented by a single vector $\mathbf{x}$ of length N^2. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K\}$ denote the entire set of training K images in the database. The average image for the set is then computed in the usual manner:

$$\mu = \frac{1}{K} \sum_{i}^{K} \mathbf{x}_i. \tag{A.1}$$

This average image is then subtracted from each vector:

$$\mathbf{y}_i = \mathbf{x}_i - \mu, \quad i = 1 \ldots K. \tag{A.2}$$

It will now be shown how these vectors can used to obtain the set of M orthonormal vectors that best represents the data distribution. The sample covariance matrix $\mathbf{C}$ is computed for the set of vectors $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K\}$ in the usual manner:

$$\mathbf{C} = \frac{1}{K}\sum_{i}^{K} \mathbf{y}_i\mathbf{y}_i^T = \mathbf{A}\mathbf{A}^T, \tag{A.3}$$

where $\mathbf{A} = [\mathbf{y}_1\ \mathbf{y}_2\ \ldots\ \mathbf{y}_K]$. The interesting result is that the eigenvectors of this covariance matrix span the optimal subspace in which to represent the data. The lth eigenvector, $\mathbf{u}_l$, and eigenvalue, λ_l, of the covariance matrix, $\mathbf{C}$, are computed in the usual manner, by maximizing:

$$\lambda_l = \frac{1}{K}\sum_{i}^{K} (\mathbf{u}_l^T\mathbf{y}_i)^2. \tag{A.4}$$

such that:

$$\mathbf{u}_l^T\mathbf{u}_j = \begin{cases} 1 & \text{if l = j,} \\ 0 & \text{if l} \neq \text{j.} \end{cases} \tag{A.5}$$

Of course, as the dimensionality of the covariance matrix is N^2, computing N^2 eigenvectors becomes intractable. However, a significant simplification can be made in the case where the number of images K is much smaller than the total dimension of the space N^2. In this case, only the top $K-1$ eigenvectors have corresponding eigenvalues with non-zero values. It can be shown that these eigenvectors can be found by first computing the $K-1$ eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{K-1}\}$ of the reduced $K \times K$ matrix, $\mathbf{A}^T\mathbf{A}$. Then, the original eigenvectors can be generated through a linear combination of the original images:

$$\mathbf{u}_j = \mathbf{A}\mathbf{v}_j, \quad j = 1\ldots K-1. \tag{A.6}$$

The complexity of the calculations has now been significantly reduced to the order of the number of images in the training set. These eigenvectors, referred to as the *principal eigenvectors*, can now be ranked in accordance with their eigenvalues. Empirically, it can be shown that the top m eigenvalues usually contain the majority of the information regarding the variation of the distribution. In [**15**], there are some suggestions as to how to operationally choose the value for m.

Once, the top m eigenvectors $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m\}$ is computed, the optimal subspace for this training set is established. The training images can then be represented in this low dimensional space through a projection operation.

APPENDIX B

Database Names

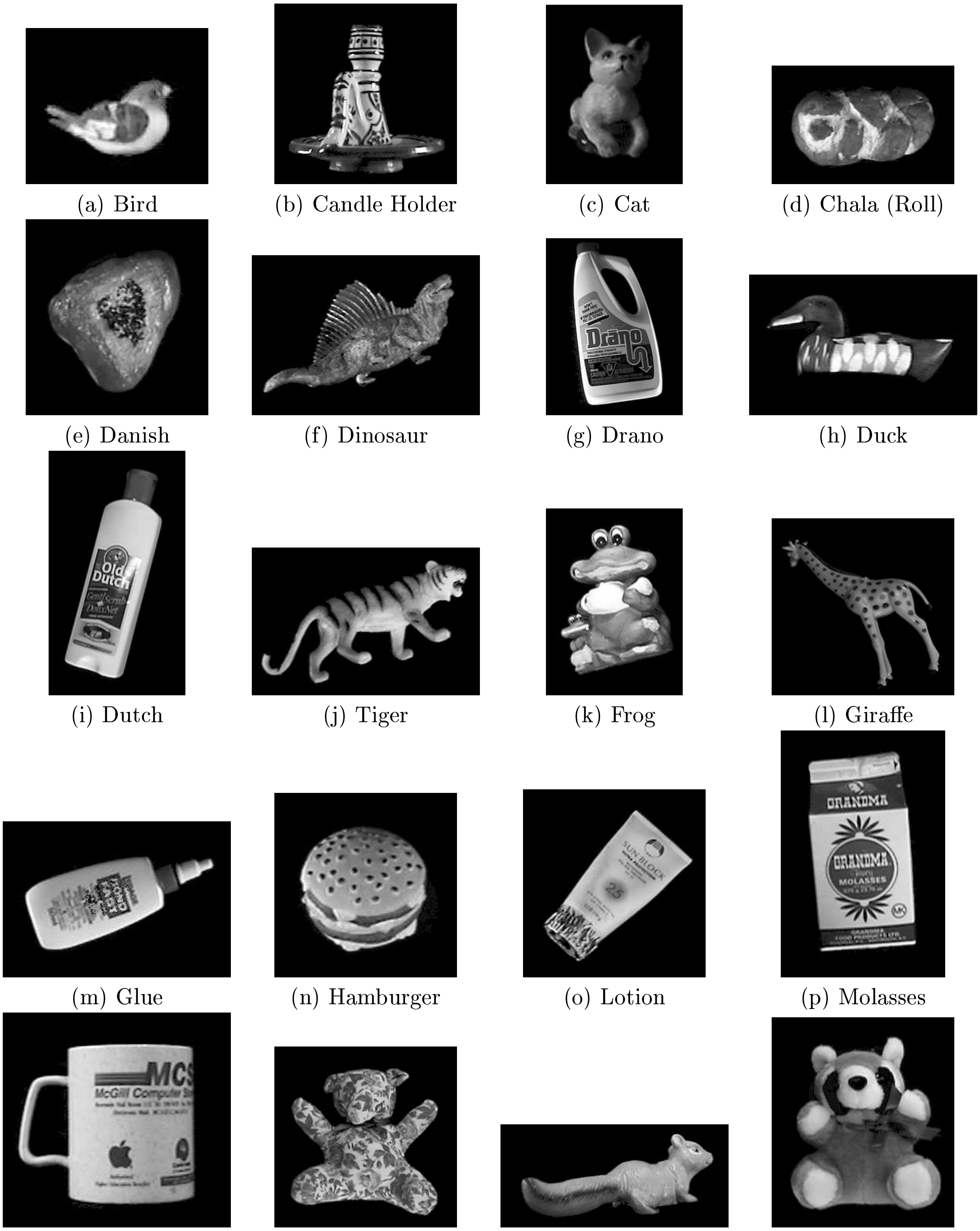
Here, one can find the list of objects placed in the database along with their abbreviated names used in the text.

