Active Recognition through Next View Planning: A Survey

Sumantra Dutta Roy [a] Santanu Chaudhury [b,*] Subhashis Banerjee [c]

[a] *Department of Electrical Engg., IIT Bombay, Powai, Mumbai - 400 076, INDIA*

[b] *Department of Electrical Engg., IIT Delhi, Hauz Khas, New Delhi - 110 016, INDIA*

[c] *Department of Comp. Sc. & Engg., IIT Delhi, Hauz Khas, New Delhi - 110 016, INDIA*

Abstract

3-D object recognition involves using image-computable features to identify 3-D object. A single view of a 3-D object may not contain sufficient features to recognize it unambiguously. One needs to plan different views around the given object in order to recognize it. Such a task involves an active sensor – one whose parameters (external and/or internal) can be changed in a purposive manner. In this paper, we review two important applications of an active sensor. We first survey important approaches to active 3-D object recognition. Next, we review existing approaches towards another important application of an active sensor namely, that of scene analysis and interpretation.

Key words: Active Vision, Computer Vision, Next View Planning, 3-D Object Recognition, Scene interpretation

1 Introduction

3-D object recognition is the process of identifying 3-D objects from their images by comparing image-based features, or image-computable representations with a stored representation of the object. (For detailed surveys of 3-D object recognition and related issues, see [1], [2]) Various factors affect the strategy used for recognition, such as the type of the sensor, the viewing transformations, the type of object, and the object representation scheme. Sensor output could be 3-D range images,

* Author for correspondence

Email addresses: sumantra@ee.iitb.ac.in (Sumantra Dutta Roy), santanuc@{ee, cse}.iitd.ac.in (Santanu Chaudhury), suban@cse.iitd.ac.in (Subhashis Banerjee).

or 2-D intensity images. 3-D range images can be obtained from the output of a light stripe range finder, for example. 2-D images may be obtained from various means such as CCD cameras, infra-red devices, X-ray images, or from other devices operating on different ranges of the electromagnetic spectrum. 3-D objects may be classified as rigid, articulated, or deformable. In this survey, we primarily concentrate on 2-D intensity images taken with cameras. This paper is restricted to the recognition of rigid 3-D objects and analysis of 3-D scenes.

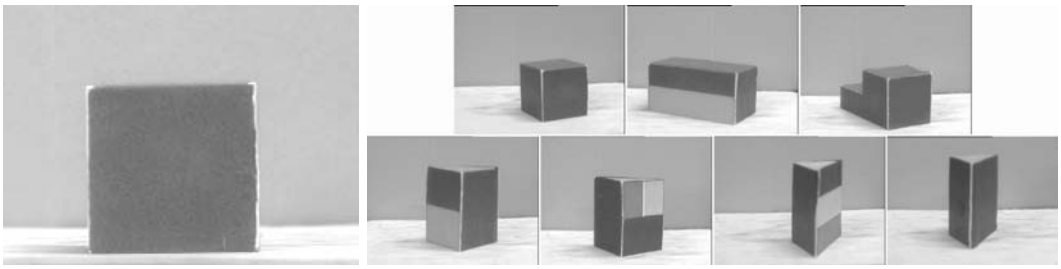
3-D object recognition from 2-D intensity images is a difficult task, primarily because of the inherent loss of information between a 3-D object and its 2-D image. The appearance of the object depends on factors such as the viewing geometry, illumination and viewpoint. The presence of noise in the feature detection process increases the difficulty of the recognition problem. The use of multiple views, instead of a single view, can make the 3-D object recognition problem more tractable.

1.1 The Need for Multiple Views

Most model-based 3-D object recognition systems consider the problem of recognizing objects from the image of a single view of an object ([1], [2], [3], [4]). Due to the inherent loss of information in the 3-D to 2-D imaging process, one needs an effective representation of properties (geometric, photometric, etc.) of objects from images which are invariant to the view point, and should be computable from image information. Invariants may be colour-based (*e.g.*, [5]), photometric (*e.g.*, [6]) or geometric (*e.g.*, [3]).

Burns, Weiss and Riseman prove a theorem in [7] that geometric invariants cannot be computed for a set of 3-D points in general position, from a single image. Invariants can only be computed for a constrained set of 3-D points. One can impose constraints on the nature of objects to compute invariants for recognition [8] – this severely restricts the applicability of the recognition system to only specific classes of objects *e.g.*, canal surfaces [9], [10], rotational symmetry [8], [11], [3], repeated structures (bilateral symmetry, translational repetition) [3], [12], [13]. While invariants may be important for recognizing some views of an object, they cannot characterize all its views – except in a few specific cases, as mentioned above. We often need to recognize 3-D objects which because of their inherent asymmetry, cannot be completely characterized by an invariant computed from a single view. For example, certain self-occluded features of an object can become visible if we change the viewpoint. In order to use multiple views for an object recognition task, one needs to maintain a relationship between different views of an object.

