Robust Sequential View Planning for Object Recognition Using Multiple Cameras

ROBUST SEQUENTIAL VIEW PLANNING FOR OBJECT RECOGNITION USING MULTIPLE CAMERAS

BY

FOROUGH FARSHIDI, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

Master of Applied Science (2005) McMaster University

(Electrical & Computer Engineering) Hamilton, Ontario, Canada

TITLE:	Robust Sequential View Planning for Object Recognition Using Multiple Cameras
AUTHOR:	Forough Farshidi B.Sc., (Electrical Engineering) Sharif University of Technology, Tehran, Iran
SUPERVISORS:	Drs. Shahin Sirouspour and Thiagalingam Kirubarajan
NUMBER OF PAGES:	xi, 107

Abstract

In this thesis the problem of object recognition/pose estimation using active sensing is investigated. It is assumed that multiple cameras acquire images from different view angles of an object belonging to a set of *a priori* known objects. The eigenspace method is used to process the sensory observations and produce an abstract measurement vector. This step is necessary to avoid the manipulation of the original sensor data, i.e. large images, that can render the sensor modelling and matching process practically infeasible.

The eigenspace representation is known to have shortcomings in dealing with structured noise such as occlusion. To overcome this problem, models of occlusions and sensor noise have been incorporated into the probabilistic model of sensor/object to increase robustness with respect to such uncertainties. The active recognition algorithm has also been modified to consider the possibility of occlusion, as well as variation in the occlusion levels due to camera movements.

A recursive *Bayesian* state estimation problem is formulated to model the observation uncertainties through a probabilistic scheme. This enables us to identify the object and estimate its pose by fusing the information obtained from individual cameras. To this end, an extensive training step is performed, providing the system with the sensor model required for the Bayesian estimation. In order to enhance the quality

of the estimates and to reduce the number of images taken, we employ active real-time viewpoint planning strategies to position cameras. For that purpose, the positions of cameras are controlled based on two different statistical performance criteria, namely the *Mutual Information* (MI) and *Cramér-Rao Lower Bound* (CRLB).

A multi-camera active vision system has been developed in order to implement the ideas proposed in this thesis. Comparative Monte Carlo experiments conducted with the two-camera system demonstrate the effectiveness of the proposed methods in object classification/pose estimation in the presence of structured noise. Different concepts introduced in this work, i.e., the multi-camera data fusion, the occlusion modelling, and the active camera movement, all improve the recognition process significantly. Specifically, these approaches all increase the recognition rate, decrease the number of steps taken before recognition is completed, and enhance robustness with respect to partial occlusion considerably.

Acknowledgement

I would like to express my deepest gratitude to my supervisors. I specially thank Dr. Shahin Sirouspour for his endless guidance and support, and for being a wonderful source of scientific insight. I am grateful to Dr. Thiagalingam Kirubarajan for his support and guidance.

I thank my friends in ECE department at McMaster University, specially my colleagues in Telerobotics, Haptics and Computational Vision Lab., Ali Shahdi, Peyman Setoodeh and Mahyar Fotoohi Ghiam for their help and friendship.

I would most especially like to thank my husband, Amir, and my parents, to whom I dedicate this thesis, for their love, support and encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and Contributions

There is a growing demand for intelligent systems that are able to adapt to their surrounding environment and work with limited human assistance. Sensor-based object recognition and classification has been an active area of research in intelligent systems and machine vision. Vision, as the well-developed component of human learning system, has become one of the most popular sensing modalities in the robotic intelligent systems [1,2]. In space explorations, autonomous robots can utilize vision-based recognition techniques to identify and localize objects in their task environment and respond accordingly. Vision systems are also widely used in robotic assembly lines for part recognition and localization. In telerobotics, smart vision systems can provide best possible views of the remote task scene to the operator based on multiple cameras images and the perceived identity of the object to be manipulated.

The concepts of localizing, tracking, and recognition of objects or the robots themselves, have been the subject of previous research [3,4], and all can be considered as

inverse problems, where the extracted feature from the sensor input is employed to determine the real world condition that has generated it [5]. This is done by inferring the closest training-set model to the input sensory data through a state estimation approach. The position, trajectory, or identity of the objects are considered to be the unknown states of the system. These states are estimated using sensory observations, and based on prior information gathered in a database that models the characteristics of the system. The idea in state estimation is to reduce the ambiguities and uncertainties inherent to the environment and to help perform the task faster, more accurately, and free of failures.

In this thesis we intend to develop active 3D multi-camera object recognition algorithms, that are robust with respect to noise and occlusion in the camera images.

Employing a single image obtained from a fixed camera location for object recognition is straightforward and has been extensively studied before [4, 6, 7]. In such systems the possibility of false recognition will be high if the features available in the image are inadequate to effectively identify the object. The main sources of problem are ambiguity and uncertainty. Ambiguity could occur when different objects present similar features from particular viewpoints, or the extracted features are the same, either because of the sensor noise or the imprecision embedded in the feature extracting method. A good example is the error caused by omitting some of the eigenvectors in the statistical eigenspace method which results in the overlap between manifolds in the eigenspace. Consequently, more information is required to resolve such ambiguities. Uncertainty occurs as a result of environment and sensor noise, as well as non-adaptive thresholds that can corrupt the output of a feature detector. These problems can be solved by moving the camera, acquiring new images, and providing

the classifier with *easy to classify* views.

In contrast with random move-and-take strategies or multiple fixed cameras [8], optimal active object recognition algorithms automatically select sensor parameters/movements before collecting new images, such that the object ambiguity is resolved with the least number of images. They are often sequential decision making processes that terminate after collecting enough evidence so that the object identity/pose can be determined with high confidence.

In the process of active recognition, maintaining a framework that translates collected information into an indicative quantity is crucial for success. This flow of information has been modelled with a probabilistic fusion scheme in our work. An approach based on the Shannon's Information Theory uses the concept of entropy to quantify the amount of uncertainty in the object state, i.e., class and pose, given the available sensory information [9].

In this thesis, sensor uncertainties are modelled within a probabilistic framework and a recursive *Bayesian* estimation problem is formulated for object classification/pose estimation. Two performance indices are used to quantify the quality of observations in the context of state estimation and subsequently choose the next best positions of the cameras. The first measure is based on the Mutual Information (MI) which leads to a problem formulation similar to that of [9], but with expansion for multi-camera systems. The second metric is based on the classic *Cramér-Rao Lower Bound* (CRLB) which provides a theoretical lower bound on achievable *Mean Squares Error* (MSE) of the estimated state. The CRLB has been extensively used in sensor management algorithms [10,11]. This information theoretic metric contains performance limits inherent to the problem and independent of any specific solution.

In this thesis we use the eigenspace method to represent objects, a technique that has been employed by other researchers in the context of computer vision [3, 6, 8]. In this approach, a compact approximate encoding of a set of template images is done in terms of a small number of orthogonal basis images. The underlying philosophy of this method is that the shape and reflectance of rigid objects are constant intrinsic properties. Some of the advantages of this method are its ability to represent the objects images by a relatively small number of coefficients, its capability to consider the combined effects of shape, pose and illumination characteristics, as well as its use of two-dimensional images in the learning and recognition phases. On the other hand, this approach is sensitive to changes in pixel values caused by illumination changes, and has shortcomings in handling structured noise such as occlusions and outliers due to its global modelling scheme [12]. This problem will be addressed in this thesis, and solutions for improving the robustness with respect to occlusion will be given.

The main contributions of this manuscript can be summarized as follows:

- While prior relevant reports in the literature have mostly considered single-camera systems, we propose novel algorithms for active *multiple* camera object recognition. Active object recognition is formulated within a probabilistic framework which makes the fusion of information from multiple cameras possible. It is anticipated that obtaining multiple simultaneous views of the object reduces the number of steps required for ambiguity resolution and increases the recognition success rate. Such an approach is also expected to be more robust with respect to partial object occlusions as the images captured from multiple viewpoints are less likely to be all occluded.

- We introduce a new performance metric based on the CRLB and compare it

with the MI for camera positioning.

- To make the eigenspace representation approach more robust to occlusion, we propose to explicitly incorporate such an uncertainty into the sensor model in the database. To this end, the images in the training database are corrupted with randomly generated occlusions before being transformed to the eigenspace. This will generate uncertainty in the eigenspace coefficients that is modelled by Gaussian distributions. Also, an online recognition algorithm is developed which is robust with respect to occlusions and sensor noise. This is achieved by proposing a novel state estimation algorithm, where the percentages of the occlusions in the images are considered as part of the object states.

- Extensive experiments are conducted for 3D active object recognition/pose estimation in presence of occlusion, employing the proposed information theoretic approaches and a two-camera active vision system shown in Figure 1.1. The results of the experiments demonstrate the significant effectiveness of the proposed approaches in increasing the recognition rate, decreasing the number of sensor actions required to achieve recognition, as well as enhancing the robustness with respect to occlusion.

1.2 Organization of the Thesis

The rest of this Thesis is organized as follows. A review of prior relevant research is given in Chapter 2. In Chapter 3, the problem of multi-camera active object recognition/pose estimation is formulated. In Chapter 4, probabilistic sensor and occlusion modelling in the eigenspace domain are discussed. In Chapter 5, the performance

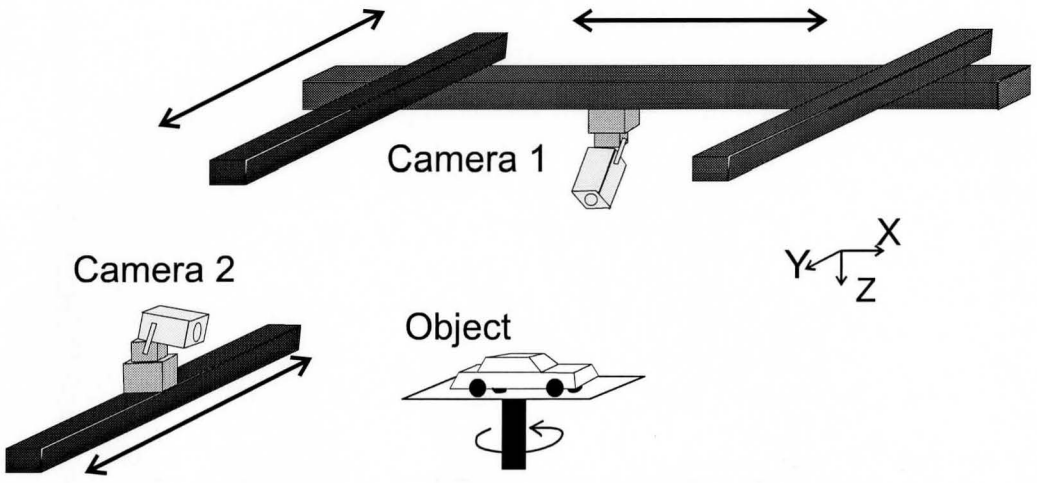

Figure 1.1: The schematic of the multi-camera vision system: Camera 1 moves in the xy plane; Camera 2 moves along the y axis. A turn-table mechanism positions the object at different angles. The bold arrows indicate movement along noted directions.

indices for camera parameter selection are presented. Chapter 6 is devoted to the description of our active vision system. Results of Monte Carlo experiments are given in Chapter 7. Finally, the thesis will be concluded in Chapter 8.

1.3 Related Publications

1. F. Farshidi, S. Sirouspour, T. Kirubarajan, "Active Multi-camera Object Recognition in Presence of Occlusion", in Proc. of IEEE/RSJ Int. Conf. of Intelligent Robots and Systems (IROS), pp. 3987-3992, 2005.

2. F. Farshidi, S. Sirouspour, T. Kirubarajan, "Robust Sequential View Planning for Object Recognition Using Multiple Cameras", submitted to IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, July 2005.

Chapter 2

Literature Survey

The aim of object recognition is to determine the object class and its pose based on information received from the sensors in the environment. No particular algorithm has yet been able to automatically learn and identify any arbitrary set of 3D objects. However, there has been a significant amount of research devoted to the problem of sensor-based object recognition and pose estimation, i.e. see [13–18]. These recognition approaches differ on several factors such as the type of the objects and sensors they are applied to, the features used to represent the objects, the matching methods used to compare these features to the ones in the database, and the decision making strategy in case of active sensor placement.

This chapter is laid out as follows. The first section describes different representation schemes employed in the computer vision literature; the second section outlines various feature matching methods; in the third section the existing information fusing techniques in literature are explored; the fourth section provides a summary of a number of active vision methods; finally a brief review on applications of Cramér-Rao Lower Bound (CRLB) is given in section five.

2.1 Representation Schemes

Ideally, to solve a large number of problems in computer vision, we need to represent our knowledge of the world in an efficient manner. The raw information received from sensors, e.g. images, are usually large to store and hard to compare. Consequently different approaches have been proposed to convert the raw data to features that are easier to manage. There has been a lot of effort in finding features that can encode the object identity efficiently [6, 7, 15, 19]. Such features must capture informative characteristics of different objects, be as discriminative as possible, and be detectable by the sensors used in different applications. At the same time, they should be expandable to different object sets and have manageable memory storage and computational power requirements. The studied features vary from the model-based ones which interpret the three dimensional (3D) shape to the appearance-based ones which employ the two-dimensional appearance of the object.

A large group of object/sensor modelling techniques employ model-based features that interpret the 3D shape of the object. This includes methods based on wire frame representation, constructive solid geometry, spatial occupancy, surface boundary representations, and aspect graphs [5, 12, 20]. In contrast, appearance-based techniques use features from the 2D appearance of the object. The appearance-based eigenspace approach that belongs to this latter category has attracted a lot of interest in the computer vision literature [3, 6, 8].

A summary of the most popular features used in the literature is given below.

2.1.1 Model-Based Representation

High-level model-based features used in the literature are usually based on intrinsic properties of parametric models, which yield stable features like algebraic surfaces and superquadrics [20–22]. These representation methods, however, are only applicable to specific classes of objects viewed from a few predetermined viewpoints. Low-level model-based features tend to match edges, curves, silhouettes, holes, zeros of curvature, and other 2D perceptual structures [23–25]. The intermediate model-based features use the characteristics of surface patches. Such features could incorporate the surface area or type (cylindrical, planar or spherical), surface normal, direction of axes of surface, centroid, and centers of spheres [26–29]. The low and intermediate features usually suffer from being sensitive to noise, unstable, and view-variant.

2.1.2 Part-Based Representation

Some researchers have considered recognizing objects through identification of their parts by partitioning every object to a collection of parts. The concept of volumetric primitives or geons (geometric ions) was first proposed by [30] for part-based recognition. In this approach, the Cartesian product of contractive shape properties yields a set of volumetric primitives. This idea has been further explored by [7, 31] in constructing single view object recognition systems. Also, an active vision system employing this feature has been introduced in [32].

Another approach for part-based recognition has employed appearance-based parts by optimally (in a Minimum Description Length sense) partitioning the image through approximating its regions using polynomial surfaces [33]. Such strategies have the advantage of being more robust with respect to occlusion and segmentation variations

compared to the whole object representation.

2.1.3 Aspect Graphs

Aspect graphs are another popular tool for representing objects in active recognition [34–39]. Aspects are defined as topologically equivalent classes of object appearances. Most relevant publications present the entire object as a set of aspects, each of which define a topologically different view of its surfaces [25]. There are also papers that use aspects to represent a small set of volumetric primitives from which each object in the data base can be reconstructed [7]. Unfortunately, these methods require large storage memory and long construction time.

2.1.4 Multidimensional Receptive Field Histograms

Multidimensional Receptive Field Histograms have also been used to describe real world objects [4]. The idea is that local structures are the key to describe objects. In [40], inspired by the previous research on the color histograms, the authors have developed a similar approach using local descriptions calculated through a vector of linear receptive fields. They use the joint statistics of these local neighborhood operators (receptive fields) to represent object appearance. For that purpose, the multidimensional receptive field histograms approximate the probability density function of the local appearances. This approach has been employed in an active recognition system in [41]. Another close metric is the Multimodal Neighborhood Signature applied to object modelling and recognition in [19].

2.1.5 Appearance-Based Parametric Eigenspace

Many recognition algorithms have successfully employed eigen-representation approaches to model the information acquired from the task environment. According to [17], such techniques can be classified into three groups, namely, feature-based eigen-shapes [42, 43], physically-based eigen-snakes [22, 44], and eigen-pictures [6, 9, 45–47] which are all based on the Karhunen-Loeve transformation. The rest of this section is devoted to a summary of the parametric eigen-picture representation.

Appearance-based eigenspace based on 2D images has been vastly employed for object modelling by prior researchers [48–51]. The basis of this approach is that the shape and reflectance properties of rigid objects are constant intrinsic properties, whereas pose and illumination are environment-dependent. The high-dimensional sensor pattern is projected to a discriminative lower dimensional subspace using the Principle Component Analysis (PCA). The close points in this subspace correspond to similar appearances of the objects. Some of the advantages of this method are its representation of the images by a few number of coefficients, its capability to consider the combined effects of shape, pose, reflectance characteristics and illumination, as well as its enabling of off-line learning and recognition through two-dimensional digital images [12]. The feature-extraction time is minimal in the parametric eigenspace methodology, compared to some of model-based methods that can become exhaustive in complex scenes.

The majority of applications based on eigenspace approach employ the gray-scale images to represent templates. In this case the lighting conditions and background have to be controlled. However in [52], this difficulty has been overcome by using

direct surface measurements instead. The authors in this work, adopt the appearance-based recognition to handle range (shape) data. The database includes some eigen-surfaces representing overall object shapes. This method has its own shortcomings as this type of data is more difficult to handle.

(i) Computation Burden:

One drawback of the eigenspace method is its sensitivity to change in pixel values, due to positioning or segmentation error, sensor noise, rotation of the image plane, illumination change or occlusion caused by other objects present in the environment. There are two basic ways to deal with this problem. The first one is to perform a scale and brightness normalization on each image prior to projection onto the eigenspace. In case of constant brightness background, a simple thresholding operation can be applied [53], whereas for moving objects spatio-temporal image filters are more appropriate to recover the moving region [6]. Alternatively an adaptive strategy has been developed in [54].

In this approach the training set is constructed by considering all changes possible. The excessive number of training images taken in different poses and illumination conditions can result in a significant computational cost for calculating the eigen-images and render the method practically infeasible [53, 55].