Abbreviated Name	Database Object
Td	Teddy bear
Te	Tea box
So	House made of soap
Fr	Frog toy
Fi	Fish toy
Di	Dinosaur toy
To	Toothpase tube (Colgate)
Ti	Tiger toy
Sq	Squirrel toy
Mo	Molasses container
Du	Duck - toilet cleaner
Da	Plastic danish
Pe	Pepper
Pa	Panda toy
Gi	Giraffe toy
Dt	Old Dutch cleaner
Dr	Liquid Drano bottle
Ct	Toy cat
Mu	Mug
Lo	Lotion container
Ha	Hamburger toy
Gl	Glue bottle
Ch	Chala (bread)
Ca	Candle holder
Bi	Bird

APPENDIX C

Database Objects

Below, one can find images of the 25 database objects used in recognition experiments based on flow images (see Section 2). Note that the objects are not seen to scale.



(a) Bird (b) Candle Holder (c) Cat (d) Chala (Roll)
(e) Danish (f) Dinosaur (g) Drano (h) Duck
(i) Dutch (j) Tiger (k) Frog (l) Giraffe
(m) Glue (n) Hamburger (o) Lotion (p) Molasses
(q) Mug (r) Teddy Bear (s) Squirrel (t) Panda Bear

FIGURE C.1. Images of the 25 Database Objects (The first 20 are shown here).

(u) Tea (v) Pepper (w) Fish (x) Soap House (y) Toothpaste

FIGURE C.2. Images of the 25 Database Objects (The last 5 are shown here).

APPENDIX D

Comparison of Navigation Strategy with Random Approach

Here, one can find the results of the experiments outlined in Section 2. Figure D.1 illustrates a comparison of the performance of the navigation strategy with a random approach, in terms of the percentage of iterations that lead to correct convergence. Here, the process was strated from arbitrary viewing positions. The pose estimation process was assumed to work perfectly. The corresponding comparison using nearest neighbour pose estimation can be found in Figure D.2.

Figure D.3 illustrates the results of experiments where the system started from ambiguous viewpoints, and perfect pose estimation was assumed. A comparison of speed of convergence is shown in Figure D.4. The analogous results using the nearest neighbour approach for pose estimation can be seen in Figures D.5 and D.6.

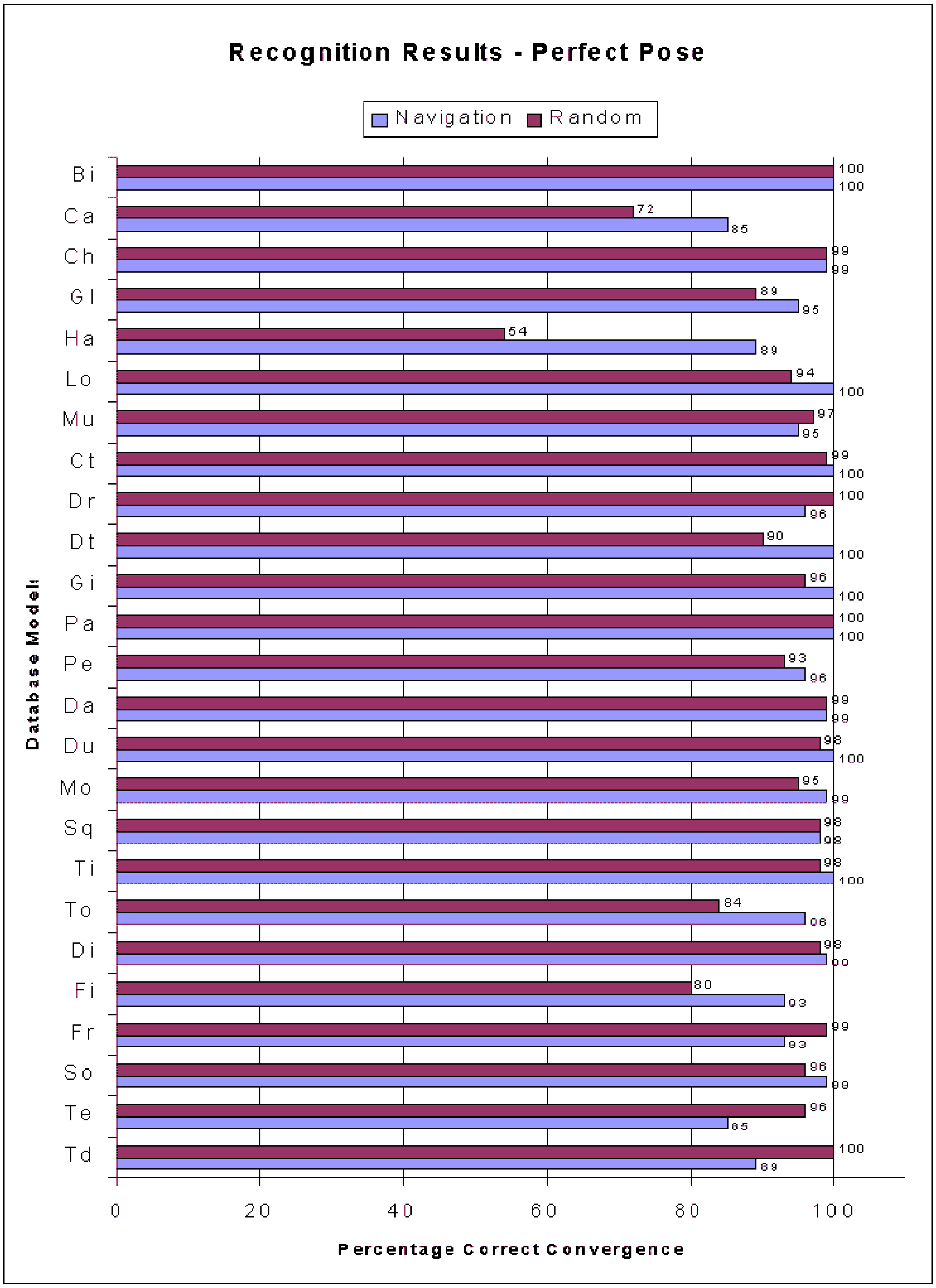


FIGURE D.1. Comparison of two strategies from arbitrary viewpoints (Perfect pose estimation).