A single view may not contain sufficient features to recognize an object unambiguously. A further complication arises if two or more objects have a view in common with respect to a feature set. Such objects may be distinguished only through a se-



(a) (b)

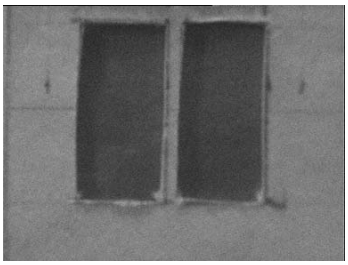
Fig. 1. (a) The given complete view of an object, and (b) the objects which this view could correspond to

quence of views. As a simple example, let us consider a set of 3-D objects with the number of horizontal and vertical lines, as features. Figure 1(a) shows a given view. On the basis of the chosen features, this view could correspond to any of the objects in Figure 1(b). In other words, with each of the objects of Figure 1(b), it is possible to obtain a view in which we would detect only two horizontal and two vertical lines. Hence, it is not possible to determine which object the given view corresponds to, given only the single view in Figure 1(a). In fact, two objects may have all views in common with respect to a given feature set, and may be distinguished only through a sequence of views. In [14], the authors cite a simple example of a sedan and a station wagon having indistinguishable front ends, but different side views.

A further complication arises when in an image, we do not have a complete view of an object. Figure 2(a) shows such an example. Such a view could have come from any of the three models, different views of which are shown in Figure 2(b), (c) and (d), respectively. Again, the identity of object cannot be uniquely determined from this one view. Further, even if the identity of the object were known, the same configuration of parts could occur at more than one place in the object. In that case, it is not possible to know the exact pose of the camera with respect to the object.

There may be another motivation for using multiple views in a recognition task. In recognizing 3-D objects from a single view, recognition systems often use complex feature sets ([2]). Complex features such as 3-D projective invariants have been proposed only for special cases so far (*e.g.*, [3]). In many cases, it may be possible to achieve the same, incurring less error and smaller processing cost using a simpler feature set and suitably planning multiple observations. A simple feature set is applicable for a larger class of objects than a model base-specific complex feature set.

Active vision involves using a sensor to take multiple views, in a purposive manner. We discuss this in detail, in the following section.



(a)

(b) (c) (d)

Fig. 2. (a) The given view of an object: only a portion of it is visible. This could have come from any of the models, different views of which are shown in (b), (c) and (d), respectively

1.2 Active Vision

An active sensor may be defined as follows:

Active Sensor An active sensor is one that can be purposively controlled. An **Active Vision** system has the ability to control the sensor parameters such as the orientation with respect to an object. Thus, vision-guided feedback may be used to position such a sensor. Such a system has other parameters that may be purposively varied, such as the focus, zoom, aperture and vergence (in two-camera system). Some specialized sensors have anthropomorphic properties, such as foveal attention mechanisms.

Some important papers proposing and elucidating the concepts of active vision and related paradigms, are the work of Aloimonos *et al.* [15]; Bajcsy *et al.* [16], [17]; Ballard [18]; Ballard and Brown [19], and Crowley [20]. A collection of representative papers in the area is [21]. Swain and Stricker [22] survey a wide gamut of vision tasks which may be performed with an active sensor. They mention that active vision broadly encompasses attention, selective sensing in space, resolution and time. This may be achieved by modifying physical camera parameters, or the way the data from the camera is processed.

The output of a camera usually contains a huge amount of data. Hence, an attention mechanism may use an active sensor for signal selection in space, velocity and distance (*e.g.*, foveal processing, tracking an object, and focusing on regions of interest). Gaze control is a possible application – active manipulation of the imaging system in order to acquire images that are directly suited to the tasks being performed. Gaze control could be used for low level vision (*e.g.*, Murray *et al.* [23], Crowley *et al.* [24]), as well as for high level vision (*e.g.*, Rimey and Brown [25]). Thus, an active mechanism could be used to overcome a limited field of view of a camera. A related application is next view planning. Cost and complexity considerations often require a system to be focused on restricted regions of a scene. Further, the current view available to a system may not even contain sufficient information for the vision task. Thus, deciding on where to look next may be task driven, feature driven, or context driven. Thus, a sequence of such sensing operations may be required. Sequential processing has the additional advantage of efficiency through directed analysis – results of each step guide subsequent steps. Object recognition and scene analysis are two example of such a vision task. Another example of an active vision task is eye-hand coordination [26].