A solution has been proposed to automatically select an optimal representative subset of views of objects, while generating the eigenspace. In [56], this process is performed online in both the training and recognition phases. The system moves the camera, takes images and evaluates the information immediately. Images are added to the subset based on the concept of Saliency, which is motivated by the definition of *Distance From Face Space* (DFFS) used to reject non-face images in [57]. Saliency

is defined as the information found in the image relative to the representation. In the eigenspace method, this information is the residual error, i.e. the energy not captured by the current basis set. During training, a high saliency indicates the need for an update in the basis, while in recognition, low saliency means the image is one of the images in the training set. A subset that entails most of the information is chosen and the initial eigenspace is constructed. The addition of any new image to this set is only required in case of high saliency. The system stores only the coefficients for the selected images and actively searches for the corresponding views in the recognition phase. Mathematically, the subset selection approach tries to rearrange the images such that the first n images are sufficient to represent the entire set, which is in contrast to the Batch method, the standard solution that finds the Singular Value Decomposition (SVD) of the complete set of images and then truncates it. For that purpose, *Greedy Algorithm* [56] chooses the average image as the initial representation. After that, in each step the algorithm selects the column most independent from the basis. The process of updating the SVD of a matrix after adding a column to it will be performed according to [58].

Another effort in simplifying the eigenspace method is [48], in which the authors propose a practical algorithm to ease the computational burden of finding the eigenvectors of a set of real world images. They propose a Spatial Temporal Adaptive (STA) method that reduces the computational demand of the approximate partial eigenvalue decomposition based on image encoding. First, spatial temporal encoding is applied to reduce the computation and storage requirements; SVD is implemented then to extract the eigenvectors. The advantage of this method over the traditional SVD algorithm becomes more evident with the increase in correlation within one

image or between subsequent images in the set.

(ii) Online Training:

A limitation of the eigenspace method stems from the fact that building an eigenspace in the off-line learning phase requires the entire raw images acquired from object templates to be present in the database. In some applications we may need to update the available object set to a larger one, or perform online training. Adding a new set of templates to the existing set without updating the eigenspace would result in inaccuracy. Efficient algorithms have been developed to eliminate this problem and extend the eigenspace without expensive computational burden [58, 59].

In [59], an approach is utilized to tackle the computational complexity of the Karhunen-Loeve (KL) transform. The Sequential Karhunen-Loeve (SKL) algorithm is proposed to calculate the KL basis. The SKL involves sequential singular value decomposition steps, yielding a low dimensional KL basis of an image sequence. At each updating step, the number of images which should be added are optimized. This method has less memory storage requirement and is faster than the Batch KL algorithm, which uses the SVD of the whole set of images to construct the eigenvector set and then truncates it. The advantages of this method are most evident in processing sequences of images where the computational delay is minimized and the database can be updated dynamically.

(iii) Illumination Sensitivity:

To cope with the illumination sensitivity problem in the eigenspace modelling strategy, experiments have been conducted in [53], evaluating a novel approach to the problem of robust object recognition by illumination planning. The goal is to

set the illumination for which the objects are most distinguishable, while measuring the similarity between objects by the correlation concept. Objects are represented by individual manifolds in each illumination condition. The similarity of two objects is determined by the minimum distance between their manifolds. The optimal light source direction is then defined by maximizing the minimum distance between the manifolds. In this approach the sensor parameters used in the recognition step are predetermined through the off-line training stage.

Reference [60] also attempts to reduce the sensitivity of the eigenspace method to the illumination condition. In [61], it has been shown that an illumination insensitive measure can be developed and used for probabilistic object recognition. Based on this and other studies, the authors in [60] have considered edges and gradients as suitable measures in coping with illumination changes. They propose a method based on eigenimages filtered by gradient-based operators, and demonstrate how to extract eigenspace coefficients from responses of local filter banks in a global eigenspace representation. They have applied this method to object recognition in case of multiple objects in different poses [60].

(iv) Sensitivity to Occlusion:

It is generally difficult to handle Occlusions, as they are usually unpredictable and random. In particular, the PCA approach is not capable of coping with structured noise such as occlusions due to its global modelling scheme. Modified versions of the method have been developed to increase robustness with respect to such uncertainties. Reference [62] has proposed a new way of calculating the coefficients of the eigenvectors to handle occlusions, outliers and varying background. Instead of projecting the entire image into the eigenspace, the authors in [62] use subsets

of image pixels to extract the coefficients through a hypothesize-and-test paradigm, while the input image in the recognition phase may contain more than one instance of the training objects. At each location in the image, multiple hypotheses are generated. The Minimum Description Length (MDL) principle is then utilized to compare different hypothesis, and selects the best ones. The output will be the number of recognized objects in the image, and the corresponding eigenspace coefficients. In [63], a robust hierarchical form of the Kalman filter derived from the Minimum Length Description(MDL) principle is implemented. Each of the hierarchical levels predicts the recognition state at a lower level and modifies its own recognition state employing the residual error between the actual lower level state and the prediction. Eventually, the filter learns an internal model of input dynamics through adapting the generative and state transition matrices at each level in order to minimize prediction errors.

Reference [3] applies the parametric eigenspace encoding technique to object tracking. The authors of this paper use and extend the eigenspace approach, robust estimation techniques and parameterized optical flow estimation (representing image motion in terms of low order polynomials). To make matching in eigenspace method more robust to outliers and occlusion, the traditional image transformation stage has been replaced by a robust estimation technique. The dot product of the image vector and the eigenvectors matrix used in the standard eigenspace method to compute the eigenspace coefficients, is equivalent to solving the problem of least squares error between the original and the reconstructed image. This approximation has been reformulated into the problem of minimizing alternative robust error norms. It is argued that these functions can tolerate outliers and occlusions better. Another novelty in

this work, is representing a limited set of canonical views, and then allowing a parameterized transformation between an image and the eigenspace. This would resolve the need to store images from *all* views of an object and *all* the viewpoints. The authors seek a *multiple views plus transformation* model of object recognition [64], meaning the matching includes both estimation of the view of object and the transformation that takes this view into the image. This is depicted within a robust estimation framework which allows both the view and the transformation to be calculated. Similar to the *brightness constancy assumption* in optical flow estimation, a *subspace constancy assumption* is defined for each object view between the eigenspace and the image. This allows for conducting experiments employing parameterized optical flow methods in recovering the transformation between the image and the eigenspace.

(v) Applications of the Parametric Eigenspace Method:

The eigenspace method has been applied for the recognition of objects [3, 8, 65], faces [57, 66–68], facial expressions [69], as well as lip reading [70–72] and other applications [3, 73]. In [74], the authors initially derived a technique to efficiently distinguish and characterize a set of well-defined patterns in the images of human face. They chose the eigenfunctions of the averaged covariance of the data set as the best coordinate system to span the data space, and named them *eigen-pictures*. In their later work [67], the authors focused on the extraction of natural symmetries (mirror images) for face patterns within the Karhunen-Loeve framework.

Motivated by the two above noted papers, Reference [57] suggested to employ such efficient representation in encoding faces and comparing the images in the recognition process. The authors localize, track, and finally recognize the subject person's head via describing the faces by a small set of 2D characteristic views instead of interpreting

the 3D geometry. The Principle Component Value encodes the variation between face images, where each eigenvector is interpreted as a *ghostly* face or an *Eigenface*. In the recognition phase, if the input image is not sufficiently close to the face space it will be rejected as a face, otherwise it will be categorized as a known or unknown person. The authors also consider updating the eigenfaces during the recognition to enhance the capability of the database.

In [66], facial texture and structure are recovered from a real-time video stream while a statistical model represents the correlation between the texture of face and its 3D structure. This model is generated as a soft mixture of eigen-feature selectors that span the variation in texture and 3D structure across a training set of laser scanned faces. The face is tracked to detect and stabilize live facial images into a canonical 3D pose. The imperfections are resolved by processing this canonical texture with the extracted statistical model.

Murase and Nayar [6] proposed to extend the eigenspace methodology used for face recognition to object recognition and pose estimation. They argue that the 2D appearance of an object depends on shape, reflectance properties, pose and illumination. The first two are intrinsic properties of the object whereas the others are dependent on the scene. This suggests the representation of 2D appearance by pose and illumination. The idea has become increasingly popular, due to its speed and ease of use. In order to enhance robustness, a pre-segmentation step will be performed before projecting the images to the eigenspace. In this work, the spatial temporal adaptive (STA) algorithm proposed in [48] has been used to generate the eigenspace.

In Reference [75], the authors have developed a face recognition system that is

insensitive to facial expressions and large variations in lightning condition. They linearly project images into a subspace based on Fisher's Linear Discriminant. This will produce well-separated classes in the low dimensional space even in the case of lighting variations or facial expressions. The computational requirements in this approach are very similar to those of the eigenfaces method. Although, the experiments have shown that this framework has a lower error rate compared to the eigenfaces method [75].

2.2 Matching Methods

All recognition algorithms require matching strategies to establish correspondence between the database and the features extracted from an unknown object. Depending on the features representing the objects, the task of this strategy could become crucial. When employing robust global features the matching process is usually straightforward as the feature strongly rules out the class implicitly. On the other hand, using simple, low-level features will do little to reduce the uncertainty on its own, demanding a more complicated matching scheme.

There has been a significant amount of research aiming to reduce the complexity of the database search in the recognition phase [28, 76]. The resulting techniques often involve extensive preprocessing on the database information instead of performing complicated inference during the recognition phase [77]. For example the *Geometric Hashing* method has been used for recognition of partially occluded objects in [78]. In this approach, a redundant representation is generated for each object through affine preprocessing and for all transformation-invariant coordinate frames. These representations are calculated again in the recognition phase for all the coordinate

frames specified by the image and then compared to the database [79].

One of the oldest matching methods is the relational approach [80] in which objects are described as graphs, with the nodes being the features and the arcs being the geometric relation among those features. The correspondence between the object and database is achieved through matching these graphs. In [81], the matching time has been reduced by using low, intermediate and global level features. The modified algorithm searches for the model graph with the largest set of matched nodes.

Another category in matching strategies is the Interpretation Tree (IT) search [26]. The IT incorporates all possible sequences of correspondence between model surface and scene surface. Each route from the root to a leaf offers a different solution to the matching problem. At the recognition phase, a sequence of IT searches are performed for each object in order to find the path that embodies a consistent matching between object and model. Reducing the search time by pruning the tree is critical in this method. The basic idea in pruning is that if an interpretation is rejected, any path rooted at that interpretation will be invalid. Different approaches have been developed to prune the tree, including constraining the range of unary (e.g. length of an edge) or binary (e.g. angle between normal vectors) feature values [23, 24, 26, 82].

A novel technique to find the pose of an object is developed in [83]. This method is based on the information theory, specifically a formulation of the Mutual Information between the model and image. The alignment of two images is carried out by maximizing the Mutual Information, while the two images do not necessarily originate from the same real world template. This approach is robust with respect to illumination changes, and requires no information about the surface properties of object, besides its shape. However, this theory becomes complicated and expensive for

implementation. Extensive experiments have been conducted to evaluate the effectiveness of the proposed approach, for example in object tracking and photometric stereo [83]. In these experiments the stochastic optimization algorithm Empirical Entropy Manipulation and Analysis (EMMA) and Parzen window densities have been employed.

In the appearance-based eigenspace method, the problem of feature matching is reduced to appearance matching instead of shape matching. Considering the sensitivity of this transformation to pixel errors, it is instructive to embed the associated uncertainties through including more observations in the training set. As a result, every object may be modelled by a *cloud* of close points in the subspace. Different techniques have been proposed to match the unknown object to one of the templates. One approach is to compare the point in the eigenspace corresponding to the observed image to all the points in the database, and choose the closest *cloud* [84]. The unknown object identity is then hypothesized as the object category corresponding to the closet cloud from the projection of the image in the eigenspace.

Based on the fact that the images are highly correlated, eigenspace points corresponding to the same object class are assumed to be dependent and be in a k-dimensional hypersurface [6]. This hypersurface includes all possible poses and illumination directions of an object, and are produced by interpolating these points, offering higher precision than that of the *cloud* representation. In [6], the standard cubic spline interpolation is used to generate continuous functions of pose and illumination. Also, each object is represented by a continuous hypersurface in its own eigenspace. The continuous parametrization is chosen to allow the recognition of objects in any pose, even those not included in the database. In the recognition phase,

the object with the minimum distance between its hypersurface and the subject point is chosen. The object category is accepted as the identity of the unknown object, only if the distance is within a predefined threshold defined by the sensor noise. In case of success, the image is then transformed to the private eigenspace of the object to ensure precise pose and illumination estimation. Using similar concepts, the authors in [85] employ manifolds to learn the lip sequences, whereas in [86] the hand gestures are modelled by trajectories in the eigenspace. In [87], the hypersurfaces are used in guiding a robot to maintain a particular view of a moving object through actively tracking it.

In [87], the manifolds are densely resampled and stored as a large set of k-dimensional points. To find the closest neighbor of an arbitrary point among a set of points, an exhaustive search can be performed on all candidate points by calculating their distances from the point. However this could be inefficient unless the number of points is limited. Alternatively, binary search or indexing can be conducted. Reference [88] has proposed a simple algorithm for nearest neighbor search in high dimensions, e.g. more than 25. The nearest point search can also be accomplished by training a regularization network like the type described in [89].

The authors in [90], interpolate the images in the image space to produce new training samples. This is applicable when not enough training images are available or a higher precision is demanded. It can also replace the interpolation of the points in the eigenspace [87]. All images will then be transformed to the eigenspace. It is shown that this approach has a superior performance compared with the approaches that interpolate the eigenspace points corresponding to a sequence of object images, in an attempt to generate new sample points.

The primary drawback in appearance matching is the need for a comprehensive training stage, i.e. acquisition of large image sets, deriving eigenspace from a large covariance matrix and building the manifolds. The efficiency of this training stage is evaluated by the number of sample images required to build accurate manifolds. The minimum number of required sample images can be determined by studying the structure of manifolds, which is closely related to the geometric and reflectance characteristics of the object. Although the manifold structure has a tight relation with the object shape, the dramatic effects of illumination condition should not be underestimated. Reference [91] has demonstrated that in the case of face recognition, illumination variations may cause more changes in the projected image than changing the subject itself. Fortunately, the changes associated with reflectance are possible to analyze. In [92], a closed-form relationship between illumination parameters and manifold structure is established under certain reflectance assumptions. In this work, the class of linear reflectance functions are employed, considering the fact that the eigenspaces are linear subspaces. It is shown that for this particular reflectance class, the structure of the illumination manifold can be completely determined from a small number of sample images. For the specific example of Lambertian surfaces with arbitrary texture, the dimensionality of the manifold is exactly three, therefore an entire illumination manifold can be generated from only three sample images [92].

The concept of manifold has been extended to maximum likelihood formulation or probability densities in the eigenspace in several different prior reports [9,93,94]. This statistical approach is based on the *closed world assumption*, meaning the input to the system is always generated by one of the templates. This assumption allows one to assign a probability distribution over object classes to each point in the eigenspace.

The likelihood of the noisy observation in each camera position is estimated as a Gaussian function.

Reference [6] has defined a specific eigenspace for each object class separately in addition to a global eigenspace, so that the pose estimation can be performed more accurately after the class is identified. Reference [55] has proposed a method for fast computation of the normalized correlation of multiple rotated templates using multiresolution eigenimages. The location and orientation of the object can be detected faster than the conventional template matching. A multiresolution image structure is used to reduce the number of rotated templates and location search area. The position and angle are coarsely obtained in a wide region for the lower resolution image. These results are then used to reduce the search area for the position, and limit the range of rotation angle of the templates to smaller neighborhoods, at the next layer.

The key to successful matching in the eigenspace approach is attaining a robust means to estimate the object position and scale in an input image, prior to projection into the eigenspace. Different strategies have been considered for this purpose. In [90], the object is detected by simple thresholding, whereas in [93] another feature detection process is employed. In [51], global search has been performed and it has been shown that eigenspace matching can be employed to perform global search under translation. It is done by comparing the eigenspace with the local subimages extracted from the input image in every image location. To incorporate scale, this idea was later extended by matching the input image at different scales using a standard eigenspace approach [93]. These exhaustive techniques can provide a coarse initial guess about the transformation in the input image. Such primary coarse search results can be refined by the novel approach proposed in [3], which is

a continuous optimization technique. As an alternative, Reference [63] proposes to estimate the eigenspace points through a Kalman filtering technique. Another work in this area, [95], is concerned with matching scaled images through a specific way of defining the eigenspace points [96]. The robustness of this approach for transforming convolved and subsampled images is demonstrated through experiments [95].

2.3 Information Fusion

The most popular methods for quantifying uncertainty, extracting, and fusing the information in active recognition structures are the Probability Theory [9, 94], the Dempster and Shafer Theory of Evidence [35], and the Fuzzy Logic [97].

Probability-based representation schemes are Bayes Nets, also known as Bayesian, Belief, or probabilistic networks. A Bayes Net [32], far more general than a 3D object modelling scheme, is a graph that represents the joint probability distribution of a set of variables. In such a graph, links represent conditional probabilities, and nodes are the variables. The Bayes rule is applied to update the probabilities of nodes [12].

One alternative to the probabilistic approach is the possibilistic scheme derived to facilitate active fusion for object recognition. First introduced by [98], the possibility theory is based on the fuzzy logic. The work of Dempster and Shafer on statistical inference and uncertain reasoning yielded the Evidence Theory [99], which is another common alternative for the modelling of uncertainty in active vision systems. In [100] a comprehensive study has been conducted for active recognition using these strategies. The formulation of view planning is based on the expected increase in recognition quality gauged by entropy-like measures. This leads to successful detection of the regions in the eigenspace in which the manifolds are most separate, in

all three strategies. An uncertain object classification is employed, as opposed to a hard decision. All these methods perform better with sensor planning versus random sensor movement, in terms of recognition rate and speed.

2.4 Active Vision Schemes

In this section a summary of the prior work on active sensor placement is given. Most of these algorithms reorder the sensors to minimize some ambiguity function, while a few reports incorporate explicit planning algorithms [12, 94, 101].

In [35], the authors have used the Dempster-Shafer theory to combine evidence collected from the unknown object in a robot work-cell, and propose an action by a multi-sensor-planning strategy, while objects are represented by aspect graphs. Sensor placement operations are analyzed based on the current estimation of the work cell, this determines the maximum ambiguity that remains after taking each action. The control action that minimizes the ambiguity is the final choice.

Reference [101] uses the Aspect-Resolution Tree built on the basis of aspect diagrams to plan multiple views in an active recognition system. Results for a vision-based sensor as well as a haptic sensor are analyzed through experiments.

An algorithm for object recognition and localization in a model-based robot vision cell is demonstrated in [102]. Based on a set of rules, the optimal next view for the sensor is chosen by predicting the results of all possible actions. The state of the work cell is defined through a state vector, where each state corresponds to a specific set of rules. The recognition task is translated as the process of rule calling and space conversion.

Reference [103] demonstrates an optimized active sensing strategy for multisensor

fusion systems. This method employs estimated error of the estimates in order to determine the best sensor location to obtain useful data. Experiments are conducted in multi-sensor multi-target tracking, fusing the information received from tactile and visual sensors.