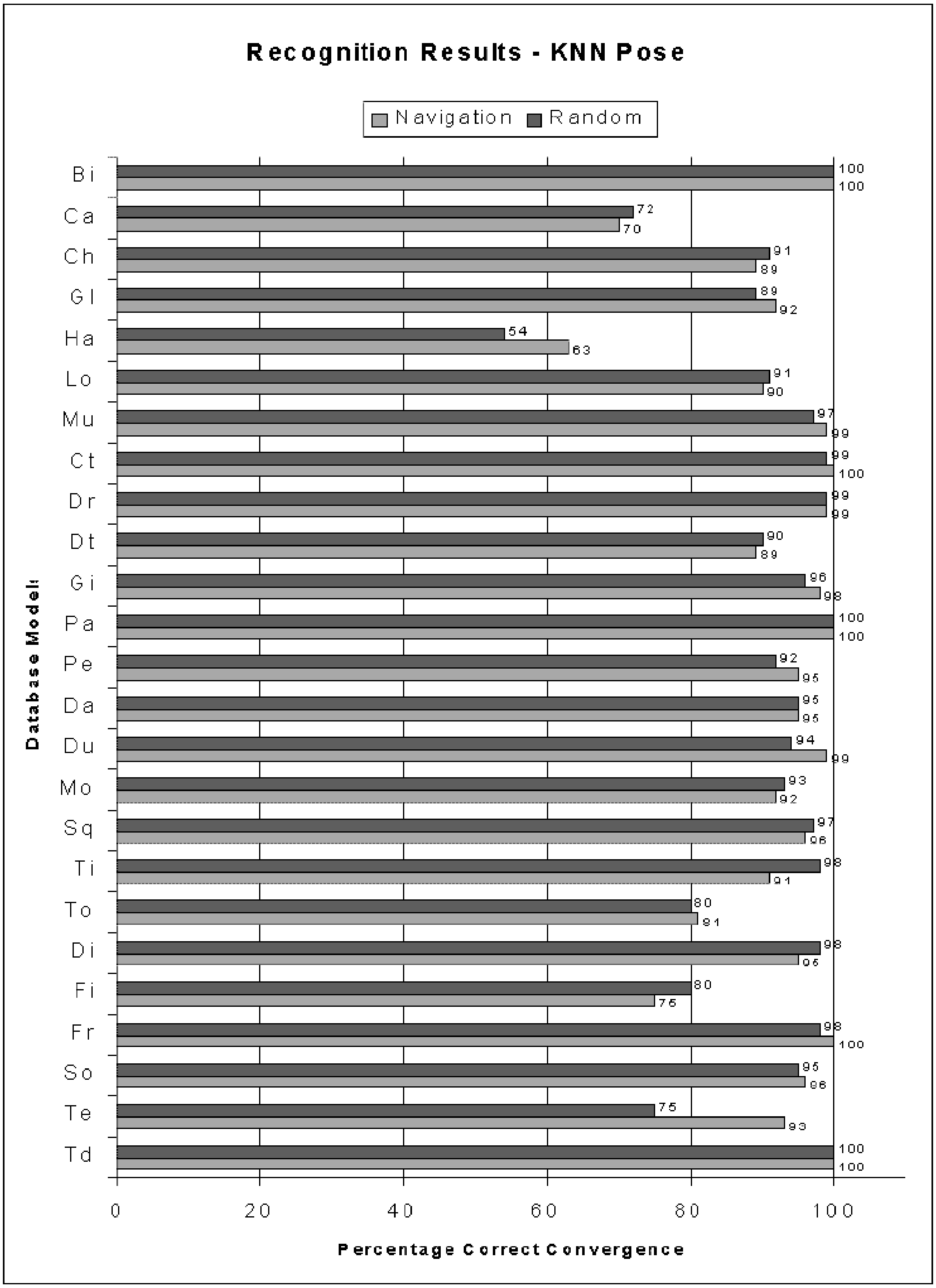


FIGURE D.2. Comparison of two strategies from arbitrary viewpoints (KNN pose estimation).