Tarabanis, Allen and Tsai [27] survey the field of sensor planning for vision tasks. They define the problem as follows: Given information about the environment (*e.g.*, the object and the sensors), and information about the vision task (*e.g.*, detection of certain object features, object recognition, scene analysis), the task at hand is to develop strategies to automatically determine parameter values in order to achieve the task, to a required degree of satisfaction. They classify problems into three classes:

(1) object feature detection
(2) object recognition and localization, and
(3) scene reconstruction.

We discuss these issues in the following sections.

1.2.1 Object Feature Detection

Object feature detection seeks to automatically determine vision sensor parameter values for which particular features satisfy particular constraints when imaged.

These features belong to a known object in a known pose [27]. In addition to the general survey on sensor planning, the authors lay specific emphasis on systems for object feature detection systems. (A separate paper [28] presents the authors' own MVP system in detail.) A related topic is planning for complete sensor coverage of 3-D objects. A recent work in the area is that of Roberts and Marshall [29], who present a viewpoint selection scheme for complete surface coverage of 3-D objects. Some important earlier work in the area include those of Cowan and Kovesi [30], Tarbox and Gottschlich [31] and Mason and Grun [32].

1.2.2 Object Recognition and Localization, and Scene Reconstruction

Given an active sensor and a set of feature detectors, the fundamental problems involved in a multiple view-based recognition system are

- the design of a suitable modeling and representation scheme, and
- an identification mechanism which can exploit properties of the sensing process and the modeling scheme.

Based on the representation scheme and the scope or nature of the recognition strategy, we classify different multi-view recognition systems into two categories:

(1) Object recognition systems, and
(2) Systems for scene analysis

In the first class of systems, we consider systems whose aim is to primarily recognize a given object and its pose. Such systems typically assume that the entire object is visible in a given view. In the second class of scene analysis systems, we consider systems whose aim is to explore and and analyze a given scene. Such a scene may contain one or more objects, known or unknown. In such cases, the entire scene to be analyzed may not be visible in one view – the sensor may 'see' only a part of it at a time. While recognition may not be a primary aim of such systems, they may involve recognition of some components of a scene. We describe these two categories in detail in Sections 2 and 4, respectively.

2 Active Object Recognition Systems

An active object recognition system uses multiple views of 3-D objects for recognition in a purposive fashion. Based upon specialized representation scheme linking multiple views of 3-D objects, different recognition schemes have been formulated for active object recognition.

Object representation schemes used for model-based 3-D object recognition systems include ([33]): wire-frame representations, constructive solid geometry-based schemes (CSG), spatial-occupancy representations (*e.g.*, voxels, octrees), surface boundary representations, generalized cone or sweep representation, skeleton representations, and aspect graphs. Appearance-based approaches (*e.g.*, [34]) to object recognition use appearance rather than shape, for matching. However, only a few of the above approaches have been used in multi-view object recognition systems. While wire-frame models have an inherent ambiguity in interpretation, feature extraction is difficult in volume or surface-based approaches. Skeleton representations and generalized cones are applicable for recognition of only a specific class of objects. Representation schemes can also be characterized on the basis of whether they represent the object as a whole, or model it in terms of its *parts*.

2.1.1 View Based Representation

Most active object recognition systems consider either of the following three representation schemes, or their variants:

- Appearance-based parametric eigenspaces
- Multidimensional Receptive Field Histograms
- Aspect graphs

These three are *view-based* – they encode information about different 2-D views of a 3-D object. Breuel [35] describes simulations and experiments on real images to suggest view based recognition as a robust and simple alternative to 3-D shape-based recognition methods. In what follows, we describe the above three view-based representation schemes. We briefly point out their use in active recognition systems – Section 2.3 describes them in detail.

Appearance-Based Parametric Eigenspaces

Murase and Nayar [34] propose the idea of appearance-based methods using parametric eigenspaces. A basic observation is that the shape and reflectance are intrinsic properties, which are constant for a rigid object. The authors propose a scheme to automatically learn 3-D objects from their appearance in 2-D images. An important advantage of this method is the ability to handle the combined effects of shape, pose, reflection properties and illumination. Furthermore, it is possible to learn appearance-based object representations off-line. For systems using such a representation, the recognition problem becomes one of appearance matching, rather than shape matching.