The work in [104], represents a sensor planning scheme for active object search using a mobile platform equipped with a depth finding method as the searcher. An optimization problem is formulated with the goal of maximizing the probability of the object detection with the minimum cost. The search space is modelled by the probability distribution of the presence of the target. Depending on the current state of the search space and the detecting capabilities of the recognition scheme, the sensor parameters are chosen for the next step. The concept of *sensed sphere* for the purpose of space modelling is proposed and utilized.

Reference [105] studies the active localization of robots from noisy sensor data in approximative world models. It proposes to employ rational criteria for setting the motion direction of the robot as well as controlling the sensor parameters, to ensure efficient localization.

In [106], a camera mounted on a robot hand is employed to take a sequence of images, in order to detect, track and estimate the location and pose of an object. A uniform statistical method based on the Expectation-Maximization (EM) algorithm is used to process the 2D views of the objects for object learning. The region of interest is extracted automatically through a motion tracking strategy. The camera parameters are set through a hand/eye-calibration method.

A knowledge-based 3D active CAD-based vision system was developed based on preprocessed and optimized Bayesian Networks built for a given set of CAD objects,

in [107]. This algorithm is sensitive to noise as 3D B-spline curves are used to represent the object rims. In this experiment, the Bayesian nets model the statistical behavior of the data.

The research in [12], proposes an online next view planning scheme for recognition of isolated 3D objects from a set of noisy feature detectors. Recognition and planning is based on a probabilistic reasoning framework. Based on the aspect graph construction approach presented in [36], an aspect graph is built from the noisy data. The hierarchical knowledge representation scheme quantifies both feature-based information on objects and the uncertainty in the recognition process. The algorithm is reactive, meaning it considers both the past and current observations to plan the next viewpoint, to best disambiguate the objects. The second experiment in [12], involves recognition of big objects that may not fit in the field of view of the camera, through a part-based knowledge representation scheme. An uncalibrated projective camera is considered, where the internal parameters of the camera may be varied. A new class of invariants for complete 3D Euclidean pose estimation, also known as inner camera invariants [108], are used, which are image-computable functions and invariant to the internal parameters of the camera. A probabilistic reasoning framework is employed for the active recognition and next view planning.

Reference [109] uses range data in a model-based shape, pose and position reconstruction. Neural networks is used to estimate Bayesian probabilities. The next view is chosen to minimize the expected ambiguity in terms of the Shannon Entropy.

In [32], the authors base their active object recognition scheme on integrating attention and viewpoint control. They present a Bayesian attention mechanism which maps objects into volumetric parts, volumetric parts into aspects, and aspects to

component faces. Aspects model a small set of volumetric part-classes, from which each object can be constructed. An augmented aspect hierarchy considers relations between boundary groups, the faces, the aspects, and the volumetric primitives themselves. In their model, each link represents a conditional probability and the entities at each level are linked to each other. The conditional probabilities coming from the augmented aspect hierarchy are used to quantify the average inferencing uncertainty.

The active recognition approach in [110] shows how the entropy maps can be computed using optical flow signatures. The entropy map provides guidance for an active observer along an optimal trajectory, by which the identity/pose of objects in the world can be inferred with confidence. Entropy maps are used to encode prior knowledge about the objects as a function of viewing point. The paper describes how these maps are computed using optical flow signatures, and how a view-planning strategy can be formulated using entropy minimization.

References [111, 112] present an active recognition and pose estimation system based on a neural network scheme to evaluate distances in global feature space. A feature space trajectory (FST) in the eigenspace generated from intensity images is employed to represent 3D views of an object. The next camera position that best resolves ambiguity is chosen by analyzing the FTSs.

Reference [41] has developed an active vision system that employs statistical representation of the object appearance through the receptive field vectors. The concept of *Transinformation* is used to choose the most discriminative viewpoint for the next step. This is achieved by evaluating the contribution of each receptive field vector. In this work, like many other ones, the knowledge about the unknown object is not *updated* after each sensor data acquisition. In fact, in all steps the *a priori*

density function of the hypotheses is assumed to be uniform, incorporating only the data achieved in the last step as opposed to the data received throughout the whole process.

In [94], the authors extend the off-line appearance-based recognition approach of [6] to an on-line 3D active object recognition scheme. They augment the parametric eigenspace with probability distributions which capture variations in the input data due to noise. The average entropy is minimized in order to select the next viewpoint.

In [113], *reinforcement learning* provides a means for viewpoint selection. A reward metric is defined, measuring how distinguishable objects are in a certain viewpoint. By maximizing this reward the best next viewpoint is determined. The mapping from the object state to the sensor actions are trained automatically through an eigenspace classification approach and approximating the reward function by Monte Carlo reinforcement learning.

In the work of [65], the process of recognition is assumed as one of sequential decision making, where the objects are learnt from their visual appearance through the eigenspace approach. The system uses a Radial Basis Function (RBF) network for the probabilistic interpretation of the input images. The *reinforcement learning* provides a tool to autonomously develop near-optimal decision making strategies for sensor placement, to disambiguate initial object hypotheses.

In [9], a framework based on the Shannon's information theory is employed to optimally select the sensor parameters for iterative state estimation in static systems, where the object class and pose are defined as the states. The camera parameters are selected to maximize the Mutual Information, which guarantees that the optimized information about the system state is obtained from the captured image.

2.5 Applications of Cramér-Rao Lower Bound (CRLB)

Cramér-rao bound has been applied to a large range of applications for example assessing the effects of different approximations in solutions offered for a certain problem. It has been also employed in system design [114] as the best achievable performance.

CRLB, defined as the inverse of the Fisher Information Matrix, is used widely in nonlinear filtering where it offers the best attainable second-order error performance, while no closed-form solution exists for the nonlinear filtering problem [115–117]. This bound has been used increasingly in the context of sensor management in radar scheduling, submarine tracking, and terrain navigation [11, 118, 119]. We will employ this performance index in the context of active multi-camera object recognition.

2.6 Summary

A review of the prior research on active vision was given is this chapter. In the following chapters the problem of active object recognition/pose estimation will be formulated and solutions will be offered in order to improve the robustness of the approach to sensor noise and occlusion.

Chapter 3

Problem Statement

In this chapter the problem of active object recognition is discussed and a solution to the multi-camera case is proposed. This chapter is organized as follows; an overview of the proposed approach is given in first section; next the problem is formulated as one of state estimation; in the third section the sequential Bayesian estimation is discussed; the contexts of state definition and transition are presented in the fourth and fifth sections respectively.

3.1 An Overview of the Approach

The application of this work is active multi-camera 3D object recognition with the possibility of partial occlusion, while the object belongs to a predefined set. The objects are described by their projections into a low dimensional space constructed based on the Principle Component Analysis, referred to as the eigenspace method in the literature.

A database is generated off-line, by accumulating the images of all objects in a

set of predefined poses. These raw images will then be processed to derive the low dimensional eigenspace. All the original images will be transformed into this space, in order to extract their signature description.

The process of the recognition is shown in Figure 3.1. This is essentially a state estimation process in which the states are defined as the class, pose and occlusion level of the object, while the occlusion levels may change throughout the experiment due to camera movement. The algorithm proceeds by recovering the high level representation of the input sensory data, and then hypothesizing the best match among the predefined models in the database. We adopt a probabilistic framework to encode the information received from the sensors in every step that leads to the Bayesian estimation approach. In each step, the probability of each reference category being the unknown object will be updated on the basis of the recently acquired data. The process is terminated if a certain category stands out in the sense of maximum likelihood.

In active object recognition, some parameters of the sensors are controllable and can be utilized to facilitate the recognition process. For example, the controllable parameters in this study are the positions of the cameras. After each image acquisition, the images from the two cameras are assessed to determine whether enough information is collected or not. If further data is required, the cameras control parameters are set by the system automatically based on the sensor model available in the database. The goal is to choose the most informative and discriminative viewpoint for the next step, in order to reduce the uncertainty and ambiguity over time. The training set and the current information about the object state are used to asses the usefulness of the camera actions.

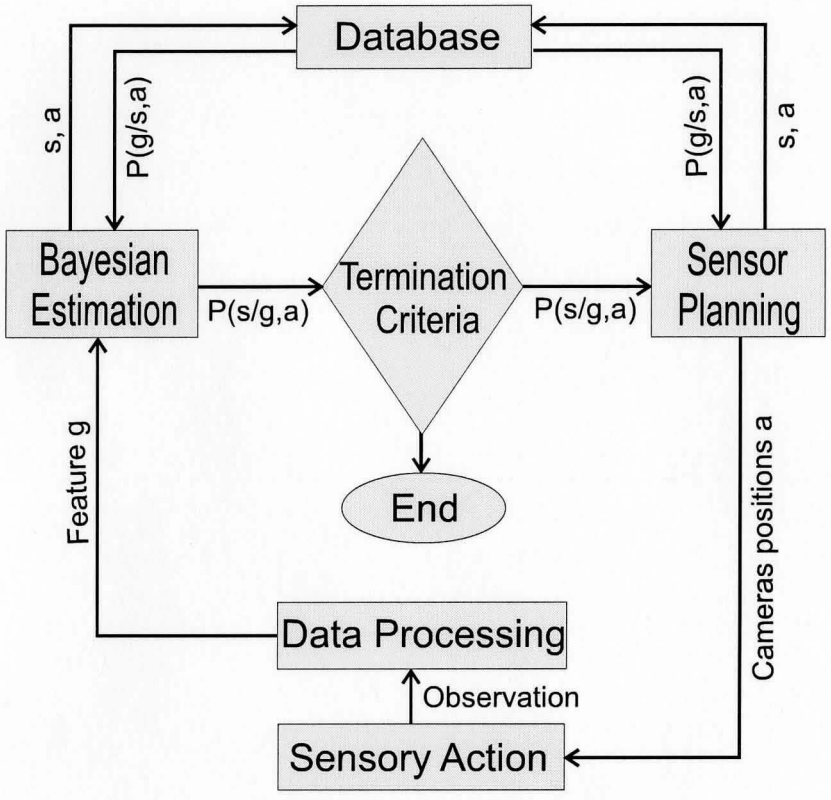

Figure 3.1: Flowchart of active object recognition. Parameter a represents the cameras positions, parameter g represents the feature extracted from the observation, and s represents the system state.

Two different approaches have been developed for the purpose of viewpoint planning. In the first technique the Mutual Information is maximized, whereas in the second strategy the lower bound of the mean square error of the state estimation, also known as Cramér-Rao Lower Bound, will be minimized.

3.2 The State Estimation Solution

As many other computer vision problems, the problem of object recognition can be formulated as that of state estimation. The goal of state estimation for a static system, like an object recognition system, is estimating the parameters of interest; whereas in dynamic systems, e.g. object tracking, it is reducing the error between the estimated and the true state over time. In this process, the information about the system state, either static or dynamic, is updated by employing the observations and the database.

In this thesis, the identity and pose of the object are assigned as unknown states/parameters of interest that must be estimated based on the camera images. In addition to the object identity and pose, we will propose to add some measure of occlusion to the object states. This will reduce the sensitivity of the recognition process to structured noise. The inclusion of occlusion levels in the state definition and their time dependence render the problem of object recognition/pose estimation into that of dynamic state estimation. All system states, i.e. the object class, its pose, and percentages of occlusion in each camera image are discretized and therefore take values from a finite set.

The noisy sensor data results in uncertainty that can be best modelled by a framework based on Shannon's Information Theory. This framework enables us to statistically model the sensor input in the database. It also allows us to describe the system state through probability distribution functions. In the beginning of the online recognition step, a uniform density function is assumed for the system state, which indicates a complete uncertainty about the environment. After each observation is made, the probability distribution of the system state is updated through a

Bayesian scheme, as shown in Figure 3.2.

3.3 Sequential Bayesian Estimation

In this work, we adopt a probabilistic framework for the flow of information throughout the recognition process. This enables us to incorporate the camera noise, occlusions, as well as the errors in the camera and object positioning mechanisms into the system model through probability distribution functions. We also consider closed world assumption, meaning the sensory data is always generated by one of the known categories in the database. This assumption allows one to assign a probability distribution over object classes to each point in the eigenspace.

Decisions with respect to identity/pose are made according to the probabilities of each object/pose in the database being the true case. This strategy is in contrast to some techniques that only provide the solution based on hard decisions, without giving any information about other possibilities. A brief overview of the sequential sensor control and decision making process for active object recognition follows next [9, 32, 100].

In the beginning of the classification process, no *a priori* knowledge about the unknown object is available which yields to the assumption of a uniform probability density function over the states. The goal is to estimate the unknown state s^n based on the observations g_0, g_1,..., g_n and sensor parameters a_0, a_1,..., a_n where n denotes the time index. At any given step n, *a priori* information about the prior state s^{n-1} is available in the form of conditional probability density function $p(s^{n-1}|g_{n-1}, a_{n-1}, \cdots, g_0, a_0)$. The sensor measurements and the state transition dynamics are used to obtain the *a posteriori* density function for the state at time n,

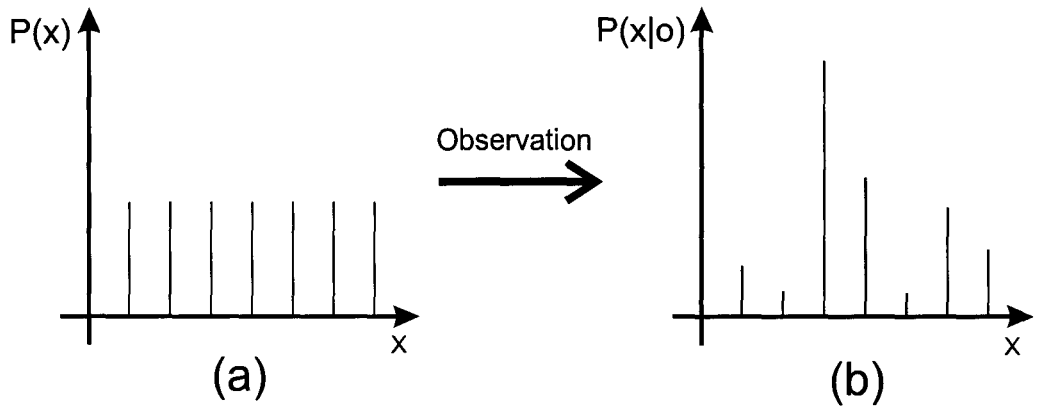

Figure 3.2: The appropriate observation reduces the uncertainty and ambiguity in the pdf of the state x; (a) the pdf of x before the observation o; (b) the pdf of x after the observation.

using the *Bayes* rule as

$$p_{post}(s^n) = \frac{p(g_n|s^n, a_n)p_{pri}(s^n)}{p(g_n|a_n, g_{n-1}, a_{n-1}, \cdots, g_0, a_0)} \tag{3.1}$$

where

$$p_{post}(s^n) = p(s^n|g_n, a_n, g_{n-1}, a_{n-1}, \cdots, g_0, a_0) \tag{3.2}$$

and

$$p_{pri}(s^n) = p(s^n|a_n, g_{n-1}, a_{n-1}, ..., g_0, a_0) \tag{3.3}$$

An important characteristic of our system is that the probabilities $p_{post}(s^{n-1})$ and $p_{pri}(s^n)$ are not necessarily equal, which can be explained through the time variance of the system state resulted from the certain definition of states in this thesis.

Assuming the state transition is governed by a *Markov process* [120], the probability of each state only depends on the state probabilities in the previous step, i.e.,

$$p_{pri}(s^n) = f(p_{post}(s^{n-1})) \tag{3.4}$$

More details on the Markov chain are given in the following sections. The sequential application of these updates would yield the probability density function (pdf) of the states at any given time. It should be noted that the observation distribution conditioned on the state and sensor position, $p(g_n|s_n, a_n)$, is required for the application of the recursive *Bayes* rule in (3.1). In other words, a probabilistic sensor model that links the observations to the states is needed. This will be generated through the off-line training step.

3.4 State Definition

It is known that the eigenspace method is sensitive to structured noise due to its global modelling approach [15, 62]. To tackle this problem, We propose to include the level of occlusion in each camera image, in the object states. We define the occlusion level as the percentage of the image corrupted due to the presence of an unwanted object in the workspace. In general, it is difficult to model the occlusion as its shape, size, and position in the image are unknown and can vary from one camera viewpoint to another. The levels of occlusions also depend on the cameras positions in the workspace w.r.t. the object and hence can vary throughout the experiment. It would be difficult to model the correlation between the occlusions in different images considering their complexity and also the large number of possible ways to obstruct the object of interest. To simplify the problem, we assume that the occlusion levels are independent from one camera pose to another and also among different cameras

images.

By considering the occlusions, the system state includes the object class, its pose, and the percentages of occlusion in each camera image. For our two-camera system, one may write

$$s^n_{i,j,k,l} = [o^n_i, \varphi^n_j, \xi^{1n}_k, \xi^{2n}_l] \tag{3.5}$$

where o^n_i is the object class i, φ^n_j is pose number j, ξ^{1n}_k is the occlusion level k in the image received from Camera 1, and ξ^{2n}_l is the occlusion level l observed in the image from Camera 2, all at time step n. Among these four states, the object class and its pose are constant throughout the entire object recognition experiment. The occlusion levels may vary depending on the camera locations. Therefore, (3.5) can be rewritten as

$$s^n_{i,j,k,l} = [o_i, \varphi_j, \xi^{1n}_k, \xi^{2n}_l] \tag{3.6}$$

3.5 State Transition

The movement of the cameras during the active recognition process would change the occlusion levels in each camera view. Due to the assumption of occlusion independence, transitions could occur between all levels of occlusion in each image. In this thesis we assume that the state transitions are governed by a *Markov* chain, since all prior information from the observations are embedded in the *a posteriori* probabilities $P_{post}(s^{n-1}_{w,x,y,z})$ corresponding to the states at $n-1$. Therefore,

$$P_{pri}(s^n_{i,j,k,l}) = \sum_{w,x,y,z} P(s^n_{i,j,k,l} | s^{n-1}_{w,x,y,z}) P_{post}(s^{n-1}_{w,x,y,z}) \tag{3.7}$$

Since the first two components of the states, i.e., object identity and pose, are time-invariant, no transition occurs between states with different class and/or pose. Therefore, the corresponding transition probabilities in the Markov chain are set to zero (see Figure 3.3). According to this model, transitions only occur between states of the same object and pose class

$$P_{pri}(s^n_{i,j,k,l}) = \sum_{y,z} P(s^n_{i,j,k,l}|s^{n-1}_{i,j,y,z}) P_{post}(s^{n-1}_{i,j,y,z}) \tag{3.8}$$

$P(s^n_{i,j,k,l}|s^{n-1}_{i,j,y,z})$ in (3.8) is constant as the transition between levels of occlusion is assumed independent of time, sensor locations, object class and pose, and uniformly distributed, i.e.,

$$P(s^n_{i,j,k,l}|s^{n-1}_{i,j,y,z}) = P_{k,l|y,z} = P_\alpha \tag{3.9}$$

3.6 Summary

In this chapter, the problem of 3D active object recognition was formulated as a state estimation one. A Bayesian estimation scheme was developed and the system state was defined. The transitions between the states are modelled through a Markov chain which is modified to follow the real world observations correctly. The following chapter demonstrates an approach for the encoding of the sensor data.