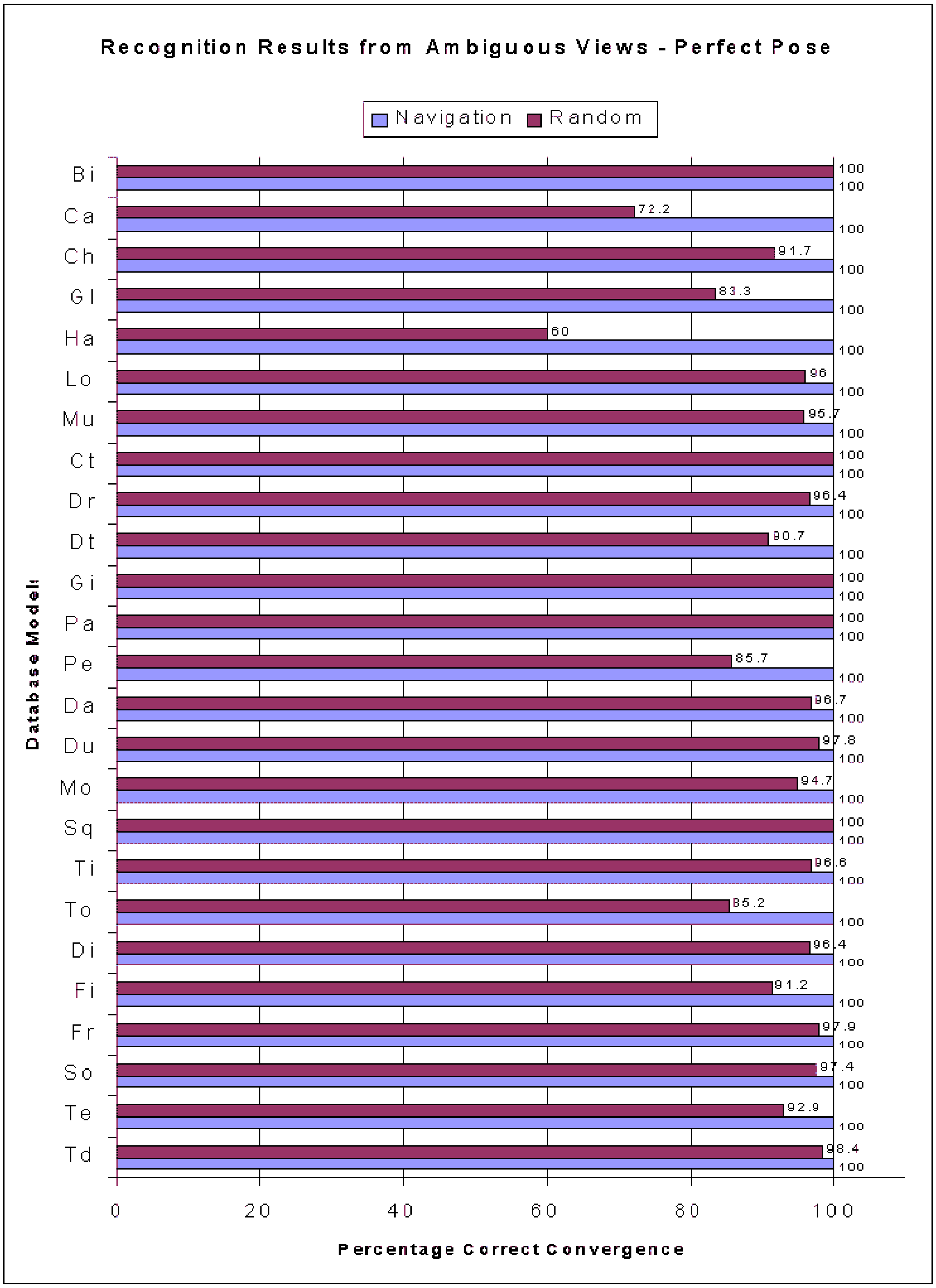
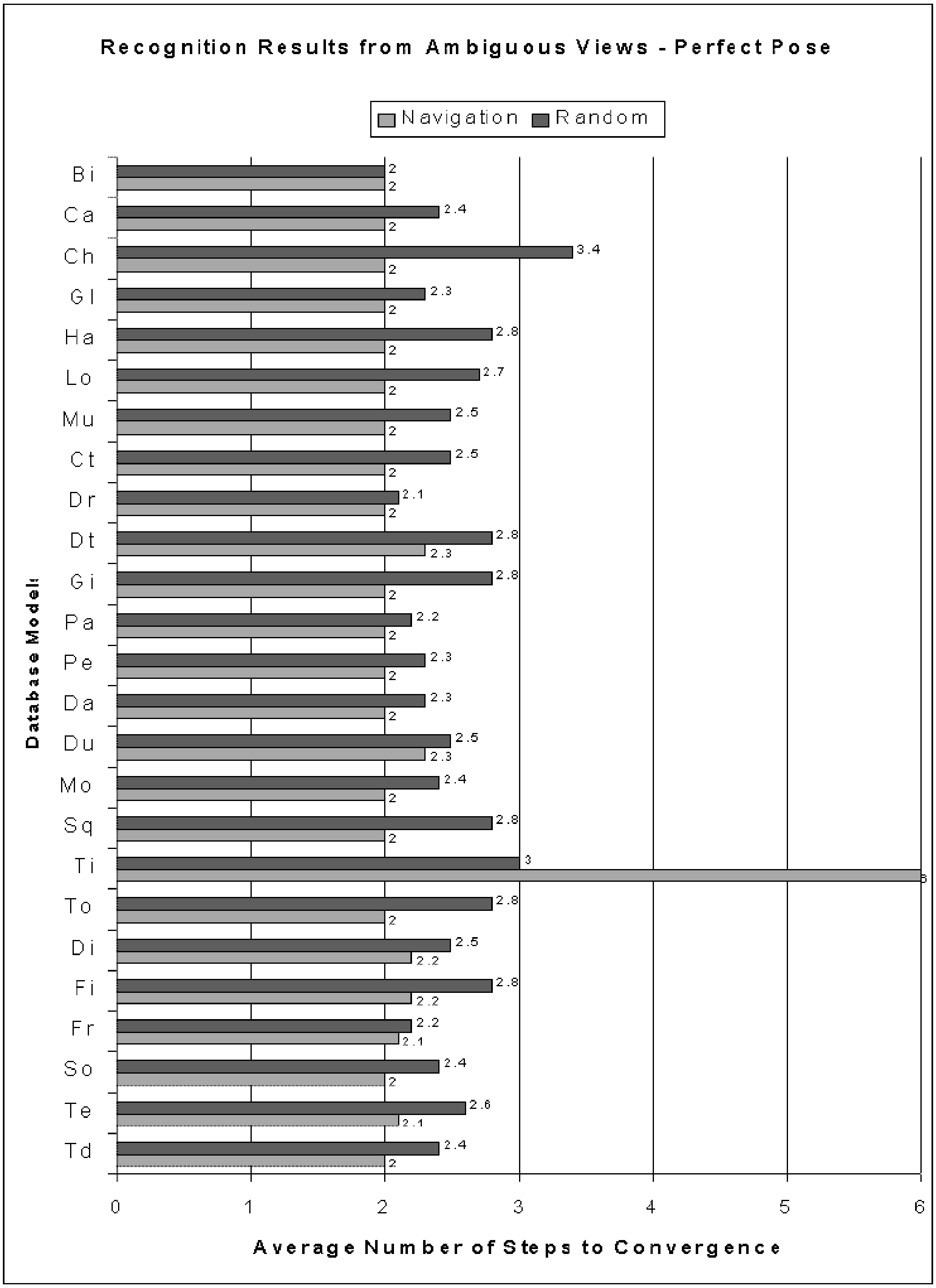


Figure D.3. Comparison of two strategies from ambiguous viewpoints (Perfect pose estimation).



The proposed strategy is compared to the random case in terms of the number of iterations that lead to convergence. Notice that here a step refers to an *algorithmic step*, i.e. a single viewpoint visited.

FIGURE D.4. Comparison of two strategies from ambiguous viewpoints (Perfect pose estimation).