Appearance-based methods require a large number of images of the object – with different poses, and illumination conditions. The images of the objects are normalized with respect to size and illumination conditions. Both in the pose and illumination space, consecutive images are correlated to a large degree. Each normalized image is written as a column vector in raster scan order. Next, the normalized images are stacked together , and the covariance matrix is calculated. The first k eigenvectors are used to represent the stacked matrix of images [34]. In the recognition phase, the image vector is projected to the eigenspace. The object which has a minimum distance between the projected image vector and its manifold, is considered to be present. The work of Borotschnig *et al.* [36] is an example of the use of parametric appearance-based information for active object recognition (Section 2.3).

Multidimensional Receptive Field Histograms

Multidimensional Receptive Field Histograms [37] are based on the idea that local structure is an important component of the appearance of an object. The local structure can be characterized by a vector of local features measured by local operators such as Gaussian derivatives or Gabor filters. The authors acquire this information from sample training images. This technique represents appearances of objects by the joint statistics of such local neighbourhood operators. Multidimensional receptive field histograms approximate the probability density function for local appearance. Their selection of features is not restricted to a particular family of objects, nor rely on a particular set of features. The features should be invariant (with respect to certain transformations), equivariant (as a function of a certain transinformation), and robust (change slowly in the presence of certain transformations). In [38] Schiele and Crowley present an active recognition system based on multidimensional receptive field histograms (Section 2.3).

Aspect Graphs

Aspect graphs are a popular representation tool for 3-D object recognition systems. Koenderink and van Doorn [39] define **aspects** as topologically equivalent classes of object appearances. Chakravarty and Freeman [40] adopt a similar approach in their definition of the 'Characteristic Views', and their uses in object recognition. Since sensors may be of different types (geometric, photometric, etc.), Ikeuchi and co-workers generalize this definition – Object appearances may be grouped into equivalence classes with respect to a feature set. These equivalence classes are aspects [41]. Thus, an aspect is a collection of contiguous sites in viewpoint space which correspond to the same set of features. We define an aspect graph as follows:

Aspect Graph An aspect graph consists of nodes which correspond to aspects. Links between nodes represent transitions from one aspect to another. A link is

often referred to as an *accidental view*, or a *visual event* [42].

Aspect graph-based and related representations include [40], [43], [44], [45], [46], [41], [47], [48], [49], [50], [51], [52], [53], [54]. Many active object recongition schemes are based on aspect graphs [55], [56], [14], [57], [58] (Section 2.3 describes these in detail). The aspect graph-based approach is more general than the other two approaches in that appearance-based information may be used to construct an aspect graph.

2.1.2 Part-Based Representations

Some object recognition systems consider the representation of an object in terms of its *parts*. Existing part-based recognition systems typically consider the object to be wholly composed of identifiable parts. Here, we review two part-based approaches. The first is based on volumetric primitives, and the second on appearance-based parts. Existing part-based recognition systems usually use information from only a single view. The works of Dickinson *et al.* [56], and our work [59] are examples of active recognition systems using a part-based representation. Here, we look at two part-based representations namely, geons and appearance-based parts:

Geons

Biederman's *Recognition by Components* theory [60] proposes the concept of volumetric primitives, called geons (short for 'geometric ions'). The Cartesian product of contractive shape properties give rise to this set of volumetric primitives. Bergevin and Levine [61], [62] propose methods of automatically extracting geons from relatively perfect 2-D line drawings. Single view-based recognition systems such as [62], [44] and [45] use geons as representation tools. In [62], Bergevin and Levine propose a system for generic object recognition from a single line drawing of an object, by parts. Dickinson and co-workers [44], [45] use an augmented aspect hierarchy using geons as the basic volumetric primitives. They use this augmented aspect hierarchy for active 3-D object recognition in [56].

Appearance-based parts

Another approach to part-based representation is that of Huang, Camps and Kanungo [63], [64]. The authors define appearance-based parts as "polynomial surfaces approximating closed, non-overlapping image regions that optimally partition the image in a minimum description length (MDL) sense." Their single view-based recognition systems consider the advantages of appearance-based representations. Additionally, the idea of recognizing parts and not the whole object gives the system robustness to occlusion and segmentation variations.

An important assumption in the above two schemes is that the object is partitioned into a set of recognizable parts. The part-based recognition system [59] considers a more general case. The paper also consider an object to be composed of a set of identifiable parts. However, the authors do not assume the entire object to be partitioned into a set of identifiable parts – there may be portions of the object which do not have any detectable features, with respect to the set of feature detectors being used. Section 2.3 briefly outlines this scheme.

2.2 Methods for Representing Uncertainty

Uncertainty in an object recognition task can be with respect to interpretation of a view of the object. It could have come from more than one object, and more than one part of the same object. Factors such as noise and non-adaptive thresholds may corrupt the output of a feature detector. In such a case, a feature detector may erroneously report a different feature from what is 'actually' present.