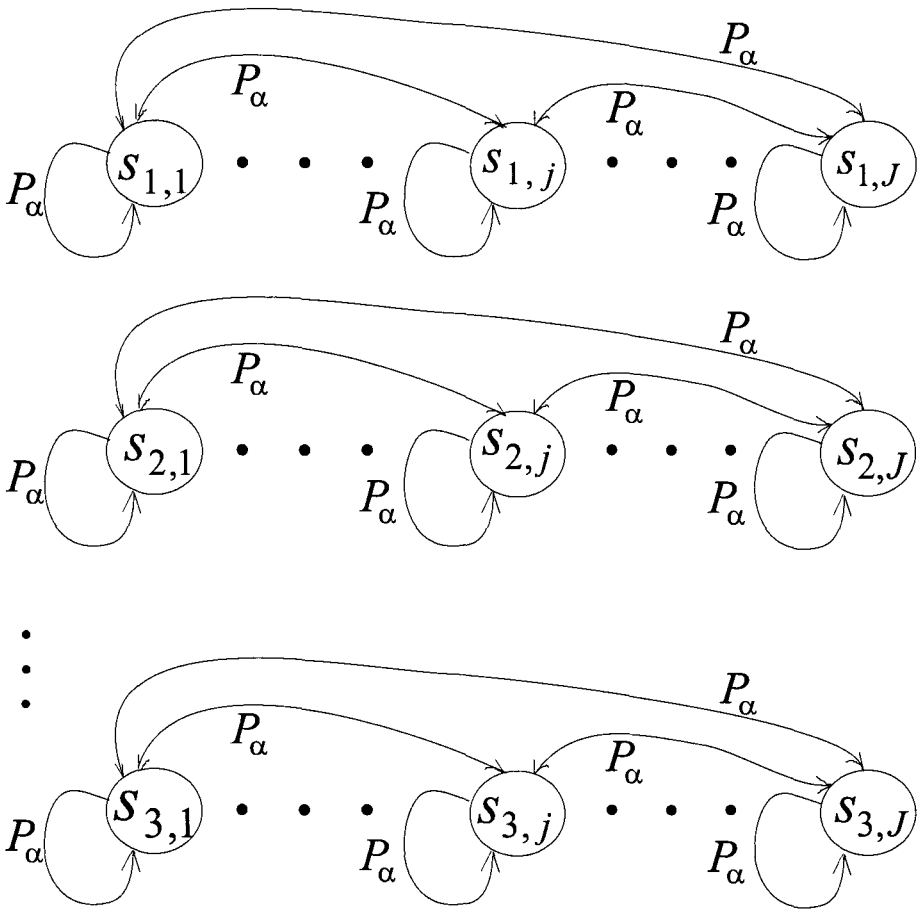

Figure 3.3: The transition between the states is governed by a Markov chain. The circles present states and the links demonstrate transitions in the direction of the arrows. I object classes and J occlusion levels (including zero percent occlusion) are assumed, and all the transition probabilities are $P_\alpha = \frac{1}{J}$.

Chapter 4

Sensor Modelling

The observations collected from the unknown object are initially in the form of raw camera images. Information processing and state estimation using these images could be a formidable task due to the large volume of data contained in each image. Alternatively, preprocessing techniques may be implemented to reduce the size of data and extract relevant features of the object.

Many recognition algorithms have successfully employed Eigen-Representation approaches to model the information acquired from task environment [6,65]. Motivated by the high correlation existing between the appearances of the object from different viewpoints, images taken from the objects are transformed to a subspace where individual features are uncorrelated. Such a subspace can be constructed by employing the Principle Component analysis (PCA) for a set of images [58].

4.1 The Eigenspace Approach

The key idea in the appearance-based parametric eigenspace is to employ a small set of images to span the image space. Images from objects can be reconstructed by a linear combination of these eigen-images and therefore, each image can be represented by a small number of coefficients. The eigen-images are extracted from a complete set of images captured under different pose and illumination conditions. For M object classes in N different poses, $M \times N$ total images are collected.

These training images are converted to vectors Y_i. Prior to the learning phase, the acquired images are normalized so the total energy in each image is unity. This ensures the independence of the observation from the light intensity [87].

$$\hat{X}_i = Y_i/\sigma_i \quad , 1 \leq i \leq M \times N \tag{4.1}$$

where,

$$\sigma_i = (\sum_{k=1}^{m \times n} (Y_i(k))^2)^{1/2} \tag{4.2}$$

and the images contain $m \times n$ pixels. In the next step, these vectors will be mean-adjusted to ensure that the eigenvector associated with the maximum eigenvalue, represents the dimension of the subspace in which the maximum image variance in the correlation sense exists [87],

$$X_i = \hat{X}_i - \bar{X} \quad , 1 \leq i \leq M \times N \tag{4.3}$$

and,

$$\bar{X} = \sum_{i=1}^{M \times N} \hat{X}_i / (M \times N) \tag{4.4}$$

The matrix $S_{mn \times MN}$ is then obtained by arranging the final images as follows,

$$S = [X_1, ..., X_N, ..., X_{M \times N}] \tag{4.5}$$

The Principle Component Analysis (PCA) can be applied to obtain the eigenvalues and corresponding eignvectors, by applying the Singular Value Decomposition (SVD) to the covariance matrix associated with S. A few eigenvectors associated with the largest eigenvalues are selected to construct the eignenspace. This yields a partial basis compared to the complete Karhunen-Loeve basis which is composed of all eigenvectors. It can be shown that such a basis is the optimal set among all bases of the same dimension in the L^2 sense. In other words, approximating the data as a linear combination of any other set of basis vectors with the same dimension will produce larger average l^2 error [48].

To calculate the eigenvectors, the covariance matrix is defined as

$$Q = SS^T \tag{4.6}$$

The matrix Q is of size $mn \times mn$ and the calculation of its eigenvectors is computationally expensive. Several remedies to this problem have been proposed in the literature. In [6], the spatial temporal adaptive (STA) algorithm proposed in [48] has been used to obtain the k most significant eigenvectors where $k \ll mn$. Nevertheless as pointed out in [58], it is difficult to achieve a robust and fast implementation of such iterative methods. Alternatively as shown in [45, 48], the eigenvectors of Q can

be computed from the eigenvectors of the much smaller matrix $T = S^T S$. Based on the definition of eigenvectors we have

$$S^T S v_i' = \lambda_i' v_i' \tag{4.7}$$

where v_i' and λ_i' are the eigenvectors and eigenvalues of T, and, v_i and λ_i the corresponding for matrix Q. By premultiplying both sides of (4.7) by S one may write

$$S S^T S v_i' = \lambda_i' S v_i' \tag{4.8}$$

This shows that Sv_i' is an eigenvector for $Q = SS^T$. The following relations have been developed using the Singular Value Decomposition framework [121],

$$\lambda_i = \lambda_i' \tag{4.9}$$

$$v_i = \lambda_i'^{-\frac{1}{2}} S v_i' \tag{4.10}$$

Once the eigenvectors are computed, the images can be projected into points g_i in the eigenspace as follows,

$$g_i = V X_i \quad 1 \leq i \leq MN \tag{4.11}$$

where $V = [v_1, v_2, ..., v_k]$ is the matrix formed by the eigenvectors so that,

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k \tag{4.12}$$

The inner product in (4.11), corresponds to the least square estimate of the eigenspace coefficient vector, g, that provides the least squares error between the original image, X, and the one approximated by the sum of the eigenvectors, defined as:

$$error = \sum_{i=1}^{m \times n} (X(i) - \sum_{j=1}^{k} g(j) v_j(i))^2 \tag{4.13}$$

To reconstruct an image from the vector g, the pseudo-inverse of V, V', can be used (see Figure 4.1),

$$V' = V^T (V V^T)^{-1} \tag{4.14}$$

$$X_i' = V' g_i + \bar{X} \tag{4.15}$$

There are approaches that consider different eigenspaces for each object, built based on the images from that certain object. This increases the precision specially for the purpose of pose estimation. Alternatively, in this thesis, we define different eigenspaces for different viewpoints of each camera, to achieve high accuracy in the recognition process.

4.2 Object Representation

Ideally, each image from a certain object at a given pose should correspond to a unique point in the eigenspace. However, factors such as camera noise, errors in camera settings, changes in the illumination condition, and errors in the camera or

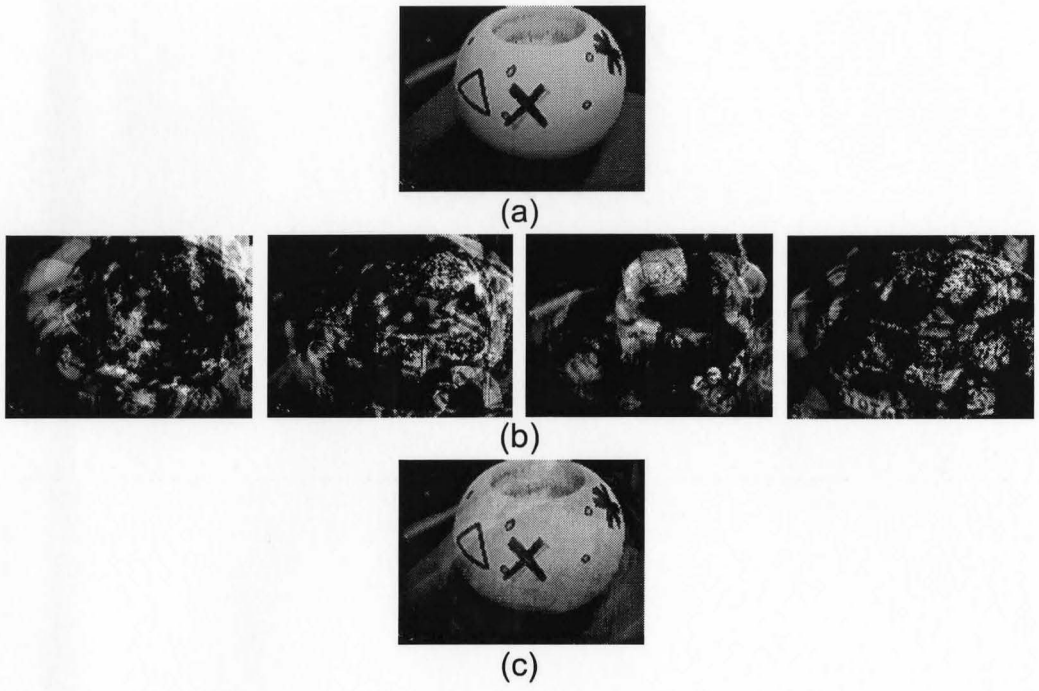

Figure 4.1: An image is projected into the eigenspace and then reconstructed; (a) the original image; (b) the most significant eigenvectors in the image eigen-representation; (c) the reconstructed image.

object position can create uncertainty in the eigenspace coefficients. The comparison between the observation and database would be unreliable if the uncertainties associated with the parameters are not considered. The reason is the lack of uniqueness in the transformation of noisy measurements, which will make the judgement difficult and prone to mistakes. These moderate unstructured errors can be modelled using a probabilistic approach, i.e., by replacing points with likelihood functions that represent each object/pose in the eigenspace. We further assume that the observation vector (eigenspace parameters) distribution for the observation of each camera, conditioned on the object state, is *Gaussian.*

Observation g is considered as a vector that includes the data obtained from the two cameras, i.e.,

$$g = \begin{bmatrix} g_1 & g_2 \end{bmatrix} \tag{4.16}$$

where g_1 and g_2 are observations from Camera 1 and 2, respectively. We assume that the camera observations as well as the occlusion levels in the images are conditionally independent of each other. Therefore,

$$p(g|s_{i,j,k,l}, a_v) = p(g_1|s_{i,j,k,l}, a_v)p(g_2|s_{i,j,k,l}, a_v) \tag{4.17}$$

with

$$p(g_1|s_{i,j,k,l}, a_v) = N(g_1, \mu_{i,j,k,v}, \sigma_{i,j,k,v}) \tag{4.18}$$

$$p(g_2|s_{i,j,k,l}, a_v) = N(g_2, \mu_{i,j,l,v}, \sigma_{i,j,l,v}) \tag{4.19}$$

Note that the first *Gaussian* function depends only on the occlusion level k and the second one just on l, due to the assumption of independence between the observations of the two cameras. The state has already been defined in (3.6),

$$s_{i,j,k,l} = \begin{bmatrix} o_i & \phi_j & \xi_k^1 & \xi_l^2 \end{bmatrix} \tag{4.20}$$

and a is the camera control vector given by

$$a = \begin{bmatrix} \theta_1 & \gamma_1 & \theta_2 \end{bmatrix} \tag{4.21}$$

where θ_1 and γ_1 are the pan and tilt angles of Camera 1, and θ_2 is the pan angle of Camera 2. Other parameters are computed as shown in Appendix C.

The Gaussian distribution is defined as follows,

$$N(q, \mu_{s,a}, \sigma_{s,a}) = \frac{1}{\sqrt{(2\pi)^n |\sigma_{s,a}|}} \exp\left\{ -\frac{1}{2}[q - \mu_{s,a}]^T \sigma_{s,a}^{-1} [q - \mu_{s,a}] \right\} \tag{4.22}$$

It should be noted that for K object classes and I camera positions, one would require $K \times I$ multivariate Gaussian distributions as defined in (4.22), to represent the observation distribution.

Maximum Likelihood estimation of the mean and variance of the Gaussian distributions are done by the projection of a large number of training samples into the eigenspace,

$$\mu_g^{M.L.} = E[g] \tag{4.23}$$

$$(\sigma_g^{M.L.})^2 = E[(g - \mu_g^{M.L.})^2] \tag{4.24}$$

In the experimental setup of this paper, two cameras are used with eight different objects, each with four possible equispaced poses between 0 and 2π. There are also 32 and 7 different positions for Camera 1 and 2, respectively. Due to the large

number of possible combinations of camera positions, object class, and object pose, it is infeasible to employ a single eigenspace. Instead, we define separate eigenspaces for individual camera locations, which greatly facilitates the application of the PCA. It should be noted that the cameras locations are always known, and having different eigenspace imposes no limitation on our work, as $p(g|s_{i,j,k,l}, a_v)$ is sufficient for the calculation of the MI and CRLB, as discussed in Chapter 5.

4.3 Robust Sensor Model

In our experiments, we consider the possibility of the presence of other objects in the workspace which can result in occlusion. Consequently, we intend to modify the sensor model and make it robust to occlusions. It is Ideal to have a general database that can handle a variety of occlusions in different shapes and levels. On the other hand, as occlusions can happen in random, unpredictable ways, they are very hard to model.

In order to avoid an exhaustive training stage, we seek a simple way to generate such a database. To this end, the *Gaussian* distributions in (4.17)-(4.19) are generated by corrupting the original images with random occlusion and noise. The distributions are conditional on the camera location, object class, object pose, and occlusion level. They are defined in the eigenspaces associated with individual camera viewpoints. Each eigenspace is generated by the application of PCA on the *original non-occluded* images of all object classes and poses, taken by one of the cameras in a fixed viewpoint. We have chosen five percentage levels of occlusion, i.e. 5%, 10%, 15%, 20%, and 25% to achieve models for different levels of occlusion. The occlusion shape is randomly selected from a set of five shapes, square, rectangle, circle, triangle and trapezoid. In

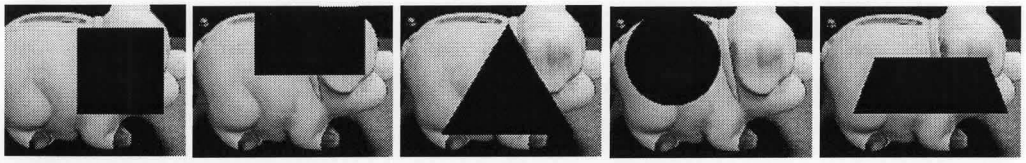

Figure 4.2: An image is randomly occluded; the area of the corrupted part is about 25% of the total area in the image.

order to generate a model that is close to reality, these shapes are placed at random locations on the object image by replacing the corresponding pixels with black pixels (see Figure 4.2).

Finally, the complete sensor model consists of (k_1+k_2) eigenspaces, and $(M.N.L.(k_1+k_2))$ *Gaussian* functions, for the case of M object classes, each in N different poses, k_1 locations for Camera 1, k_2 locations for Camera 2, and L occlusion levels.

4.4 Summary

The interpretation of the real world observations through an eigenspace approach was discussed in this chapter. The occlusion was incorporated into the database and the state estimation scheme. In the following chapter, two different measures will be introduced for selecting the best next camera locations in an active multi-camera vision system.

Chapter 5

Selection Criteria for Camera Movements

The key feature of active sensing is to use the existing information and plan the next data acquisition such that the most relevant information pertaining to the object state is obtained. The outcome of such a process is dependent on the metric used for the evaluation of the quality of observations. In this thesis, we consider two different indices for selecting the best next camera positions, namely the MI and CRLB. The camera movements are chosen to obtain the most discriminative sensor observations by optimizing performance measures based on these metrics. This will consequently enhance the quality of state estimation by improving the measurements. These performance measures are discussed below.

5.1 Mutual Information (MI)

In order to quantify the quality of information, for a random variable x the notion of entropy is defined as [120]

$$H(x) = -\int_x p(x)\log(p(x))dx \tag{5.1}$$

This represents the amount of uncertainty associated with the pdf $p(x)$ of random variable x. In an unambiguous situation the entropy is zero and for a completely ambiguous case of uniform distribution the entropy is maximized. The Conditional Mutual Information of random variables x, y and z, is defined as a measure of the reduction in the uncertainty [120],

$$I(x;y|z) = H(x|z) - H(x|y,z) \tag{5.2}$$

In our case, the state s^n of the object is the variable the uncertainty of which we intend to reduce. The observation vector g and the state s are related through the likelihood function $p(g|s,a_n)$, where a_n represents the vector of cameras' positions; a_n can be employed as a control variable to reduce the entropy and therefore the uncertainty in s^n due to the observation g. In this thesis, the Mutual Information (MI) is defined as,

$$I(s^n;g|a_n) = H(s^n|a_n) - H(s^n|g,a_n) \tag{5.3}$$

This metric is zero if the states and the observation are uncorrelated and reaches its maximum when the observation can definitely resolve the states ambiguity.