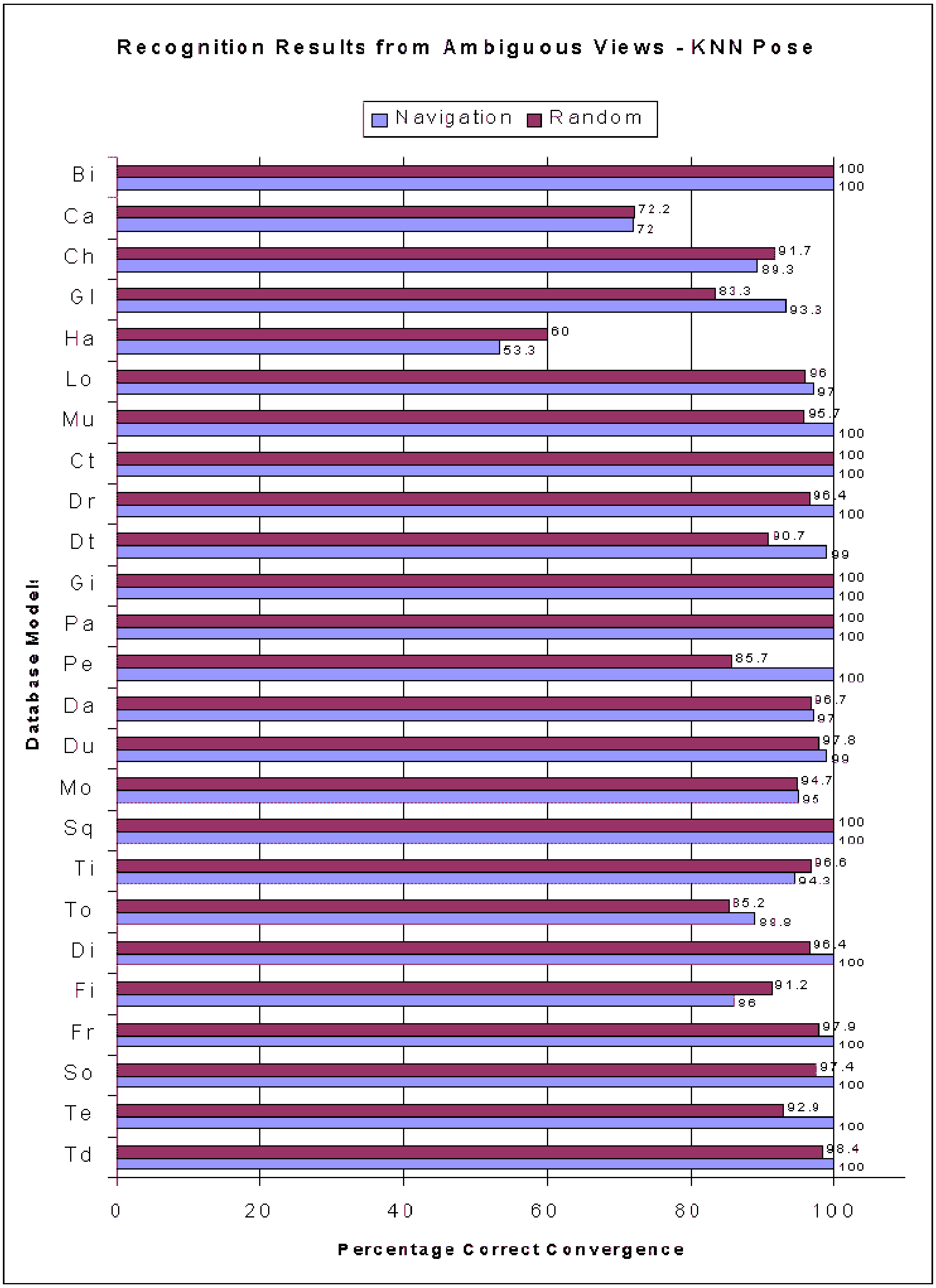


FIGURE D.5. Comparison of two strategies from ambiguous viewpoints (KNN pose estimation).

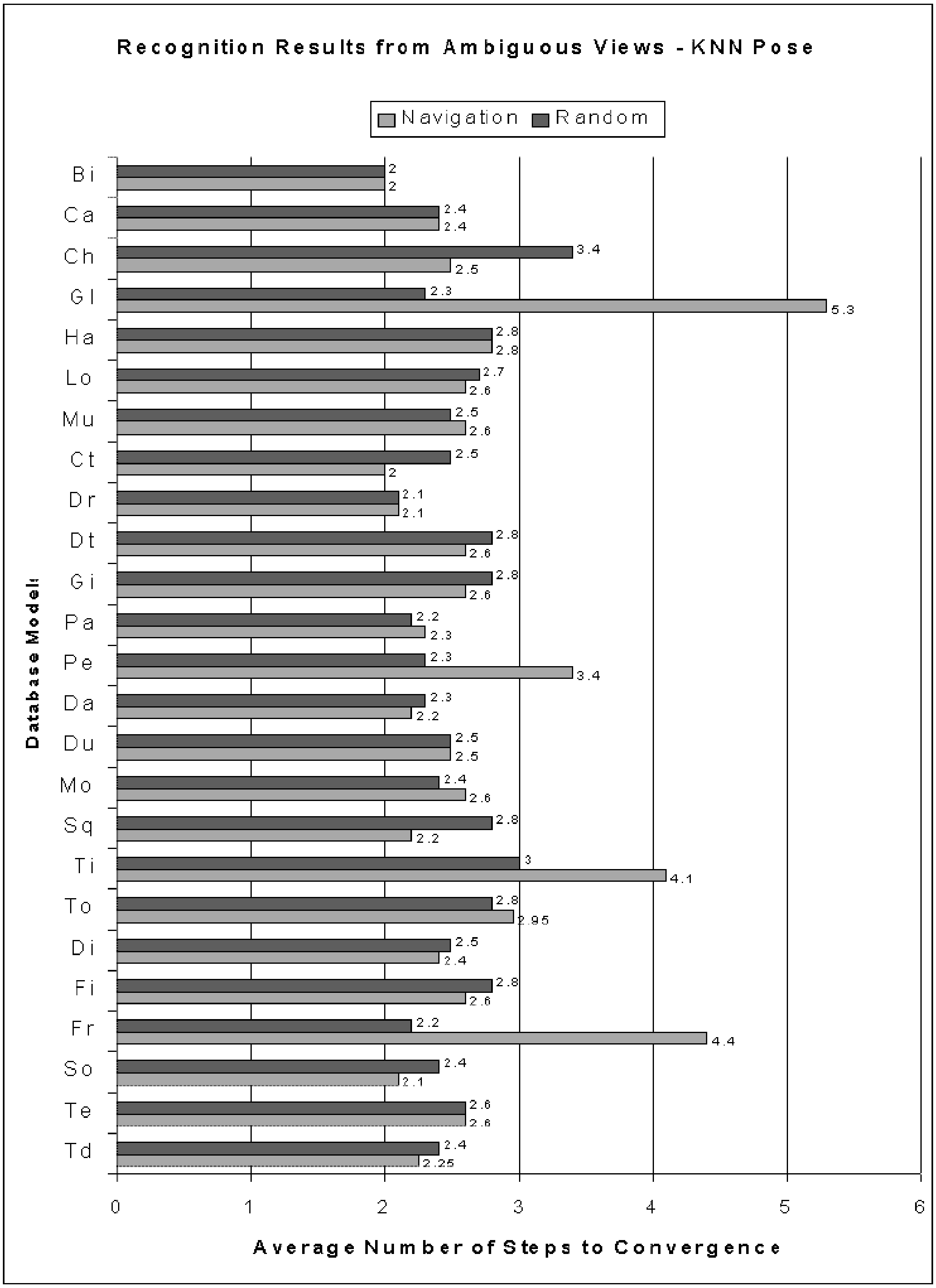


FIGURE D.6. Comparison of two strategies from ambiguous viewpoints (KNN pose estimation).