Common methods for representing uncertainty are probability theory, the Dempster-Shafer theory [65], [66], and fuzzy logic [67]. A representation scheme based on probability theory is a Bayes Net [68]. (Bayes nets, and their variants are also known as Belief networks, Bayesian networks, and probabilistic networks.) However, a Bayes net is a far more general AI-based representation scheme (as against the above schemes specifically used for modeling 3-D objects). A Bayes net is a graph which represents the joint probability distribution of a set of variables. Nodes in the graph represent variables, and directed links represent conditional probabilities. The Bayes rule is used for updating the probabilities of nodes having a particular label, given that successor nodes have particular labels. Dickinson *et al.* [56] use a variant of a Bayes net for their recognition system (Rimey and Brown [25]. use Bayes nets for scene analysis), while Hutchinson and Kak [55] use the Dempster-Shafer theory to represent uncertainty. Some scene analysis systems use fuzzy logic (*e.g.*, [69]).

In the following section, we discuss different characteristics about uncertainty representation schemes, in conjunction with the recognition strategies.

2.3 Recognition Strategies

We now present recognition strategies for some important active 3-D object recognition schemes. We classify these on the basis of the next view planning strategy:

(1) Systems which take the next view to minimize an ambiguity function, and
(2) Systems incorporating explicit planning algorithms

We discuss different schemes as follows. All except the last three (described in detail, below) belong to the first category. Our own work on active 3-D object recognition uses a planning scheme in order to take the next view [57], [58], [59] (brief description in Sections 2.3 and 2.3).

Hutchinson and Kak

In their work on planning sensing strategies in a robot work cell with multi-sensor capabilities, Hutchinson and Kak [55] use an aspect graph to represent information about the objects in their model base. They present a system for dynamically planning sensing strategies, based on the current best estimate of the world. They automatically propose a sensing operation, and then determine the maximum ambiguity which would remain if the operation were applied. The system then selects the operation which minimizes the remaining ambiguity. They use the Dempster-Shafer theory to combine evidence and analyze proposed operations.

Liu and Tsai

Liu and Tsai [70] describe a multiple view-based 3-D object recognition system. Their setup has two cameras and a turntable. They use silhouettes as features. The system first reduces ambiguity by taking images from above the turntable to normalize the shape of the top view, position the object centroid, and align the principal axis of the object. The system then takes a side view, and analyzes its features. Then the object is rotated by 45^o. This system repeats the above process, till the object is recognized.

Callari and Ferrie

Callari and Ferrie [71] base their active object recognition system on mode-based shape, pose and position reconstructions from range data. They estimate Bayesian probabilities with neural networks. Their system takes the next view based on the move which minimizes the expected ambiguity in terms of Shannon entropy.

Schiele and Crowley

Schiele and Crowley [38] develop an analogy between object recognition and the transmission of information through a channel based on the statistical representation of 2-D object appearances. They use multidimensional receptive field histograms. Transinformation enables the determination of the most discriminative

Verification Steps	recognition	number of errors
0	98.22%	64
1	99.08%	33
2	99.97%	1

Table 1
Some experimental results for the active recognition system of Schiele and Crowley [38] on the Columbia image database of 100 3-D objects (Table 1, page 254 in the paper).

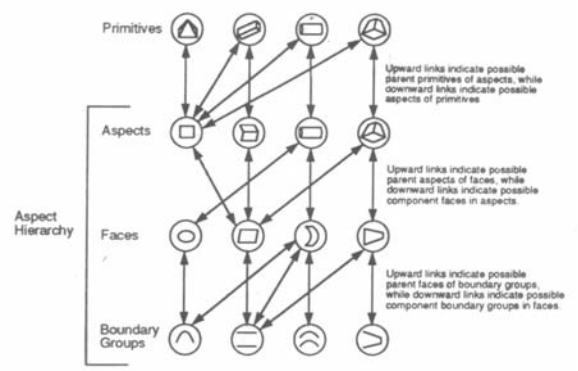


Fig. 3. The augmented aspect hierarchy: This is Figure 2 in [56], page 243.

viewpoints. The proposed strategy moves the camera to the most discriminant viewpoint of the hypothesized object. The authors show results of using their recognition strategy on the Columbia database of 100 3-D objects. Table 1 summarizes some of the results.

Dickinson et al.

Dickinson and co-workers [56] present an active object recognition scheme which integrates attention and viewpoint control. Their representation scheme is similar to that of Biederman [60]. The system uses an augmented aspect hierarchy as their data structure (Figure 3). Aspects are used to model the (typically small) set of volumetric part-classes from which each object in the database is constructed. The augmented aspect hierarchy considers relations between boundary groups (representing all subsets of contours bounding the faces), the faces, the aspects, and finally, the volumetric primitives themselves. The entities at each level are linked with one another. Each link is associated with a conditional probability.