A set of camera positions that can best resolve such ambiguity can be obtained

by solving the following optimization problem:

$$a_n = \arg\max_a I(s^n; g|a) \tag{5.4}$$

This optimization problem is based on the assumption that the distribution of the feature expected to be observed after the sensory action, will depend only on the last control action and not on the previous ones. The philosophy behind this approach is another assumption that the reduction in the overall uncertainty of the system could lead to a better state estimation, without having to improve the state estimator.

In our case, the state space is discrete and the MI is given by,

$$I(s^n; g|a_n) = \sum_{i,j,k,l} P(s^n_{i,j,k,l}|a_n) \int_g p(g|s^n_{i,j,k,l}, a_n) \log\left(\frac{p(g|s^n_{i,j,k,l}, a_n)}{p(g|a_n)}\right) dg \tag{5.5}$$

This indicates that, the next camera positions are chosen based on the available observations, and the conditional probability density function $p(g|s^n_{i,j,k}, a_n)$ and $p(g|a_n)$. An approach for generating the observation likelihood $p(g|s^n_{i,j,k,l}, a_n)$ based on the training data set was discussed in Chapter 4. Also, $p(g|a_n)$ in (5.5) can be calculated based on the fact that,

$$\sum_{i,j,k,l} P(s^n_{i,j,k,l}|g, a_n) = 1 \tag{5.6}$$

which results in,

$$p(g|a_n) = \sum_{i,j,k,l} P(s^n_{i,j,k,l}|a_n) p(g|s^n_{i,j,k,l}, a_n) \tag{5.7}$$

based upon the following equation,

$$P(s^n_{i,j,k,l}|g, a_n) = P(s^n_{i,j,k,l}|a_n)p(g|s^n_{i,j,k,l}, a_n)/p(g|a_n) \tag{5.8}$$

The main hurdle in solving the optimization problem in (5.4) is its computational cost that is caused by the integration over g in (5.5). One can quantize the feature space, when possible. For example in [9], the mean gray value of the image pixels is used as the feature, which can be easily quantized as follows,

$$I(s^n; g|a_n) = \sum_{s^n} P(s^n) \sum_{g_i} p(g_i|s^n, a_n) \log \frac{p(g_i|, a_n)}{p(g_i|a_n)} \tag{5.9}$$

Such an approach would be infeasible for our application however, because of the large dimension of the problem which could lead to exhaustive computation. Instead, we compute the MI in (5.5) via Monte Carlo simulation, i.e.

$$I(s^n; g|a_n) = E_{P(s^n)} \left[E_{p(g|s^n, a_n)} \left[\log \left(\frac{p(g|s^n, a_n)}{p(g|a_n)} \right) \right] \right] \tag{5.10}$$

According to the the Law of Large Numbers, the expected value of the random variable $f(x)$ may be computed by sampling it using the corresponding distribution $p(x)$ [122], i.e.,

$$E_{p(x)}[f(x)] = \underset{n \longrightarrow \infty}{\text{Lim}} \frac{1}{n} \sum_{i=1}^{n} f(x_i) \tag{5.11}$$

Since $p(g|s^n, a_n)$ is independent of the sensor observations, the calculations in the recognition step can be reduced by the off-line computation of the inner expected value in (5.10), after the training phase is completed. For that purpose, we wish to

reformulate (5.10) as follows,

$$I(s^n; g|a_n) = E_{P(s^n)}(E_{s^n}) \tag{5.12}$$

$$E_{s_i^n} = E_{p(g|s_i^n, a_n)}\left[\log\left(\frac{p(g|s_i^n, a_n)}{p(g|a_n)}\right)\right] \tag{5.13}$$

The convergence of this sequential decision making process is given in the Appendix A. It is also argued that this approach is optimal in the uncertainty reduction sense [9]. For a fixed *a priori* pdf, i.e. constant $H(s^n|a_n)$ in (5.3), the MI depends only on the conditional entropy. This indicates that maximizing the Mutual Information corresponds to minimizing the conditional entropy $H(s^n|g, a_n)$. It can be concluded from (5.3) that the change in the uncertainty depends on the current observation, but it can not be claimed that the uncertainty will be reduced after every step of the decision making process. However, based on the definition of Mutual Information, the uncertainty will decrease after a number of steps.

5.2 Cramér-Rao Lower Bound (CRLB)

Let g_n be the observation vector, s^n be the parameter vector to be estimated, and $\hat{s}^n$ be an unbiased estimate of s^n, i.e.,

$$E[\hat{s}^n(g_n)] = s^n \tag{5.14}$$

where the expectation is with respect to s^n. The covariance of the $\hat{s}^n$ has a theoretical lower bound, related to the likelihood function as follows [122],

$$C_n = E\left[(\hat{s}^n - s^n)(\hat{s}^n - s^n)^T\right] \geq J_n^{-1} \tag{5.15}$$

The inequality denotes that the matrix $C_n - J_n^{-1}$ is a positive semi-definite matrix; J_n is the Fisher Information Matrix given by,

$$J_n = -E\left[\nabla_{s^n} \nabla'_{s^n} \log p(s^n, g_n)\right] \tag{5.16}$$

or equivalently,

$$J_n = E\left[\{\nabla_{s^n} \log p(s^n, g_n)\} \{\nabla_{s^n} \log p(s^n, g_n)\}^T\right] \tag{5.17}$$

where,

$$\nabla_x = \left[\frac{\partial}{\partial x_1}, ..., \frac{\partial}{\partial x_n}\right] \tag{5.18}$$

The Fisher Information is a measure of knowledge which expresses the capability of the estimator in estimating a parameter, and measures the state disorder of a system or phenomenon. Equation (5.15) is known as the Cramér-Rao inequality, and is a measure of intrinsic uncertainties when no external source of noise is present (see Appendix B for the proof). This equation indicates an increase in the estimation quality with increase in the Fisher Information, making it a valuable candidate for a quality metric in assessing the estimation procedure [123]. The Cramér-Rao Lower Bound (CRLB), defined as the inverse of the Fisher information matrix, provides the variance for an efficient estimator.

5.2.1 A Recursive Formulation for Posterior CRLB

In case of state estimation, CRLB is also known as the Posterior Cramér-Rao Lower Bound (PCRLB), and is very difficult to calculate directly from the definition. Alternatively, the Fisher information matrix can be calculated using Riccati-like recursions [115, 117],

$$J_n = D_{n-1}^{22} - D_{n-1}^{21}[J_{n-1} + D_{n-1}^{11}]^{-1} D_{n-1}^{12} + J_g(n) \tag{5.19}$$

where,

$$\begin{aligned}
D_{n-1}^{11} &= E\left[\left\{\nabla_{s^{n-1}} \log p(s^n|s^{n-1})\right\}\left\{\nabla_{s^{n-1}} \log p(s^n|s^{n-1})\right\}^T\right] \\
D_{n-1}^{22} &= E\left[\left\{\nabla_{s^n} \log p(s^n|s^{n-1})\right\}\left\{\nabla_{s^n} \log p(s^n|s^{n-1})\right\}^T\right] \\
D_{n-1}^{12} &= E\left[\left\{\nabla_{s^{n-1}} \log p(s^n|s^{n-1})\right\}\left\{\nabla_{s^n} \log p(s^n|s^{n-1})\right\}^T\right] \\
D_{n-1}^{21} &= (D_{n-1}^{12})^T \\
J_g(n) &= E\left[\left\{\nabla_{s^n} \log p(g_n|s^n)\right\}\left\{\nabla_{s^n} \log p(g_n|s^n)\right\}^T\right]
\end{aligned} \tag{5.20}$$

and the initial Fisher Information Matrix J_0 is defined as

$$J_0 = E\left[\left\{\nabla_{s_0} \log p(s_0)\right\}\left\{\nabla_{s_0} \log p(s_0)\right\}^T\right] \tag{5.21}$$

Note that J_n is dependent on the camera control parameters a, as a consequence of the dependency of g_n and s^n on a. Since the state vector is discrete in our system, the PCRLB cannot be directly computed through this formulation, as it is based on continuous variables and there is a need to calculate the gradients with respect to the state in (5.20). However, we employ a new continuous state vector based on the

discrete version to enable the evaluation of the PCRLB,

$$p(s_c^n) = \sum_{s_d^n} P(s_d^n) N(s_c^n, s_d^n, \sigma_s) \tag{5.22}$$

and

$$N(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(x-\mu)^2/2\sigma^2) \tag{5.23}$$

where s_c^n and s_d^n represent the continuous and discrete state at time n, respectively. The variance of the normal distribution σ_s is chosen small and therefore, the mixed continuous distribution is concentrated around the initial discrete state values, generating a distribution as close to the discrete case as possible (see Figure 5.1). The joint probability distribution density of observation g_n and state s_c^n is given by

$$p(g_n, s_c^n | a_n) = \sum_{s_d^n} P(s_d^n) N(s_c^n, s_d^n, \sigma_s) p(g_n | s_d^n, a_n) \tag{5.24}$$

The next camera positions can be selected by solving the following optimization problem:

$$a_n = \underset{a}{\operatorname{argmin}}\ J_n^{-1}(a) \tag{5.25}$$

which guarantees the least lower bound on the estimation error.

The optimization problems in (5.4) and (5.25) are both solved through exhaustive search on the control action domain. One of the reasons for using this strategy is that the domain of control actions is limited, so the exhaustive search is feasible and would not slow down the algorithm. If one considers continuous sensor locations, as opposed to discrete predefined sensor positions, one may define a single eigenspace

for all sensor positions. An optimization problem can be defined in that space then, which is not possible in our multi-space case, as there is a high possibility of having local minimums as a result of complications associated with employing numerous eigenspaces.

Considering our assumptions about the state transitions, the first two terms in (5.19) are independent of the control action a_n, and hence the optimization problem in (5.25) can be simplified by replacing J_n with $J_g(n)$. Also,

$$J_g(n) = E\left[\left\{\nabla_{s^n} \log(\frac{p(g_n, s_c^n|a_n)}{p(s_c^n|a_n)})\right\}^2\right] \tag{5.26}$$

which will result in

$$J_g(n) = E\left[\left\{\frac{\sum P(s_d^n)D(s_c^n, s_d^n, \sigma_s)p(g_n|s_d^n, a_n)}{\sum P(s_d^n)N(s_c^n, s_d^n, \sigma_s)p(g_n|s_d^n, a_n)} - \frac{\sum P(s_d^n)D(s_c^n, s_d^n, \sigma_s)}{\sum P(s_d^n)N(s_c^n, s_d^n, \sigma_s)}\right\}^2\right] \tag{5.27}$$

where all $\sum$'s are over s_d^n, the expectation is with respect to g_n and s_c^n, and

$$D(x, \mu, \sigma) = -\frac{(x-\mu)}{\sigma^2}N(x, \mu, \sigma) \tag{5.28}$$

We employ the Vegas Adaptive Monte Carlo Algorithm [124,125] for calculating the expectation in (5.27). The implementation of this method is not as straightforward as that of the MI as the integrals involved must be computed during the recognition phase. Therefore, recognition based on the PCRLB will require more real-time computational resources compared with the one based on the MI.

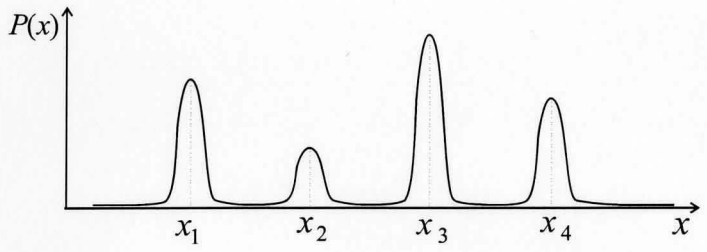

Figure 5.1: The discrete state is approximated by a continuous one; the probability distribution function of the new state is constructed by employing normal functions centered at the discrete states.

5.3 Summary

In this chapter, the context of efficient active camera positioning was discussed based on two different metrics, MI and CRLB. Computational strategies were developed to make the use of these measures possible. Experimental results for these two measures will be given in Chapter 7. In the next chapter, the experimental setup, used in the experiments implemented in Chapter 7 of this thesis, will be described.

Chapter 6

Vision System Description

We have developed an active multi-camera vision system, which is shown in Figures 1.1 and 6.1. This structure allows a relatively large motion range for two pan/tilt camera units. The first pan/tilt unit is mounted on an x-y gantry frame, while the second unit moves along the y-axis on a linear track on the side of the workspace. The two-camera system provides seven degrees of motion, three linear and four rotational. The object is placed on a turn-table mechanism and is always kept on the lines-of-sight of cameras by adjusting the positions of the linear track and the gantry system. This reduces the number of degrees of freedom to three, i.e., the pan angle for the side camera and the pan and tilt angles for the top camera. It is assumed that the object is randomly chosen from a set of *a priori* known objects, i.e. a *closed world assumption.* The goal is to determine the object identity and pose, which is held at a constant position by the turn-table, through individual trials. The objects maybe partially occluded during the recognition process. This objective should be achieved by acquiring and processing images from both cameras in multiple steps.

In the following sections, a brief overview of the equipment used in the experiments

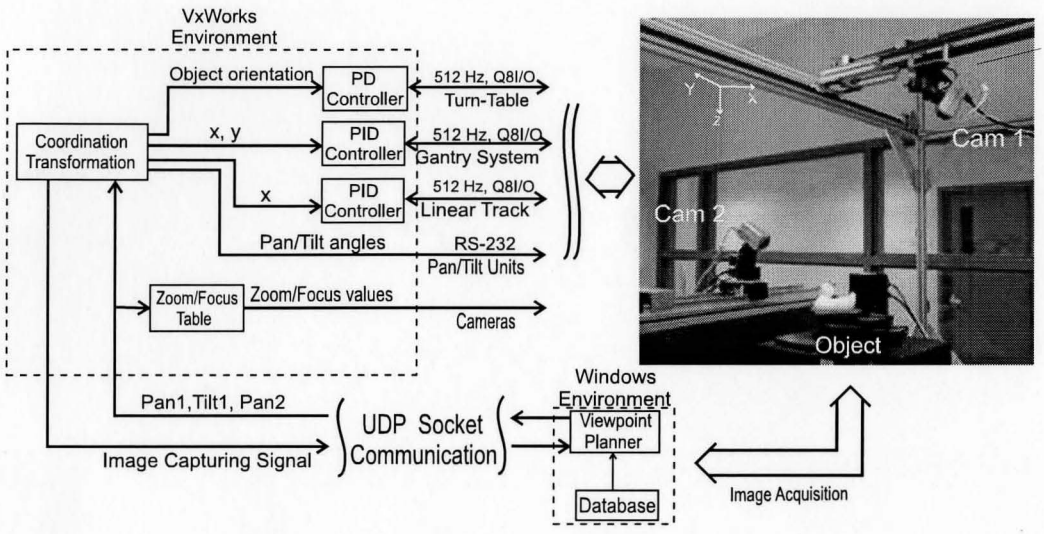

Figure 6.1: The experimental setup.

is given.

6.1 Gantry System

Camera 1 mounted on the pan-tilt unit, is held and positioned by a frame structure, which provides two degrees of motion along the x and y axes as can be seen in Figure 6.1. The Gantry system is equipped with two DC motors, and optical encoders measure the x and y positions with an accuracy of 3.04mm positioning resolution.

Proportional-Integral-Derivative (PID) controllers are implemented to control the x and y motion, under Tornado/VxWorks RTOS with a control rate of 512Hz. Q8 multifunction data acquisition boards provide interface between the PC and the real-time operating systems.

6.2 Linear Track

Camera 2 is located on a linear track as shown in Figures 6.1 and 6.2. A DC motor moves the camera, while its position along the track is measured by incremental rotary encoders installed on the motor shafts. The encoders produce 4,096 counts per revolution which approximately yield a 0.01mm linear position measurement resolution. The device is controlled through a Proportional-Integral-Derivative (PID) controller under Tornado/VxWorks RTOS with a control rate of 512Hz.

6.3 Turn-Table

The object is held and positioned by a planar twin pantograph unit from Quanser Consulting which serves as the turn-table (see Figure 6.2). This device provides three active degrees of motion, i.e., two translations and one rotation using four electric motors; in our work only rotational movement is required. The motor shaft angles are measured by optical encoders that produce 20,000 counts per revolution. A Proportional-Derivative (PD) controller is used to control the angular position of the device based on its kinematics. This controller is implemented under Tornado/VxWorks RTOS with the control rate of 512Hz.

6.4 Pan-Tilt Units

The pan-tilt units shown in Figures 1.1 and 6.3, are PTU-D46-17.5 from Directed Perception, and provide precise control of position, speed and acceleration. Each unit includes two step motors for the pan and tilt control, with speed over 300°/second

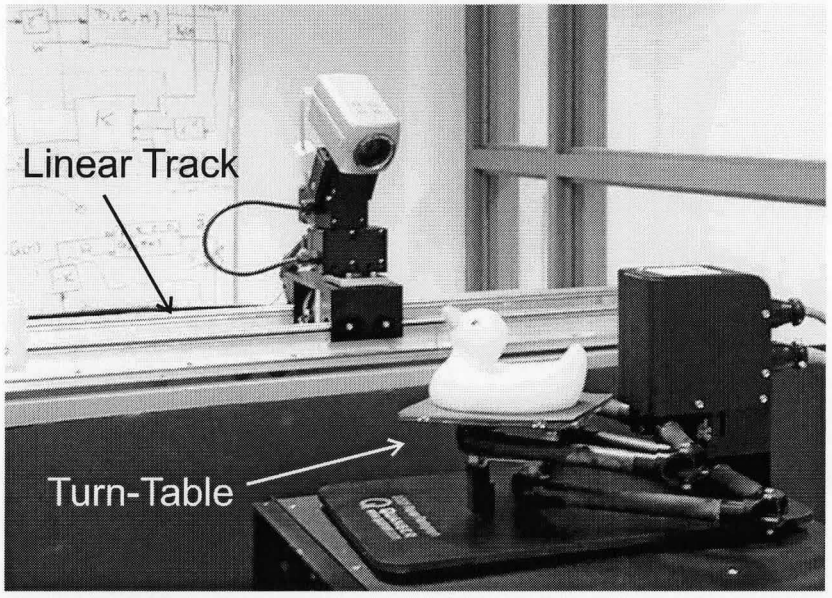

Figure 6.2: Camera 2 is positioned by a linear track; a twin pantograph unit serves as the turn-table.

and resolution of 3.086 arc minute(.0514^o). For communication, RS-232 serial port is employed between the pan-tilt and the controller PC, as seen in Figure 6.1. The RS-485 network can be used to control multiple pan-tilt units.