APPENDIX E

Entropy Map Breakdown

Here, one can find a detailed breakdown of the components of the entropy map of each object in the database. The maps were broken down into 4 components in accordance to their entropy values as follows:

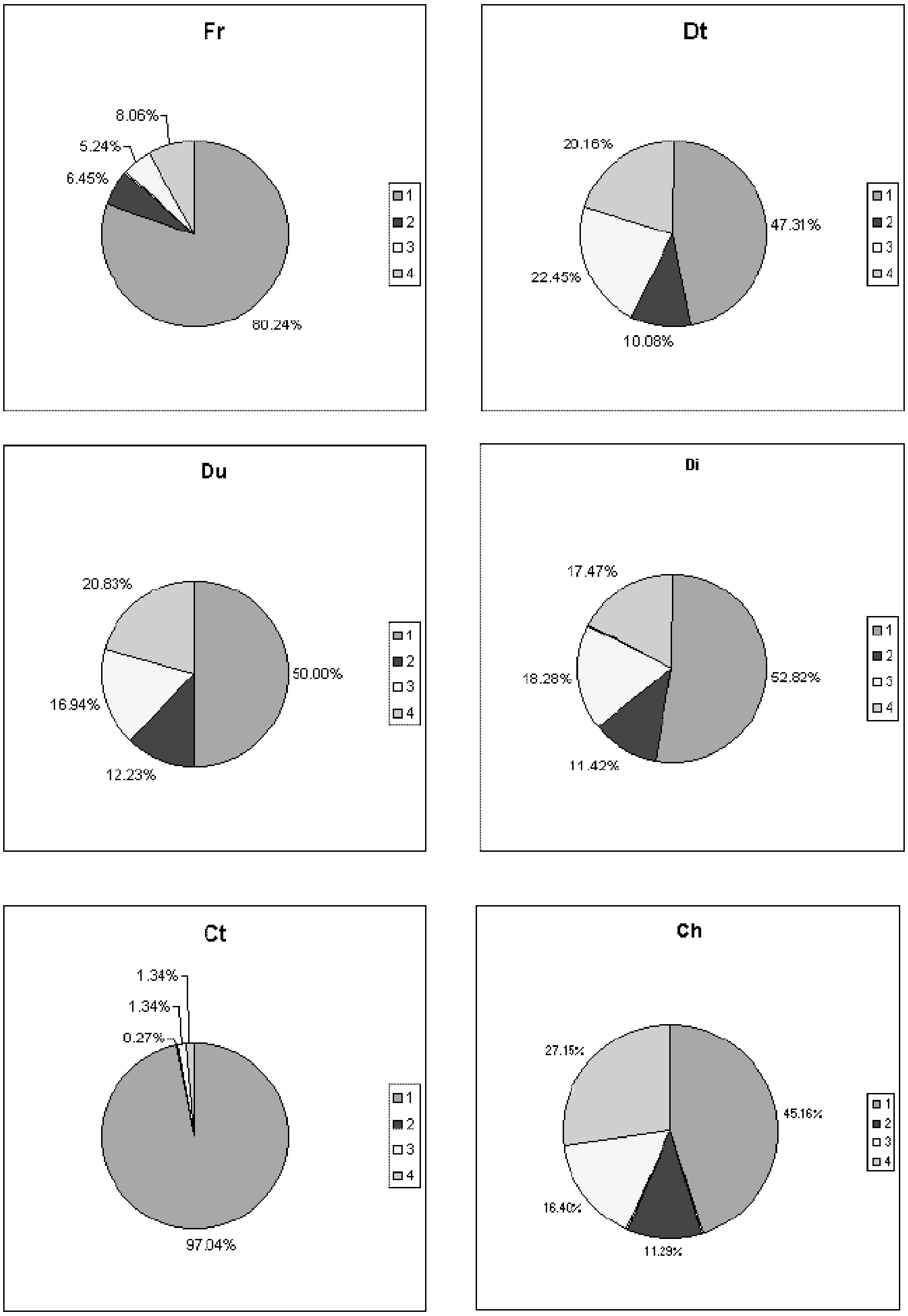
case1: non-ambiguous and correct identification,

case2: ambiguous and correct identification,

case3: ambiguous and incorrect identification,

case4: non-ambiguous and incorrect identification.

Here, an ambiguous viewpoint is defined to have an value exceeding 0.5.



Displayed above are pie charts indicating the breakdown of the entropy maps into categories: (1) non-ambiguous (entropy < 0.5) and correct identification, (2) ambiguous and correct identification, (3); ambiguous and incorrect identification, (4) non-ambiguous and incorrect identification.

FIGURE E.1. Breakdown of entropy maps.