Dickinson *et al.* present a case for using regions. They use conditional probabilities captured in the augmented aspect hierarchy to define a measure of average inferencing uncertainty. On the basis of this, they conclude that the value of this parameter for faces is less than that for boundary groups. It is pointed out that the advantage would be realizable if the cost of extracting the features corresponding to the two are comparable. Their attention mechanism exploits the augmented aspect hierar-

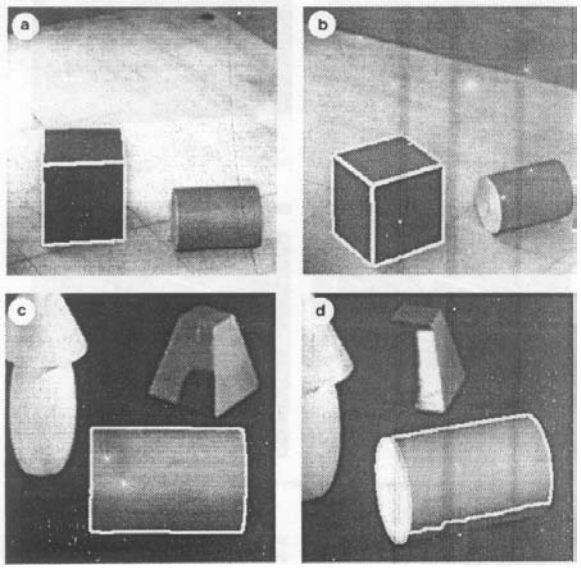


Fig. 4. Moving the sensor to disambiguate volumes 1 (block) and 5 (cylinder) in the system of Dickinson *et al.*: This is Figure 16 in [56], page 257

chy to map target objects down to target faces. Target faces are in turn, compared to image faces. In selecting which recovered face to focus attention on, they use a decision theoretic approach using a Bayesian framework. They use a structure known as the aspect prediction graph to drive the sensor to a new position from which an object's part can be disambiguated. Figure 4 shows results of moving the camera to disambiguate between two objects.

Borotschnig et al.

Borotschnig *et al.* [36] present an active 3-D object recognition system that uses appearance-based information. They extend the idea of the off-line system of Murase and Nayar [34] to an on-line case. They use a parametric eigenspace, and augment it with probability distributions – to capture possible variations in the input images due to errors. Their system chooses as the next view a move, which minimizes the average entropy.

Gremban and Ikeuchi

Gremban and Ikeuchi [14] present a scheme for planning multiple views in an object recognition task. They use Aspect-Resolution Trees built on the basis of aspect diagrams for planning multiple observations for object recognition. The authors show results for a vision-based sensor, and a haptic sensor, and give examples of recognition of objects based on sensors to detect specularities and their properties. These specularities determine the aspects of the object. They consider this to be a challenging domain, since a single image of a specular object yields very little

Sampling Interval	1^o	2^o	4^o	6^o	8^o	10^o
Total tests	100	100	100	100	100	100
Correct	99	97	69	46	26	34
Incorrect	1	3	11	24	12	9
Unresolvable	0	0	20	30	62	57

Table 2
Aspect resolution results for the system of Gremban and Ikeuchi [14]: Table 2, page 66 in the paper

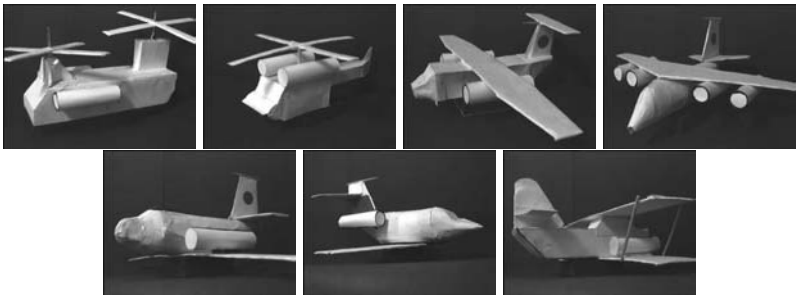


Fig. 5. Aircraft models: one of the sets of models used for experimentation with our first system.

information about the overall object shape (specularity being a local phenomenon). Further, many different object poses can yield the same pattern of specularities. The authors show recognition results with three objects. Table 2 shows some recognition (aspect resolution) results for an object - a stylized jet.