6.5 Cameras

The cameras used in the experiments are Sony DFW-VL500 model, Figure 6.3, a digital model which adopts the IEEE1394-1995 standard. In our work the cameras are adjusted to 15 fps and (120x160) pixel resolution, with the image color format YUV(4:4:4). Images are initially saved in the *ppm* format, and then converted to gray-scale *png* format in MATLAB, using the Image Processing Toolbox. The focus and zoom parameters are controlled based on predefined tables. These tables are

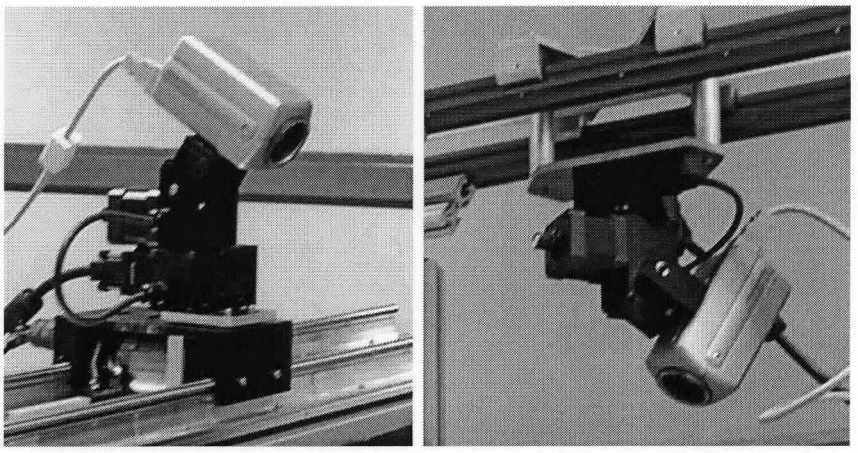

Figure 6.3: Pan-tilt units with the cameras attached; the left picture shows a sample positioning of Camera 2, whereas the picture on the right depicts the positioning of Camera 1.

defined in the beginning of the work in order to keep the focus and zoom values in each camera location constant throughout the experiments. Using this strategy, it is possible to avoid image segmentation, as further discussed in Chapter 7.

6.6 Software Development

The control algorithm has been implemented using the Matlab Realtime Workshop Toolbox and Tornado 2.2/VxWorks 5.5 RTOS by WindRiver. The control rate is set to 512Hz. Q8 multifunction I/O boards by Quanser Consulting collect sensory information and output the control commands. The image acquisition is performed under Windows operating system, in C++ and using the CMU 1394 Digital Camera Driver library. The interface between Windows and VxWorks is maintained through a UDP protocol.

The communication protocol between the Windows and VxWorks sides is as follows (see Figure 6.1); The Windows machine sends the desired pan and tilt angles for Camera 1 and the pan angle for Camera 2 to the VxWorks machine. These parameters are then converted into Cartesian locations, as discussed in Appendix C, and proper control commands are sent to the equipment. After the cameras are located in the desired positions, the VxWorks computer sends a *capture image* signal to the Windows computer. The images are taken based on commands sent from Windows to the cameras. These images are processed, and through updating the pdf of the object states it is decided if enough information is collected. If there is enough evidence to decide about the identity/pose of the object in front of the cameras, the algorithm exits and the final decision about the object identity is declared. On the other hand, if enough information is not collected, the Windows side continues by running the view-planning algorithm. The best next camera positions are chosen and sent to the VxWorks side; this cycle is repeated for at most seven sensory actions.

6.7 Summary

The experimental setup along with the algorithms for implementing the experiments were discussed. In Chapter 7, the results of various experiments will be given and comparisons among the results of different approaches will be made.

Chapter 7

Experimental Results

Experiments have been conducted to evaluate the effectiveness of the proposed occlusion modelling and multi-camera object recognition/pose estimation algorithms. The experimental setup, described in Chapter 6, is shown in Figure 6.1. Eight objects are considered in the experiments with 4 different pose angles each 90 degrees apart. The objects are shown in Figure 7.1. It should be noted that the object pairs (5,6) and (7,8) appear exactly similar from most viewing angles. Such selection of objects can demonstrate the advantage of active recognition over passive recognition as the ambiguity in the object identity may not be resolved by a single image taken from a fixed viewpoint. Such scenarios can also motivate the use of multiple cameras which can provide simultaneous views of the object from different viewing angles.

The images used in training and recognition experiments are 8-bit gray-scale. Most appearance-based methods segment and scale the collected images before the application of the PCA. While this approach works well for single objects against a uniform background, it could break down in the presence of structured noise. In particular, occlusions that extend beyond the object border can significantly alter

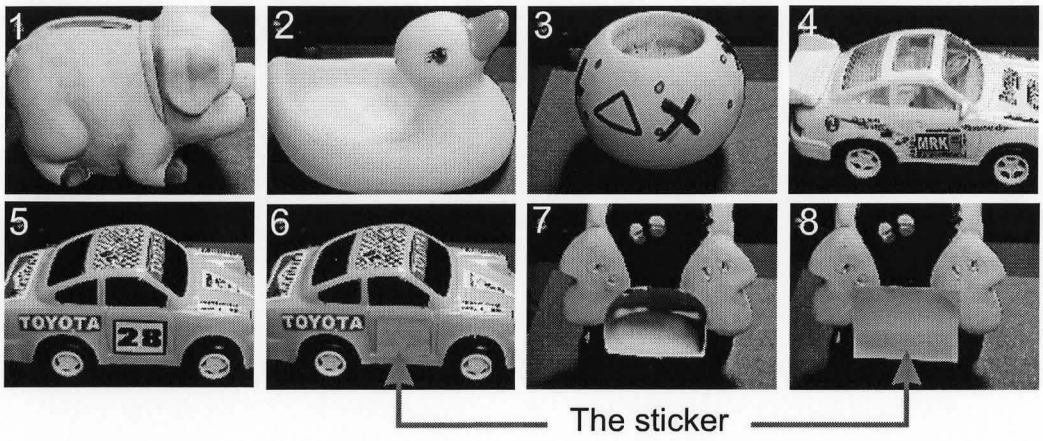

Figure 7.1: The objects used in the experiments.

the segmentation outcome. In this paper, we set the camera zoom parameter to an *a priori* known value at each camera position. These values have been chosen such that most relevant information is acquired from the object of interest while minimizing the effect of outliers in the scene.

Two training sets have been constructed by taking 32 images in each camera location from the eight objects at four different poses. This has been repeated for 32 locations of Camera 1 and seven locations of Camera 2, resulting in 1248 images in total. In the first set, each image has been populated to 100 new images by adding Gaussian brightness noise and random translational pixel shifts. The second set includes the occluded images generated according to the procedure outlined in Chapter 4. The parameters of the Gaussian distributions for the eigenspace coefficients in (4.18) and (4.19) are estimated using the Maximum Likelihood (ML) estimation. Four sets of experiments have been conducted as follows.

7.1 Multi-camera Recognition of Non-occluded Objects

In this case, the multi-camera recognition algorithms were employed to estimate class/pose of non-occluded objects. The first database that does not include the occlusion model was used in this experiment. For each object class, approximately 100 Monte Carlo recognition experiments have been conducted using three different strategies of MI-based, CRLB-based, as well as random movements of the cameras. The recognition process is terminated if the maximum probability of the object class/pose reaches the threshold $P_{th} = 0.95$ or after a maximum number of seven steps. A recognition is declared successful if both identity and pose are correctly estimated.

The results of the first set of experiments are summarized in Tables 7.1, 7.2, and 7.3, where the average number of steps for object recognition/pose estimation, the average probability of recognition, i.e., the highest object/pose probability, and the recognition/pose estimation success rate are given. As can be seen from the data, the CRLB and MI methods perform similarly and can successfully classify the objects in almost all cases. The success rate is lower in the random strategy particularly in the cases of the ambiguous object pairs (5,6) and (7,8). The average number of steps required for recognition is also higher when using the random camera movement approach compared to those of the two other methods.

One explanation for the acceptable results of the random movement approach is the use of a rich feature vector such as eigenspace coefficients. Similar observation has been reported in [126]. The use of weaker features such as the quantized mean

gray value could highlight the difference between the the random strategy and the other two sensor planing approaches.

Mutual Information Method

Object	Recog. Rate(%)	Mean exit prob.	Mean no. of views
1	100.0	0.98	1.6
2	100.0	0.99	1.9
3	100.0	0.99	2.2
4	100.0	0.98	1.6
5	100.0	0.97	3.5
6	100.0	0.97	3.6
7	100.0	0.83	4.3
8	97.9	0.86	4.7
Average	99.7	.95	2.9

Table 7.1: The results of the first set of experiments, based on MI maximization.

CRLB Method

Object	Recog. Rate(%)	Mean exit prob.	Mean no. of views
1	100.0	0.99	1.8
2	100.0	0.98	1.8
3	100.0	0.98	2.3
4	100.0	0.97	1.9
5	100.0	0.97	3.8
6	100.0	0.96	4.2
7	100.0	0.84	4.9
8	99.0	0.81	4.5
Average	99.9	.94	3.2

Table 7.2: The results of the first set of experiments, based on CRLB minimization.

7.2 Multi-camera Recognition of Occluded Objects

These experiments are similar to those in the previous case with the exception that during the recognition process the objects are occluded with pieces of cardboard. The

Random Method			
Object	Recog. Rate(%)	Mean exit prob.	Mean no. of views
1	100.0	0.96	2.8
2	100.0	0.99	3.3
3	90.6	0.98	4.7
4	96.9	0.91	2.8
5	90.6	0.83	5.5
6	89.6	0.92	6.6
7	82.3	0.80	5.9
8	76.0	0.79	6.5
Average	90.8	.90	4.8

Table 7.3: The results of the first set of experiments, moving the cameras randomly.

algorithms still use the first database without the occlusion model. The occlusions are such that they corrupt the images of both cameras at their initial positions and therefore successful recognition in the first step is unlikely. The occlusion in the images are of random shapes, and do not necessarily fit in any predetermined class of shapes. However, the occlusion percentages in the images are comparable to the values used in the second training set. The results of these experiments are summarized in Tables 7.4 and 7.5.

Significant degradation in the performance is observed as seen from the success rate column. This is mostly due to the sensitivity of the appearance-based modelling to structured noise as expected. The results of the two strategies are comparable in this case as well.

Mutual Information Method			
Object	Recog. Rate(%)	Mean exit prob..	Mean no. of views
1	45.8	.97	4.2
2	39.6	.98	4.1
3	53.1	.96	3.6
4	51.0	.99	2.1
5	42.7	.97	3.9
6	57.3	.95	3.7
7	58.3	.92	3.3
8	49.0	.86	4.4
Average	49.6	.95	3.7

Table 7.4: The results of the second set of experiments, based on MI maximization.

CRLB Method			
Object	Recog. Rate(%)	Mean exit prob..	Mean no. of views
1	43.8	.99	3.0
2	34.4	.98	3.3
3	47.9	.99	2.9
4	58.3	.98	3.2
5	49.0	.99	2.9
6	51.0	.99	3.5
7	52.1	.89	3.9
8	47.9	.94	4.1
Average	48.1	.97	3.4

Table 7.5: The results of the second set of experiments, based on CRLB minimization.

7.3 Multi-camera Recognition of Occluded Objects with Occlusion Model

The third set of experiments were conducted with our proposed algorithms that incorporate the levels of occlusion into the system states in (3.6). The second training set was used in the experiments while the recognition scenarios and the occlusions remained unchanged from Case (ii). According to the results in Tables 7.6 and 7.7, a

significant improvement is observed in the success rates of both MI-based and CRLB-based approaches compared with those of the previous case. This enhancement has been gained by incorporating a model of occlusion into the algorithms.

A particularly challenging scenario is depicted in Figure 7.2, in which Object 6 from the ambiguous pair of (5,6) is occluded in both camera views using two pieces of cardboard one from the side and one folded from the top. The discriminating side that carries the sticker can only be viewed by Camera 1 and even then only in a limited number of positions. The system correctly identifies the object class/pose after taking four pairs of images. The object probabilities summed over different poses are given in Table 7.8 for each recognition step. It is interesting to note the evolution of the classification probabilities through the recognition steps. The discriminating factor between Objects 5 and 6, the sticker, is absent of from the images of both cameras in the first step. This results in an initial confusion as to the identity/position of the object. However the system positions Camera 1 in a suitable place to make the recognition possible. After the sticker is exposed to Camera 1 in the second step, it is kept in the view of this camera until the system successfully recognizes Object 6 at Step 4.

7.4 Multi-camera Recognition vs. Single-Camera Recognition

To investigate the advantages of the multi-camera recognition approach over a single-camera one, we have applied the MI recognition algorithm to a single-camera case for non-occluded objects using Camera 1. The average mean number of steps before

Mutual Information Method			
Object	Recog. Rate(%)	Mean exit prob..	Mean no. of views
1	100.0	0.95	3.7
2	100.0	0.96	3.5
3	100.0	0.91	4.3
4	100.0	0.97	4.2
5	95.8	0.85	4.4
6	100.0	0.81	5.9
7	96.9	0.87	5.3
8	97.9	0.75	5.7
Average	98.8	0.88	4.6

Table 7.6: The results of the third set of experiment, based on MI maximization.

CRLB Method			
Object	Recog. Rate(%)	Mean exit prob..	Mean no. of views
1	100.0	0.95	3.6
2	100.0	0.96	3.3
3	99.0	0.83	4.1
4	100.0	0.94	2.9
5	97.1	0.81	4.9
6	97.9	0.88	5.2
7	95.8	0.86	5.3
8	99.0	0.71	6.1
Average	98.6	0.87	4.4

Table 7.7: The results of the third set of experiment, based on CRLB minimization.

a successful recognition increased to 5.1 from 2.9 in Table 7.1, while the success rate was still high at 97.8%.

Repeating the experiments with occluded objects in Case (iii) and incorporating only images from Camera 1, the recognition rate dropped to 66.6%, with an average number of steps of 5.8, and an average exit probability of 0.74 for the correctly classified objects. These numbers clearly demonstrate the significance of adding the the second camera specially when dealing with occluded objects.

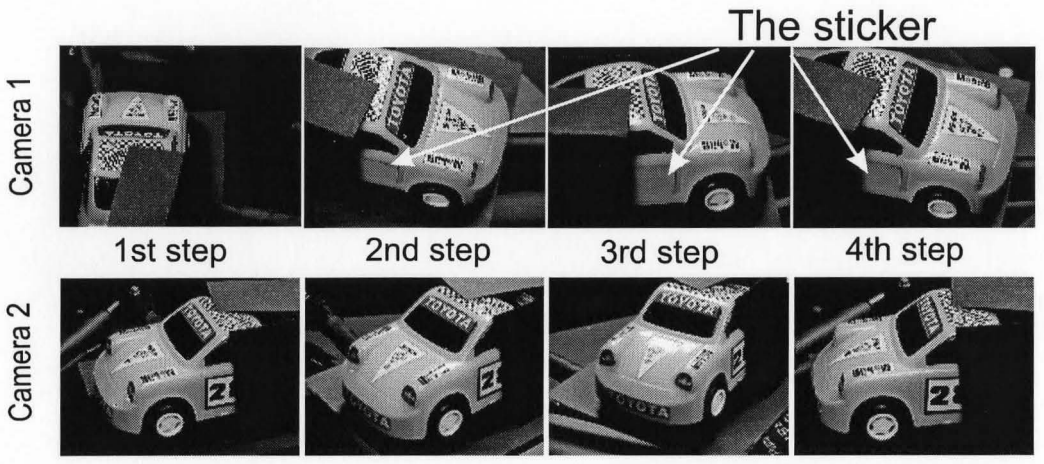

Figure 7.2: A trial for recognizing Object 6.

Step	Initial prob.	Step 1	Step 2	step 3	Step 4
Obj. 1	.1250	.1065	.0040	.0001	0.000
Obj. 2	.1250	.0333	.0007	0.000	0.000
Obj. 3	.1250	.0090	0.000	0.000	0.000
Obj. 4	.1250	.2062	.0791	.0062	.0008
Obj. 5	.1250	.3722	.3912	.1993	.0119
Obj. 6	.1250	.2553	.5250	.7944	.9873
Obj. 7	.1250	.0086	0.000	0.000	0.000
Obj. 8	.1250	.0088	0.000	0.000	0.000

Table 7.8: Object probabilities for the trial in Figure 7.2.

7.5 Summary

In this chapter, the results of extensive experiments were given. It was shown that the proposed approaches, i.e. active camera movement, multi-camera data fusion, and occlusion modelling, all have significant roles in enhancing the recognition. This improvement is evident from the increase in the recognition rate, the decrease in the number of steps required for the recognition, as well as more robustness with respect to structured noise, i.e. occlusion. Comparable results were gained from the MI and

CRLB metrics, although the CRLB method requires more processing time.

Chapter 8

Conclusions and Future Work

In this thesis, we studied the problem of active appearance-based object recognition/pose estimation. The main contributions of this work can be outlined as follows,

- While relevant prior work in the literature use single camera images, in this thesis an optimal *multi-camera* active recognition algorithm was developed.
- The information collected by multiple cameras was incorporated into performance measures based on the Mutual Information and the Cramér-Rao Lower Bound. This is the first time the CRLB has been used for active object recognition.
- The Principle Component Analysis was employed to process the images and produce the sensor observation vectors in the eigenspace. To reduce sensitivity with respect to structured noise, a model of occlusion was incorporated into the database by corrupting original images. Also, a novel definition for the object states was introduced that incorporates occlusion as time-variable terms in the state.

- Extensive experiments were conducted with a two-camera system, to assess the performance of the proposed algorithms. It has been evident from the results of these experiments that the features of proposed methodology, i.e., sensor planing, sensor fusion, and occlusion modelling can all significantly enhance the outcome of the recognition process. This improvement includes increment in the recognition rate, decrement in the number of steps required for recognition, and more robustness with respect to structured noise, i.e. occlusion. The two methods based on the MI and CRLB performed similarly in all scenarios. However, the MI-based requires less computations during the recognition stage as most of its calculations are performed off-line.

The following directions may be pursued for future research,

- The sensor models can be revised in order to consider factors such as correlation between images of the two cameras, occlusions in the images of the two cameras, as well as occlusions from one viewpoint to another viewpoint.

- The limiting assumption of constant *a priori* known camera zoom used in this paperwork shall be relaxed.

- The problem of localizing the object before recognition can be considered by relaxing the assumption of constant *a priori* known object position in the workspace.

- The problem can be expanded to the case of recognizing multiple objects with the possibility of them occluding each other.