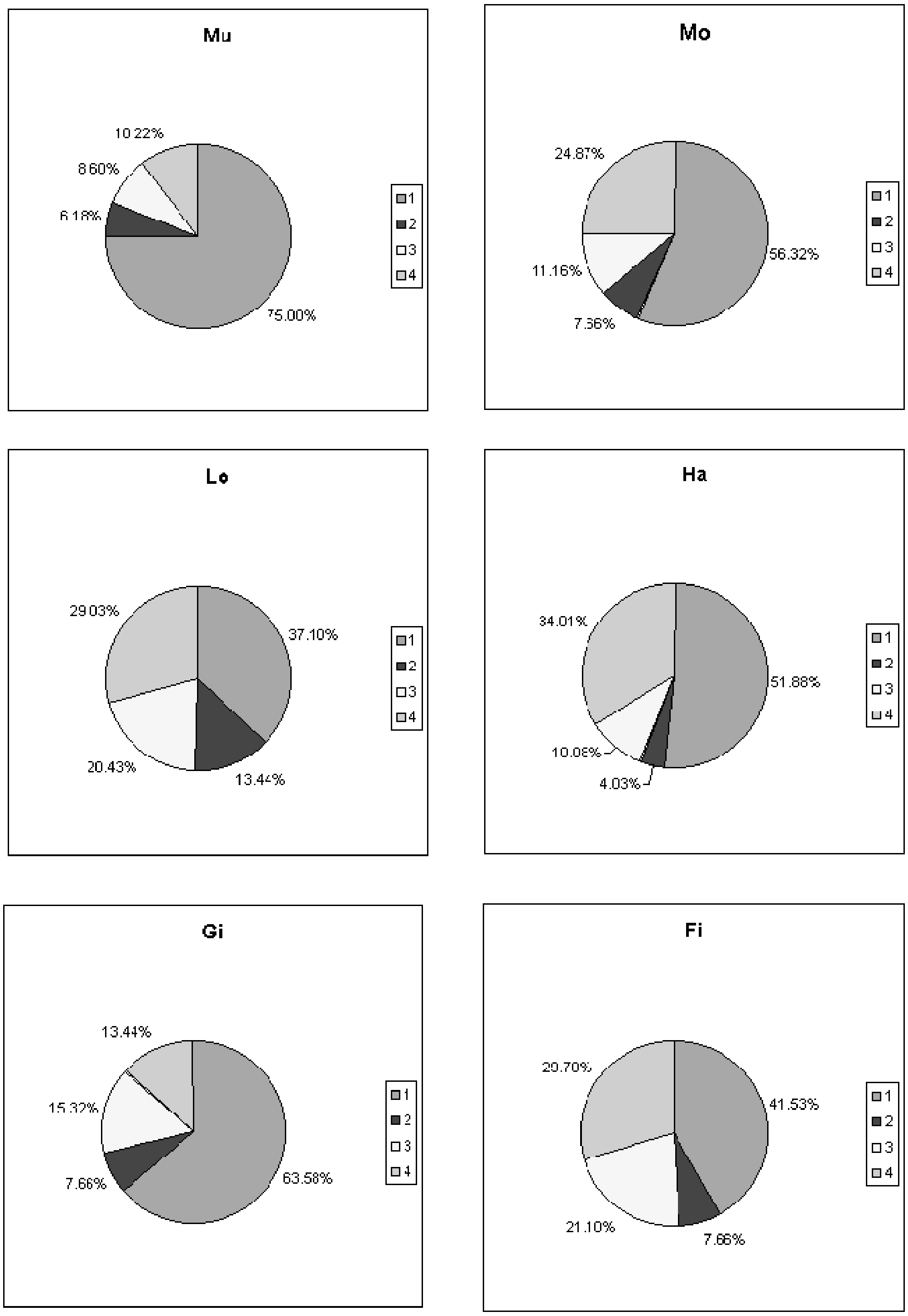


FIGURE E.2. Breakdown of entropy maps (cont.).

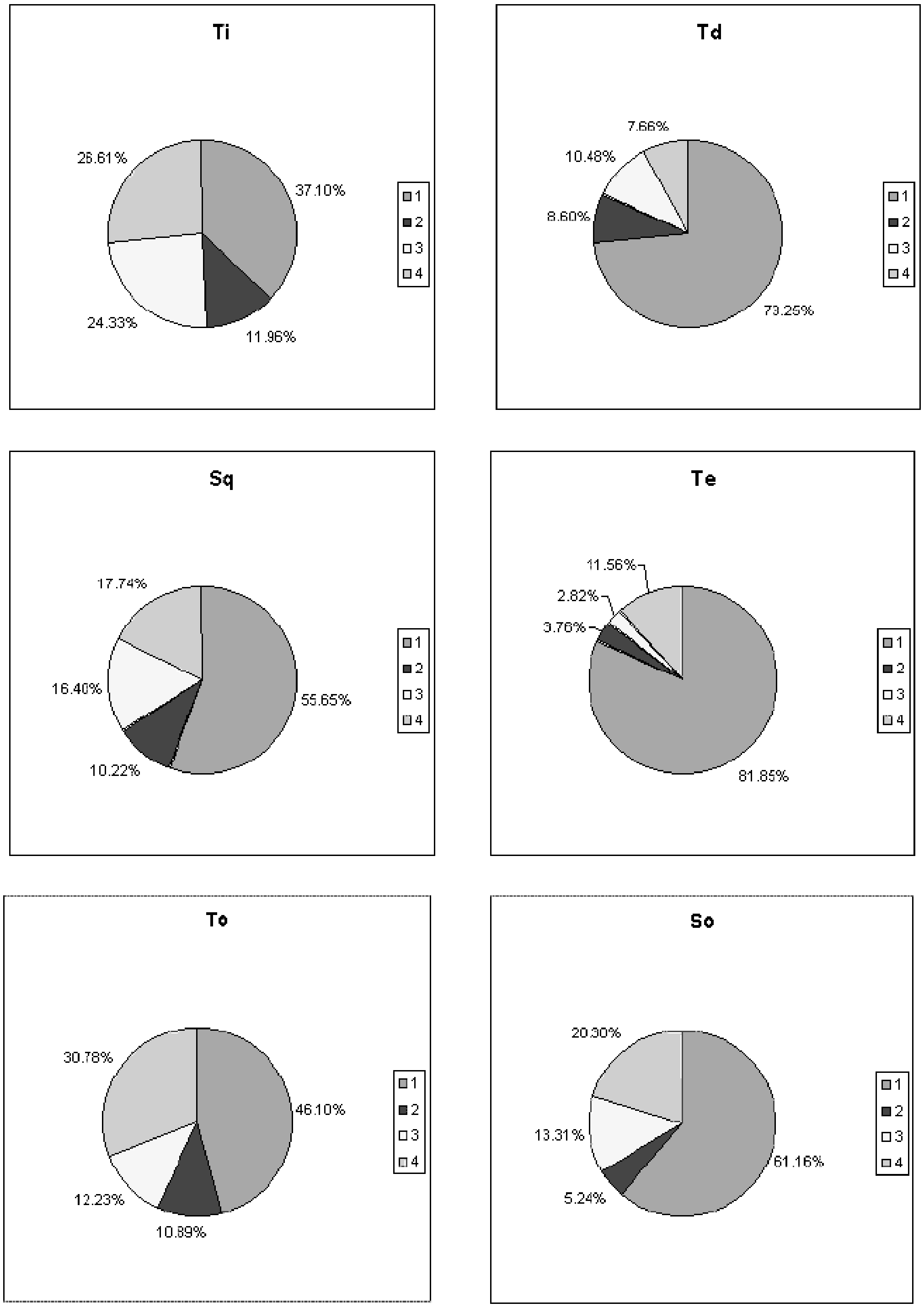


FIGURE E.3. Breakdown of entropy maps (cont.).