Aspect Graph-based Modeling and Recognition using Noisy Sensors

We propose a new on-line recognition scheme based on next view planning for the identification of an isolated 3-D object using a set of noisy feature detectors. We use our aspect graph construction scheme [54] to construct an aspect graph, given noisy feature data. The scheme uses a probabilistic reasoning framework for recognition and planning. Our hierarchical knowledge representation scheme encodes feature-based information about objects as well as the uncertainty in the recognition process. This is used both in the probability calculations as well as in planning the next view. The planning process is reactive – the system uses both the past history as well as the current observation to plan a move which best disambiguates between competing hypotheses about the identity of the object. Results clearly demonstrate the effectiveness of our strategy for a reasonably complex experimental set. Figure 5 shows a data set with fairly complex shapes, with a large degree of interpretation ambiguity corresponding to a view. We use very simple features: the number of horizontal and vertical lines in an image of the object, and

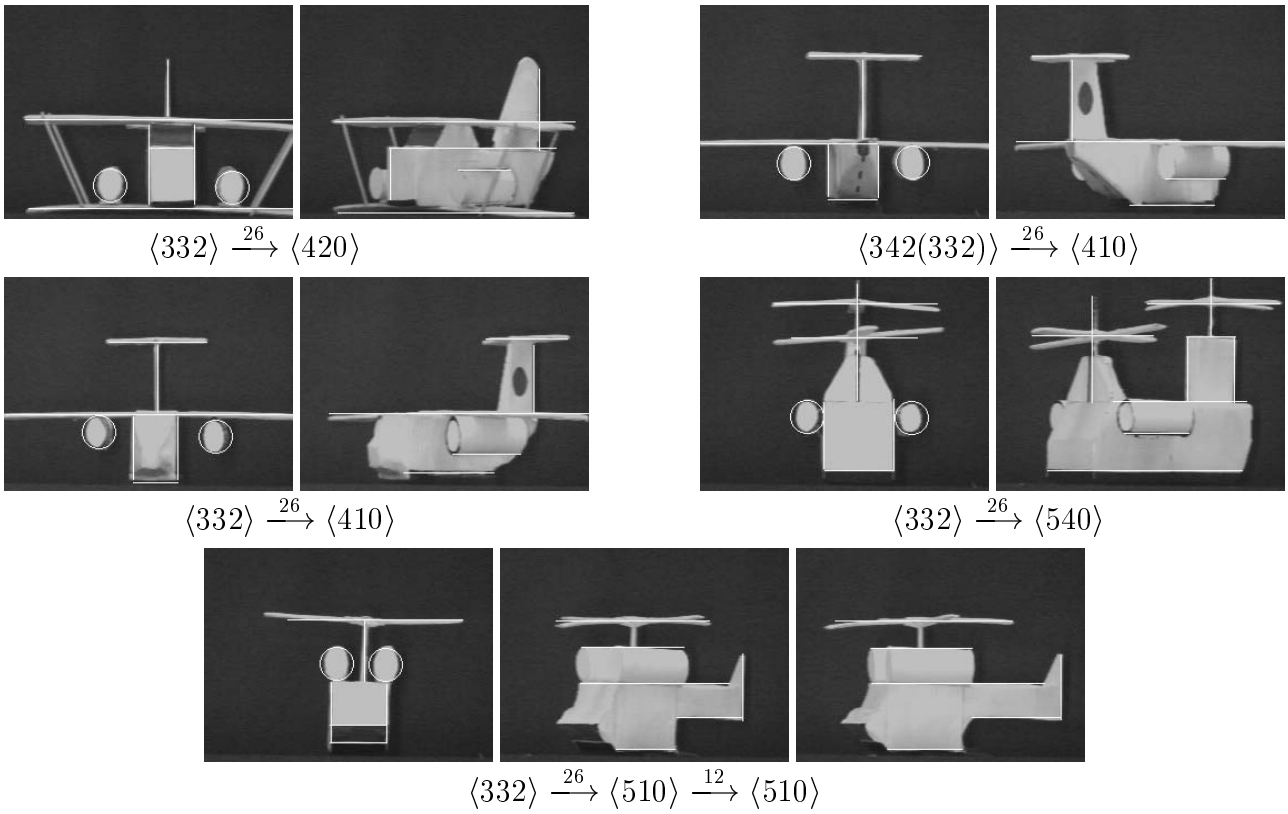


Fig. 6. Some experiments with our first system on a set of aircraft models, with the same initial view with respect to the feature set used. The numbers above the arrows denote the number of turntable steps. (The figure in parenthesis shows an example of recovery from feature detection errors)

the number of circles. Figure 6 shows examples of experiments with objects from the aircraft model base. The initial view in each of these examples has the same features: 3 horizontal and vertical lines each, and 2 circles. In Figure 6(b), the shadow of the left wing on the fuselage of the aircraft, the feature detector detects 4 vertical lines instead of 3, the correct number. Our error modeling and correction scheme enables the system to recover from this feature detection error. Papers [57] and [58] describe different stages of the work, and its different aspects.