- The context of tracking moving objects can be considered also, as an extension of the object recognition case.

Appendix A

Information Theory Related Concepts

A.1 Definitions

Entropy: The entropy $H(x)$ of a discrete random variable x is defined as [126],

$$H(x) = -\sum p(x) \log p(x) \tag{A.1}$$

Conditional Entropy: The conditional entropy $H(x|y)$ for two random variable x and y, when $(x, y) \sim p(x, y)$, is defined as

$$H(x|y) = -\sum p(y_0) H(x|y = y_0) = -\sum \sum p(x, y) \log p(x|y) \tag{A.2}$$

Mutual Information: The mutual information, indicates the reduction in the uncertainty of one random variable, due to information embedded in the other one,

$$I(x;y) = \sum\sum p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = H(x) - H(x|y) \tag{A.3}$$

Conditional Mutual Information: The conditional mutual information of the two random variables x and y, given z is given as,

$$I(x;y|z) = H(x|z) - H(x|y,z) \tag{A.4}$$

A.2 Proof of Equation (5.5)

$$\begin{aligned}
&I(s;g|a) = H(s|a) - H(s|g,a)\\
&= -\sum_s P(s|a)\log P(s|a) + \sum_s \int_g p(s,g|a)\log(p(s|g,a))\\
&= -\sum_s P(s|a)\log P(s|a) + \sum_s \int_g P(s)p(g|s,a)\log(p(s|g,a))\\
&= \sum_s P(s|a)\left[-\log P(s|a) + \int_g p(g|s,a)\log p(s|g,a)\right]\\
&= \sum_s P(s|a)\left[-\int_g p(g|s,a)\log P(s|a) + \int_g p(g|s,a)\log p(s|g,a)\right]\\
&= \sum_s P(s|a)\int_g p(g|s,a)\log\left(\frac{p(s|g,a)}{P(s|a)}\right)
\end{aligned} \tag{A.5}$$

and since,

$$p(s|g,a) = \frac{p(s,g|a)}{p(g|a)} = \frac{P(s|a)p(g|s,a)}{p(g|a)} \tag{A.6}$$

we have,

$$I(s;g|a) = \sum_s P(s|a)\int_g p(g|s,a)\log\left(\frac{p(g|s,a)}{p(g|a)}\right) \tag{A.7}$$

A.3 Proof of Convergence for the Sequential Decision Making Process

The Mutual Information minimization performed throughout the active vision process has been originally used in [9] for single sensor vision, and has been expanded to the case of multi-sensor vision in this thesis. The authors in [9] have provided the proof for the convergence of the approach. For that purpose the following theorem and corollaries will be required (corollary 1 is required to prove corollary 2),

Theorem 1: The sequential decision process forms a Markov chain.

Proof: A Markov chain $(S, G, A, p_{tr}(s|a, s^{'}))$ is defined and utilized, where S is the set of states $\{s_1, s_2, ..., s_n\}$, G is the set of possible observations $\{g_1, g_2, ..., g_m\}$, A is the set of control action $\{a_1, a_2, ..., a_l\}$, and the state transition probability is defined as,

$$p_{tr}(s|a, s^{'}) = \sum_{s} p(s^{'}|s, a, g)p(g|s, a) \tag{A.8}$$

where,

$$p(s^{'}|s, a, g) = \begin{cases} 1 & s^{'} = \underset{s}{argmax}\ p(s|a, g) \\ 0 & otherwise \end{cases} \tag{A.9}$$

$p(s|a, g)$ is the *a posteriori* pdf of the state after observing g under control action a. Evidently, the state transition is performed only if the maximum in the *a posteriori* pdf switches from one class to the other. The action is chosen based on a logic similar to the one in (5.4)

Corollary 1: Assume the two distributions $p_1(x_n)$ and $p_2(x_n)$ on a Markov chain

at time n, the Kullback-Leibler distance between the two distributions will never increase over time,

$$D(p_1(x_n) \parallel p_2(x_n)) \geq D(p_1(x_{n+1}) \parallel p_2(x_{n+1})) \tag{A.10}$$

where the Kullback-Leibler distance between two probability mass function $p(x)$and $q(x)$ is defined as,

$$D(p \parallel q) = \sum p(x) \log \frac{p(x)}{q(x)} \tag{A.11}$$

Corollary 2: Relative entropy between a distribution $p(x_n)$ and a stationary distribution $p(x)$ of a Markov chain, $D(p(x_n) \parallel p(x))$, decreases with n (time step).

Theorem 2:(Convergence for the Sequential Decision Making Process) the sequential decision process, based on Bayes rule for calculating the *a posteriori* probability and the maximum Mutual Information criterion for the purpose of sensor action selection, will converge.

Proof: Based on Theorem 1, the sequential decision making process forms a Markov chain. From corollary 2, it can be concluded that the distribution over the state of the Markov chain converges to a point, where the distance to all stationary distributions of the Markov chain in minimized. In case of unique stationary distribution, the final distribution will be equal to the stationary distribution.

It can be argued whether the Markov chain has only one stationary distribution. For example, in cases like having two copies of the same object in the data base it could be expected to have two stationary distributions. This will cause the sequential decision making process to converge to a point with minimum distance to both of

these distributions. Whether the stationary distribution of the Markov chain is the right distribution or not remains unsolved, while the correct distribution is assumed to be

$$P(x_i) = \begin{cases} 0 & i \neq n \\ 1 & i = n \end{cases} \tag{A.12}$$

when object number n is in front of the camera.

Appendix B

Proof of Cramér-Rao Lower Bound (CRLB)

Consider g as the observation vector, s as the non-random scalar parameter to be estimated, and $\hat{s}(g)$ as an unbiased estimate of it [122],

$$E[\hat{s}(g)] = s \tag{B.13}$$

The likelihood function $p(g|s)$, is assumed to have absolutely integrable first and second derivatives with respect to x. From (B.13) we have,

$$\int_{-\infty}^{+\infty} (\hat{s}(g) - s)p(g|s)dg = 0 \tag{B.14}$$

which yields,

$$\frac{d}{ds}\int_{-\infty}^{+\infty} (\hat{s}(g) - s)p(g|s)dg = 0 \tag{B.15}$$

and as a result,

$$\int_{-\infty}^{+\infty}(\hat{s}(g)-s)\frac{\partial p(g|s)}{\partial s}dg-\int_{-\infty}^{+\infty}p(g|s)dg=0 \tag{B.16}$$

where the second integral is equal to one and we have

$$\int_{-\infty}^{+\infty}(\hat{s}(g)-s)\frac{\partial p(g|s)}{\partial s}dg=1 \tag{B.17}$$

which can be rewritten as,

$$\int_{-\infty}^{+\infty}(\hat{s}(g)-s)\frac{\partial \log p(g|s)}{\partial s}p(g|s)dg=1 \tag{B.18}$$

and finally,

$$\int_{-\infty}^{+\infty}\left[(\hat{s}(g)-s)\sqrt{p(g|s)}\right]\left[\frac{\partial \log p(g|s)}{\partial s}\sqrt{p(g|s)}\right]dg=1 \tag{B.19}$$

We define,

$$\begin{aligned} F&=(\hat{s}(g)-s)\sqrt{p(g|s)}\\ G&=\frac{\partial \log p(g|s)}{\partial s}\sqrt{p(g|s)} \end{aligned} \tag{B.20}$$

For real valued functions u and v, the *Schwarz* inequality is defined as,

$$\langle u,v\rangle \leq \|u\|\|v\| \tag{B.21}$$

where,

$$\langle u, v \rangle = \int_{-\infty}^{+\infty} u(t)v(t)dt \tag{B.22}$$

and

$$\|u\| = [\langle u, u \rangle]^{1/2} \tag{B.23}$$

Applying the *Schwarz* inequality to (B.19) will result in

$$1 \leq \left[\int_{-\infty}^{+\infty} (\hat{s}(g) - s)^2 p(g|s)dg\right]^{1/2} \left[\int_{-\infty}^{+\infty} \left(\frac{\partial \log p(g|s)}{\partial s}\right)^2 p(g|s)dg\right]^{1/2} \tag{B.24}$$

consequently,

$$\int_{-\infty}^{+\infty} (\hat{s}(g) - s)^2 p(g|s)dg \geq \left[\int_{-\infty}^{+\infty} \left(\frac{\partial \log p(g|s)}{\partial s}\right)^2 p(g|s)dg\right]^{-1} \tag{B.25}$$

or in other words,

$$E\left[(\hat{s}(g) - s)^2\right] \geq \left\{E\left[\left(\frac{\partial \log p(g|s)}{\partial s}\right)^2\right]\right\}^{-1} \tag{B.26}$$

which is identical to (5.15), and completes the proof for CRLB for a non-random scalar parameter [122].

Appendix C

The Workspace Analysis

The outputs of the sensor planning algorithm are the pan and tilt angles for the first unit and the pan angle for the second unit. These angles are then converted into Cartesian locations of both cameras, as well as the tilt angle for the second camera. These parameters are calculated based on the structure of the workspace. Assume indices $c1$, $c2$ and o represent the locations corresponding to Camera 1, 2 and the object, respectively. Also let θ and ν to be the pan and tilt angles for the pan-tilt units, respectively. The following equations can be written for the first camera,

$$\tan\theta_{c1} = \Delta x_{c1}/\Delta y_{c1} \tag{C.27}$$

$$\tan\nu_{c1} = \Delta z_{c1}/(\Delta y_{c1}^2 + \Delta x_{c1}^2)^{1/2} \tag{C.28}$$

where,

$$\Delta x_{c1} = x_{c1} - x_o \tag{C.29}$$

$$\Delta y_{c1} = y_{c1} - y_o \tag{C.30}$$

$$\Delta z_{c1} = z_{c1} - z_o \tag{C.31}$$

and parameters x, y and z are shown in Figure C.1. The object is always held in a constant location by the turn-table. None of the cameras can move along the z axis, Camera 2 is not provided with any movement along the x axis too. These constraints result in predefined and known object location, as well as constant z_{c1}, z_{c2}, and x_{c2}. This results in the following reformulation of (C.27) and (C.28), writing the unknown parameters in terms of the known ones,

$$x_{c1} = x_o + \Delta z_{c1} \cos\theta_{c1} / \tan\nu_{c1} \tag{C.32}$$

$$y_{c1} = y_o + \Delta z_{c1} \cos\theta_{c1} / (\tan\nu_{c1} \tan\theta_{c1}) \tag{C.33}$$

And for the second camera shown in Figure C.1, one may write,

$$\tan\theta_{c2} = \Delta y_{c2} / \Delta x_{c2} \tag{C.34}$$

$$\tan\nu_{c2} = \Delta z_{c2} / (\Delta y_{c2}^2 + \Delta x_{c2}^2)^{1/2} \tag{C.35}$$

where Δx_{c2}, Δy_{c2}, and Δz_{c2} are defined as,

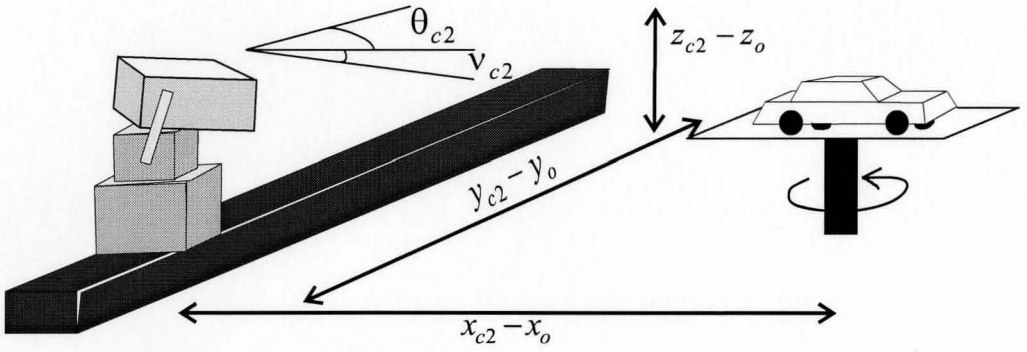

Figure C.1: The pan and tilt angles for camera number two are shown, as well as the quantities Δx_{c2}, Δy_{c2}, and Δz_{c2}.

$$\Delta x_{c2} = x_{c2} - x_o \tag{C.36}$$

$$\Delta y_{c2} = y_{c2} - y_o \tag{C.37}$$

$$\Delta z_{c2} = z_{c2} - z_o \tag{C.38}$$

Consequently we have,

$$y_{c2} = y_o + \tan\theta_{c2}\Delta x_{c2} \tag{C.39}$$

$$\nu_{c2} = \arctan(\Delta z_{c2}/(\Delta y_{c2}^2 + \Delta x_{c2}^2)^{1/2}) \tag{C.40}$$

which gives the unknown parameters in terms of the known ones.

Bibliography

[1] S. Chen and Y. Li, "A Method of Automatic Sensor Placement for Robot Vision in Inspection Tasks," in *Proc. IEEE Int. Conf. on Robotics and Automation*, vol. 3, pp. 2545–2550, 2002.

[2] B. Nelson and P. Khosla, "Integrating Sensor Placement and Visual Tracking Strategies," in *Proc. of IEEE Conf. on Robotics and Automation*, vol. 2, pp. 1351–1356, 1994.

[3] M. J. Black and A. D. Jepson, "EigenTracking: Robust Matching and Tracking of Articulated Objects Using A View-Based Representation," *Int. Journal of Computer Vision*, vol. 26, pp. 63–84, 1998.

[4] B. Schiele and J. Crowley, "Probabilistic Object Recognition Using Multidimensional Receptive Field Histograms," in *Proc. of the 13th Int. Conf. on Pattern Recognition*, vol. 2, pp. 50–54, 1996.

[5] T. Arbel and F. Ferrie, "On the Sequential Accumulation of Evidence," *Int. Journal of Computer Vision*, vol. 43, pp. 205–230, 2001.

[6] H. Murase and S. Nayar, "Learning and Recognition of 3D Objects from Appearance," in *Proc. of IEEE Workshop on Qualitative Vision*, pp. 39–50, 1993.

[7] S. J. Dickinson, A. Pentland, and A. Rosenfeld, "Qualitative 3D Shape Reconstruction Using Distributed Aspect Graph Matching," in *Proc. of the Third Int. Conf. on Computer Vision*, pp. 257–262, 1990.

[8] A. Selinger and R. Nelson, "Appearance-Based Object Recognition Using Multiple Views," in *Proc. of the 2001 IEEE Comput. Society Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. I–905–I–911, 2001.

[9] J. Denzler and C. M. Brown, "Information Theoretic Sensor Data Selection for Active Object Recognition and State Estimation," *IEEE Tran. on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 145–157, 2002.

[10] X. Lin, Y. Bar-Shalom, and T. Kirubarajan, "Multisensor-Multitarget Bias Estimation for General Asynchronous Sensros," in *Proc. of the Seventh Int. Conf. on Information Fusion*, pp. 243–250, 2004.

[11] l. Leibowicz, P. Nicolas, and L. Ratton, "Radar/ESM Tracking of Constant Velocity Target: Comparison of Batch (MLE) and EKF Performance," in *Proc. of the Third Int. Conf. on Information Fusion*, pp. TuC2–3–TuC2–8, 2000.

[12] S. Roy, S. Chaudhury, and S. Banerjee, "Active Recognition through Next View Planning: A Survey," *Pattren Recognition*, vol. 37, pp. 429–446, 2004.

[13] B. Schiele and J. Crowley, "Where to Look Next and What to Look for," in *Proc. of the 1996 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp. 1249–1255, 1996.

[14] D. Wilkes and J. Tsotsos, "Active Object Recognition," in *Proc. of IEEE Comput. Society Conf. on Computer Vision and Pattern Recognition*, pp. 136–141, 1992.

[15] B. Schiele and J. Crowley, "Recognition without Correspondence Using Multidimensional Receptive Field Histograms," *International Journal of Computer Vision*, vol. 36, pp. 31–50, 2000.

[16] X. He, B. Benhabib, K. Smith, and R. Safaee-Rad, "Optimal Camera Placement for An Active Vision System," in *Proc. of IEEE Int. Conf. on Systems, Man, and Cybernetics*, vol. 1, pp. 69–74, 1991.

[17] S. Sclaroff and A. P. Pentland, "Modal Matching for Correspondence and Recognition," *IEEE Trans. on Pattern analysis and Machine Intelligence*, vol. 17, pp. 545–561, 1995.

[18] V. Ferrari, T. Tuytelaars, and L. V. Gool, "Appearance-Based Object Recognition Using Multiple Views," in *Proc. of The 2004 IEEE Comput. Society Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. II–105 – II–112, 2004.

[19] J. Matas, D. Koubaroulis, and J. Kittler, "The Multimodal Neighbourhood Signature for Modeling Object Color Appearance and Applications in Object Recognition and Image Retrieval," *Computer Vision and Image Understanding*, vol. 88, pp. 1–23, 2002.

[20] T. Arbel, P. Whaite, and F. Ferrie, "Recognizing Volumetric Objects in The Presence of Uncertainty," in *Proc. of the 12th Int. Conf. on Pattern Recognition*, pp. 470–476, 1994.

[21] D. Keren, D. Cooper, and J. Subrahmonia, "Describing Complicated Objects by Implicit Polynomials," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 38–53, 1994.

[22] A. Pentland and S. Sclaroff, "Closed-form Solutions for Physically Based Shape Modeling and Recognition," *IEEE Trans. on Pattern analysis and Machine Intelligence*, vol. 13, pp. 715–729, 1991.

[23] W. Grimson, "On The Recognition of Curved Objects," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 632–643, 1989.

[24] W. E. L. Grimson and T. Lozano-Prez, "Localizing Overlapping Parts by Searching The Interpretation Tree," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 9, pp. 469–482, 1987.

[25] Y. Lamdan, J. Schwatrtz, and H. Wolfson, "On Recognition of 3D Objects from 2D Images," in *Proc. of IEEE Int. Conf. on Robotics and Automation*, pp. 243–250, 1988.

[26] P. Flynn and A. Jain, "BONSAI: 3D Object Recognition Using Constrained Search," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 1066–1075, 1991.

[27] A. Jain and R. Hoffman, "Evidence-Based Recognition of 3D Objects," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 10, pp. 783–802, 1988.

[28] W.-Y. Kim and A. Kak, "3D Object Recognition Using Bipartite Matching Embedded in Discrete Relaxation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 224–251, 1991.

[29] T. Fan, G. Medioni, and R. Nevatia, "Recognizing 3D Objects Using Surface Descriptions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 1140–1157, 1989.

[30] I. Biederman, "Human Image Understanding: Recent Research and A Theory," in *Papers from the second workshop Vol. 13 on Human and Machine Vision II*, pp. 13–57, 1986.