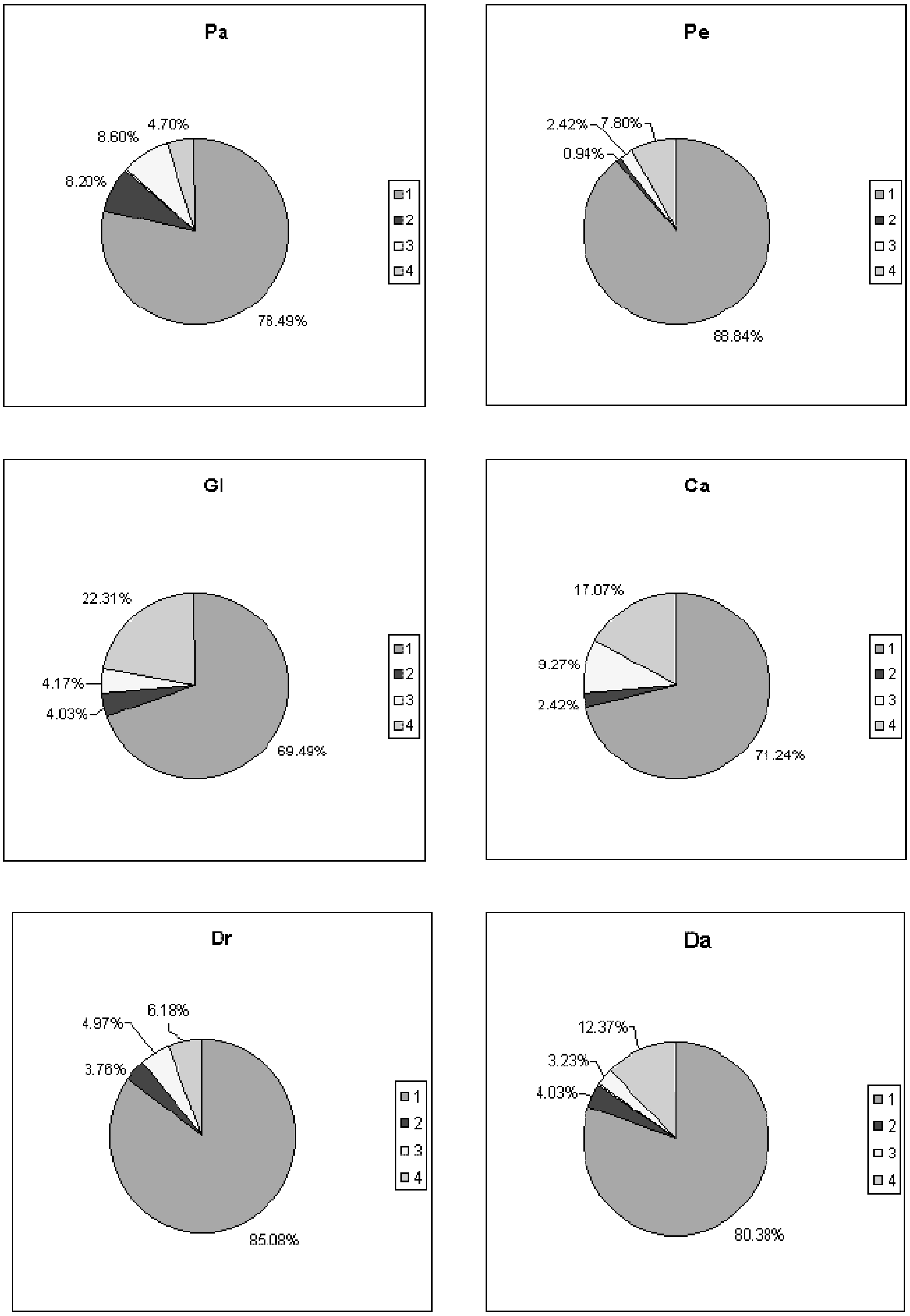


FIGURE E.4. Breakdown of entropy maps (cont.).

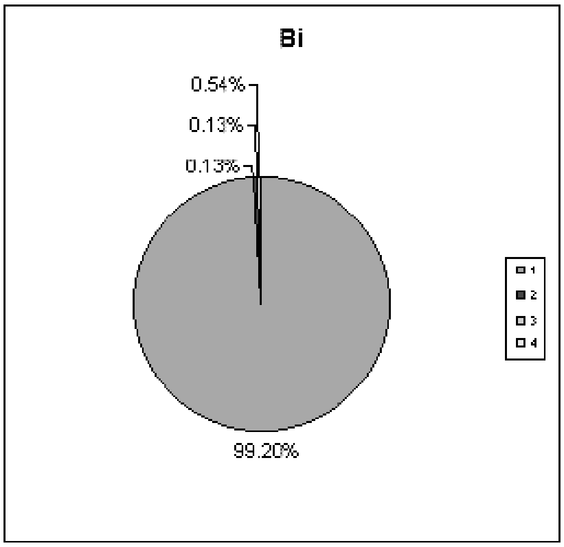


FIGURE E.5. Breakdown of entropy maps (cont.).

APPENDIX F

Average Pose Error Using Nearest Neighbour Approach

Below, one can find a chart describing the average pose error (in degrees) for each object in the database, using the nearest neighbour approach.

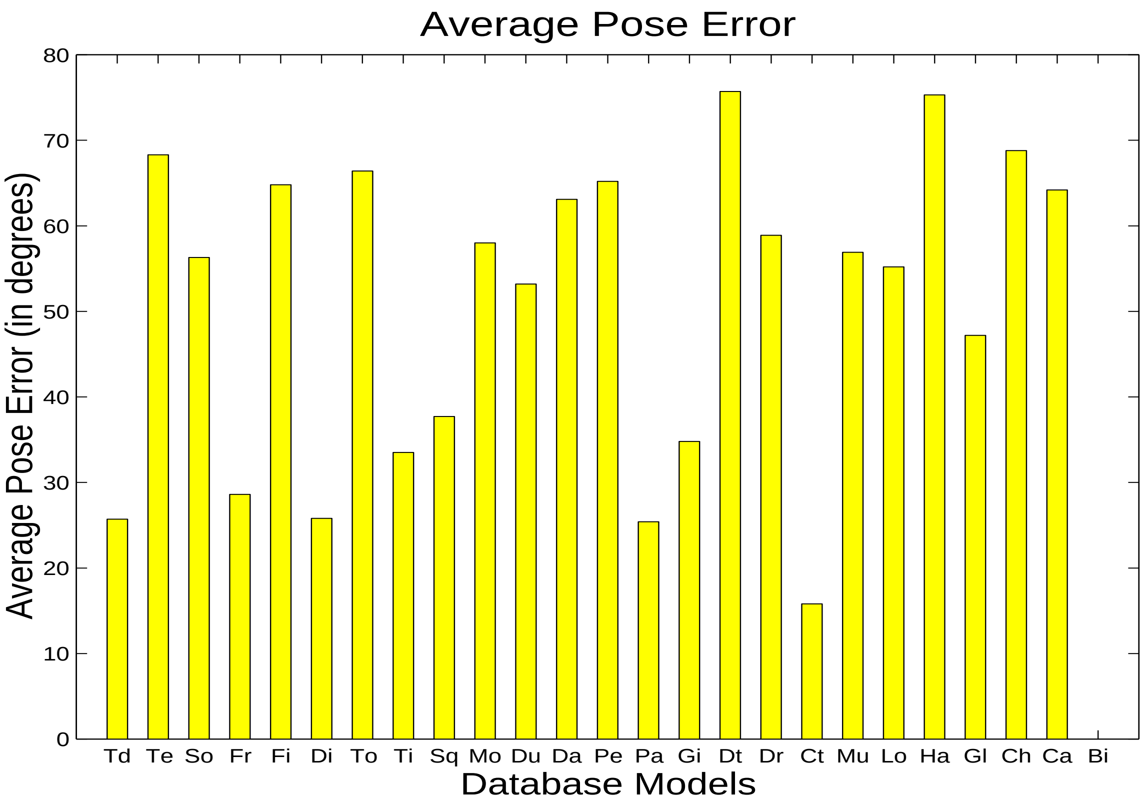


FIGURE F.1. Average pose error for each object in database.

Document Log:

Manuscript Version 0

Typeset by $\mathcal{AMS}$-LaTeX — 30 March 2000

Tal Arbel

Center for Intelligent Machines, McGill University, 3480 University St., Montréal (Québec) H3A 2A7, Canada, *Tel.* : (514) 398-2185

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

Typeset by $\mathcal{AMS}$-LaTeX