[31] R. Bergevin and M. Levine, "Generic Object Recognition: Building and Matching Coarse Descriptions from Line Drawings," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 19–36, 1993.

[32] S. Dickinson, H. Christensen, J. Tsotsos, and G. Olofsson, "Active Object Recognition Integrating Attention and Viewpoint Control," *Computer Vision Image Understanding*, vol. 67, pp. 239–260, 1997.

[33] C. Huang, O. Camps, and T. Kanungo, "Object Recognition Using Appearance-Based Parts and Relations," in *Proc. of the 1997 Conf. on Computer Vision and Pattern Recognition*, pp. 877–883, 1997.

[34] S. Roy, S. Chaudhury, and S. Banerjee, "Isolated 3D Object Recognition Through Next View Planning," *IEEE Trans. on Systems, Man and Cybernetics, Part A*, vol. 30, pp. 67–76, 2000.

[35] S. Hutchinson and A. Kak, "Planning Sensing Strategies in A Robot Work Cell With Multi-Sensor Capabilities," *IEEE Trans. on Robotics and Automation*, vol. 5, pp. 765–783, 1989.

[36] S. Roy, S. Chaudhury, and S. Banerjee, "Aspect Graph Construction with Noisy Feature Detectors," *IEEE Trans. on Systems, Man and Cybernetics, Part B*, vol. 33, pp. 340–351, 2003.

[37] D. Eggert and K. Bowyer, "Computing the Orthographic Projection Aspect Graph of Solids of Revolution," in *Proc. of Workshop on Interpretation of 3D Scenes*, pp. 102–108, 1989.

[38] D. Eggert, K. Bowyer, C. Dyer, H. Christensen, and D. Goldgof, "The Scale Space Aspect Graph," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 1114–1130, 1993.

[39] C. Cyr and B. Kimia, "3D Object Recognition Using Shape Similiarity-Based Aspect Graph," in *Proc. of the Eighth IEEE Int. Conf. on Computer Vision*, vol. 1, pp. 254–261, 2001.

[40] M. Swain and D. Ballard, "Indexing Via Color Histograms," in *Proc. of the Third Int. Conf. on Computer Vision*, pp. 390–393, 1990.

[41] B. Schiele and J. Crowley, "Transinformation for Active Object Recognition," in *the Sixth Int. Conf. on Computer Vision*, pp. 249–254, 1998.

[42] T. Cootes, D. Cooper, C. Taylor, and J. Graham, "Trainable Method of Parametric Shape Description," *Image and vision Computing*, vol. 10, pp. 289–294, 1992.

[43] A. Baumberg and D. Hogg, "Learning Flexible Models from Image Sequences," in *Proc. of the Third European Conf. on Computer Vision*, pp. 299–308, 1994.

[44] F. Bookstein, "Principal Warps: Thin-Plate Splines and The Decomposition of Deformations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 567–585, 1989.

[45] D. Paulus, C. Drexler, M. Reinhold, M. Zobel, and J. Denzler, "Active Computer Vision System," in *Proc. of the Fifth IEEE Int. Workshop on Computer Architectures for Machine Perception*, pp. 18–27, 2000.

[46] D. Schuurman and D. Capson, "Robust Direct Visual Servo Using Network-Synchronized Cameras," *IEEE Trans. on Robotics and Automation*, vol. 20, pp. 319–334, 2004.

[47] J. Krumm, "Eigenfeatures for Planar Pose Measurement of Partially Occluded Objects," in *Proc. of IEEE Comput. Society Conf. on Computer Vision and Pattern Recognition*, pp. 55–60, 1996.

[48] H. Murase and M. Lindenbaum, "Partial Eigenvalue Decomposition of Large Images Using Spatial Temporal Adaptive Method," *IEEE Tran. on Image Processing*, vol. 4, pp. 620–628, 1995.

[49] M. Covell and C. Bregler, "Eigen-points," in *Proc. IEEE Int. Conf. on Image Processing*, vol. 3, pp. 471–474, 1996.

[50] U. Ahlrichs, J. Fischer, J. Denzler, C. Drexler, H. Niemann, E. Noth, and D. Paulus, "Knowledge Based Image and Speech Analysis for Service Robots,"

in *Proc. of Integration of Speech and Image Understanding*, vol. 26, pp. 21–47, 1999.

[51] M. Turk and A. Pentland, "Face Recognition Using Eigenfaces," in *Proc. of IEEE Comput. Society Conf. on Computer Vision and Pattern Recognition*, pp. 586–591, 1991.

[52] R. Campbell and P. Flynn, "Eigenshapes for 3D Object Recognition in Range Data," in *Proc. of IEEE Comput. Society Conf. on Computer Vision and Pattern Recognition*, pp. 505–510, 1999.

[53] H. Murase and S. Nayar, "Illumination Planning for Object Recognition Using Parametric Eigenspaces," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 1219–1227, 1994.

[54] D. P. Valaparla and V. Asari, "An Adaptive Technique for The Extraction of Object Region and Boundary from Images with Complex Environment," in *the 30th Applied Imagery Pattern Recognition Workshop*, pp. 194–199, 2001.

[55] S. Yoshimura and T. Kanade, "Fast Template Matching Based on The Normalized Correlation by Using Multiresolution Eigenimages," in *Proc. of IROS'94*, pp. 2086–2093, 1994.

[56] J. Winkeler, B. Manjunath, and S. Chandrasekaran, "Subset Selection for Active Recognition," in *Proc. of IEEE Comput. Society Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 511–516, 1999.

[57] M. Turk and A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neoroscience*, vol. 3, pp. 71–86, 1991.

[58] S. Chandrasekaran, B. Manjunath, Y. Wang, J. Winkeler, and H. Zhang, "An Eigenspace Update Algorithm for Image Analysis," *Graphical Models and Image Processing*, vol. 59, pp. 321–332, 1997.

[59] A. Levy and M. Lindenbaum, "Sequential Karhunen-Loeve Basis Extraction and Its Application to Images," in *Proc. of IEEE Int. Conf. on Image Processing*, vol. 2, pp. 456–460, 1998.

[60] H. Bischof, H. Wildenauer, and A. Leonardis, "Illumination Insensitive Eigenspaces," in *Proc. of the Eighth IEEE Int. Conf. on Computer Vision*, vol. 1, pp. 233–238, 2001.

[61] H. Chen, P. Belhumeur, and D. Jacobs, "In Search of Illumination Invariants," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, 2000.*, vol. 1, pp. 254–261, 2000.

[62] A. Leonardis and H. Bischof, "Robust Recognition Using Eigenimages," *Computer Vision and Image Understanding*, vol. 78, pp. 99–118, 2000.

[63] R. Rao, "Dynamic Appearance-Based Recognition," in *Proc. of IEEE Comput. Society Conf. on Computer Vision and Pattern Recognition*, pp. 540–546, 1997.

[64] M. Tarr and S. Pinker, "Mental Rotation and Orientation-Dependence in Shape Recognition," *Cognitive Psychology*, vol. 21, pp. 233–282, 1989.

[65] L. Paletta and A. Pinz, "Active Object Recognition by View Integration and Reinforcement Learning," *Robotics and Autonomous Systems*, vol. 31, pp. 71–86, 2000.

[66] T. Jebara, K. Russell, and A. Pentland, "Mixtures of Eigenfeatures for Real-Time Structure from Texture," in *the Sixth International Conf. on Computer Vision*, pp. 128–135, 1998.

[67] M. Kirby and L. Sirovich, "Application of The Karhunen-Loeve Procedure for The Characterization of Human Faces," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 103–108, 1990.

[68] B. Moghaddam, W. Wahid, and A. Pentland, "Beyond Eigenfaces: Probabilistic Matching for Face Recognition," in *Proc. of the Third IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 30–35, 1998.

[69] M. Black and Y. Yacoob, "Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion," *Int. Journal of Computer Vision*, vol. 25, pp. 23–48, 1997.

[70] C. Bregler and Y. Konig, "Eigenlips for Robust Speech Recognition," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, pp. II/669–II/672, 1994.

[71] N. Li, S. Dettmer, and M. Shah, "Visually Recognizing Speech Using Eigen Sequences," *MBR97*, pp. 337–363, 1997.

[72] M. Kirby, "Low-Dimensional Processing of Still And Moving Images," in *Conf. Record of The Twenty-Sixth Asilomar Conf. on Signals, Systems and Computers*, pp. 1026–1030, 1992.

[73] A. Pentland, R. Picard, and S. Sclaroff, "Photobook: Content-Based Manipulation of Image Databases," *Int. Journal of Computer Vision*, vol. 18, pp. 233–254, 1996.

[74] L. Sirovich and M. Kirby, "Low Dimensional Procedure for The Characterization of Human Faces," *Journal of the Optical Society of America*, vol. 4, pp. 519–524, 1987.

[75] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711–720, 1997.

[76] F. Stein and G. Medioni, "Structural Indexing: Efficient 3D Object Recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14, pp. 125–145, 1992.

[77] P. Flynn and A. Jain, "3D Object Recognition Using Invariant Feature Indexing of Interpretation Tables," in *1991 Workshop on Directions in Automated CAD-Based Vision*, pp. 115–123, 1991.

[78] Y. Lamdan and H. Wolfson, "Geometric Hashing: A General And Efficient Model-based Recognition Scheme," in *the Second Int. Conf. on Computer Vision*, pp. 238–249, 1988.

[79] W. Grimson and D. Huttenlocher, "On The Sensitivity of Geometric Hashing," in *Proc. of the Third Int. Conf. on Computer Vision*, pp. 334–338, 1990.

[80] A. Kak, A. Vayda, R. Cromwell, and W. K. adn C. Chen, "Knowledge-Based Robotics," in *Proc. of 1987 IEEE Int. Conf. on Robotics and Automation*, pp. 637–646, 1987.

[81] T. Fan, G. Medioni, and R. Nevatia, "Segmented Descriptions of 3D Surfaces," *IEEE Journal of Robotics and Automation*, vol. 3, pp. 527–538, 1987.

[82] P. Flynn and A. Jain, "CAD-Based Computer Vision: from CAD Models to Relational Graphs," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 114–132, 1991.

[83] P. Viola and W. WellsIII, "Alignment by Maximization of Mutual Information," *International Journal of Computer Vision*, vol. 24, pp. 137–154, 1997.

[84] L. Paletta, M. Prantl, and A. Pinz, "Learning Temporal Context in Active Object Recognition Using Bayesian Analysis," in *Proc. of the 15th Int. Conf. on Pattern Recognition*, vol. 1, pp. 695–699, 2000.

[85] C. Bregler and S. Omohundro, "Surface Learning with Applications to Lipreading," in *Proc. of Neural Information Processing Systems*, pp. 43–50, 1993.

[86] A. Bobick and A. Wilson, "A State Based Approach to The Representation and Recognition of Gesture," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 1325–1337, 1997.

[87] S. Nayar, H. Murase, and S. Nene, "Learning, Positioning, and Tracking Visual Appearance," in *Proc. IEEE Int. Conf. on Robotics and Automation*, vol. 4, pp. 3237–3244, 1994.

[88] S. Nene and S. Nayar, "A Simple Algorithm for Nearest Neighbor Search in High Dimensions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 989–1003, 1997.

[89] T. Poggio and F. Girosi, "Networks for Approximation and Learning," in *Proc. of the IEEE Conf.*, vol. 78, pp. 1481–1497, 1990.

[90] H. Murase and S. Nayar, "Learning by a Generation Approach to Appearance-Based Object Recognition," in *Proc. of the 13th Int. Conf. on Pattern Recognition*, vol. 1, pp. 24–29, 1996.

[91] Y. Adini, Y. Moses, and S. Ullman, "Face Recognition: The Problem of Compensating for Changes in Illumination Direction," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 721–732, 1997.

[92] S. Nayar and H. Murase, "Dimensionality of Illumination in Appearance Matching," in *Proc. of IEEE Conf. on Robotics and Automation*, vol. 2, pp. 1326–1332, 1996.

[93] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 696–710, 1997.

[94] H. Borotschnig, L. Paletta, and A. Pinz, "Appearance-Based Active Object Recognition," *Image and Vision Computing*, vol. 18, pp. 715–727, 2000.

[95] H. Bischof and A. Leonardis, "Robust Recognition of Scaled Eigenimages through a Hierarchical Approach," in *Proc. of the Eighth IEEE Comput. Soc. Conf. on Computer Vision and Pattern Recognition*, pp. 664–670, 1998.

[96] H. Bischof, A. Leonardis, and F. Pezzei, "A Robust Subspace Classifier," in *Proc. of the 14th Int. Conf. on Pattern Recognition*, vol. 1, pp. 114–116, 1998.

[97] G. Oriolo, G. Ulivi, and M. Vendittelli, "Real-Time Map Building and Navigation for Autonomous Robots in Unknown Environments," *IEEE Trans. on Systems, Man and Cybernetics, Part B*, vol. 28, pp. 316–333, 1998.

[98] L. Zadeh, "Fuzzy Sets as A Basis for A Theory of Possibility," *Fuzzy Sets Syst.*, vol. 1, pp. 3–28, 1978.

[99] G. Shafer, *A Mathematical Theory of Evidence.* Princeton University Press, 1976.

[100] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz, "A Comparison of Probabilistic, Possibilistic and Evidence Theoretic Fusion Schemes for Active Object Recognition," *Computing*, vol. 62, pp. 293–319, 1998.

[101] K. Gremban and K. Ikeuchi, "Planning Multiple Observations for Object Recognition," *Int. J. Comput. Vision*, vol. 12, pp. 137–172, 1994.

[102] H. Liu and X. Lin, "Model-Based Next View Planning by Using Rules-Automatic Feature Prediction and Detection," in *Proc. of IEEE Comput. Society Conf. on Computer Vision and Pattern Recognition*, pp. 773–776, 1994.

[103] T. Mukai and M. Ishikawa, "An Active Sensing Method Using Estimated Errors for Multisensor Fusion Systems," *IEEE Tran. on Industrial Electronics*, vol. 43, pp. 380–386, 1996.

[104] Y. Ye and J. Tsotsos, "Where to Look Next in 3D Object Search," in *Proc. of Int. Symp. on Computer Vision*, vol. 2, pp. 539–544, 1995.

[105] W. Burgard, D. Fox, and S. Thrun, "Active Mobile Robot Localization by Entropy Minimization," in *Proc. of the Second EUROMICRO workshop on Advanced Mobile Robots*, pp. 155–162, 1997.

[106] J. Denzler, R. Bess, J. Hornegger, H. Niemann, and D. Paulus, "Learning, tracking and recognition of 3D objects," in *Proc. of IEEE/RSJ/GI Int. Conf. on Intelligent Robots and Systems*, vol. 1, pp. 89–96, 1994.

[107] B. Krebs, B. Korn, and M. Burkhardt, "A Task Driven 3D Object Recognition System Using Bayesian Networks," in *the Sixth Int. Conf. on Computer Vision*, pp. 527–532, 1998.

[108] M. Werman, S. Banerjee, S. Roy, and Q. Maolin, "Robot Localization Using Uncalibrated Camera Invariants," in *Proc. of IEEE Comput. Society Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 353–359, 1999.

[109] F. Callari and F. Ferrie, "Autonomous Recognition: Driven by Ambiguity," in *Proc. of IEEE Comput. Society Conf. on Computer Vision and Pattern Recognition*, pp. 701–707, 1996.

[110] T. Arbel and F. Ferrie, "Viewpoint Selection by Navigation through Entropy Maps," in *Proc. of the Seventh Int. Conf. on Computer Vision*, vol. 1, pp. 248–254, 1999.

[111] M. Sipe and D. Casasent, "Global Feature Space Neural Network for Active Object Recognition," in *Int. Joint Conf. on Neural Networks*, vol. 5, pp. 3128–3133, 1999.

[112] M. Sipe and D. Casasent, "Feature Space Trajectory Methods for Active Computer Vision," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1634–1643, 2002.

[113] F. Deinzer, J. Denzler, and H. Niemann, "Viewpoint Selection - A Classifier Independent Learning Approach," in *Proc. of the Fourth IEEE Southwest Symposium Image Analysis and Interpretation*, pp. 209–213, 2000.

[114] A. Nehorai and M. Hawkes, "Performance Bounds for Estimating Vector Systems," *IEEE Trans. on Signal Processing*, vol. 48, pp. 1737–1749, 2000.

[115] M. Hernandez, B. Ristic, A. Farina, and L. Timmoneri, "A Comparison of Two Cramér-Rao Bounds for Nonlinear Filtering with $Pd < 1$," *IEEE Trans. on Signal Processing*, vol. 52, pp. 2361–2370, 2004.

[116] A. Farina, B. Ristic, and L. Timmoneri, "Cramer-Rao Bound for Nonlinear Filtering with $Pd < 1$ and Its Application to Target Tracking," *IEEE Trans. on Signal Processing*, vol. 50, pp. 1916–1924, 2002.

[117] P. Tichavsky, C. H. Muravchik, and A. Nehorai, "Posterior Cramér-Rao Bounds for Discrete-Time Nonlinear Filtering," *IEEE Trans. on Signal Processing*, vol. 46, pp. 1386–1396, 1998.

[118] M. Hernandez, T. Kirubarajan, and Y. Bar-Shalom, " Multisensor Resource Deployment Using Posterior Cramer-Rao Bounds," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 40, pp. 399–416, 2004.

[119] S. Paris and J.-P. L. Cadre, "Planification for Terrain-Aided Navigation," in *Proc. of the Fifth IEEE Int. Conf. on Information Fusion*, vol. 2, pp. 1007–1014, 2002.

[120] A. Papoulis and S. Pillai, *Probability, Random Variables and Stochastic Processes.* New York: McGraw-Hill, 2002.

[121] H. Murakami and B. V. Kumar, "Efficient Calculation of Primary Images From a Set of Images," *IEEE Tran. on Pattern Analysis and Machine Intelligence*, vol. 4, pp. 511–515, 1982.

[122] Y. Bar-Shalom, X. Li, and T. Kirubarajan, *Estimation, Tracking and Navigation: Theory, Algorithms and Software.* New York: John Wiley and Sons, 2001.

[123] B. R. Frieden, *Physics from Fisher Information, A Unification* . Cambridge, UK: Cambridge University Press, 1998.

[124] G. Lepage, "A New Algorithm for Adaptive Multidimensional Integration," *Journal of Computational Physics*, vol. 27, pp. 192–203, 1978.

[125] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C, The Art of Scientific Computing.* New York: Press Syndicate of the Cambridge University, 2002.

[126] J. Denzler and C. Brown, *Optimal Selection of Camera Parameters for State Estimation of Static Systems: An Information Theoretic Approach.* New York: Technical Report, Computer Science Department, The University of Rochester, 2000.