Recognizing Large 3-D Objects through Next View Planning using Inner Camera Invariants

Our second system uses a new on-line scheme for the recognition and pose estimation of a *large* isolated 3-D object, which may not entirely fit in a camera's field of view. We consider an uncalibrated projective camera, and consider the case when the internal parameters of the camera may be varied either unintentionally, or on purpose. We use a new class of invariants for complete 3-D Euclidean pose estimation – *Inner Camera Invariants*, image-computable functions which are invariant to the internal parameters of a camera [72]. We propose a part-based knowledge repre-

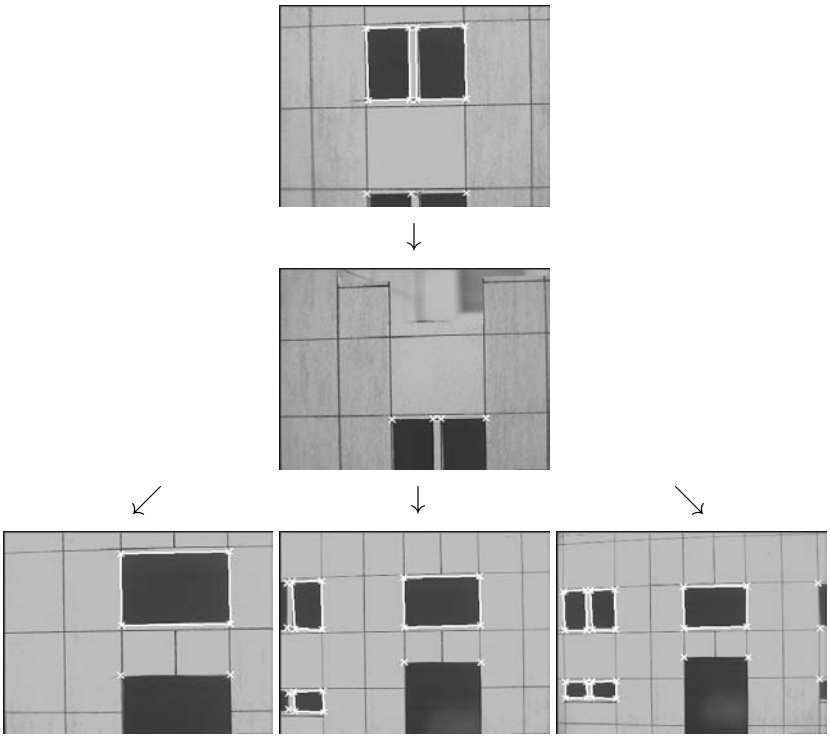


Fig. 7. For the same first two views, we progressively zoom-out the camera in three stages. (a), (b) and (c) depict the three views which the camera sees, for the third view. This does not affect the recognition system in any way – the system identifies the object and the camera pose accurately in each case.

sentation scheme. We consider a very general definition of the word 'part' – A view of an object contains 2-D or 3-D **parts** (which are detectable using 2-D or 3-D projective invariants, for example), and other 'blank' or 'featureless' regions (which does not have features detectable by the feature detectors). Thus, according to our formulation, an object is composed of parts, but is not partitioned into a collection of parts. The scheme uses a probabilistic reasoning framework for recognition and next view planning. We show results of successful recognition and pose estimation even in cases of a high degree of interpretation ambiguity associated with the initial view. We have experimented with a set of architectural models. Figure 2(a) shows such an example. Such a view could have come from any of the three models, different views of which are shown in Figure 2(b), (c) and (d), respectively. Figure 7 shows an example of correct object recognition and pose estimation, in a case when the internal parameters of the camera change – a zoom-out operation, in this case. (Details in [59]).

3 A Comparative Analysis of Active Object Recognition Systems

Active recognition systems have been proposed which can work with different assumptions about the nature of the sensors and the environment, the degrees of freedom between the object and the sensor, and the object models themselves. We discuss the following issues with respect to different active 3-D object recognition systems:

(1) *Features used for modeling and view recognition*
While many approaches such as those of Hutchinson and Kak [55] and Liu and Tsai [70] use geometric features, the scheme of Gremban and Ikeuchi [14] is independent of the features used. The latter present results with geometric and photometric information. Our work on isolated 3-D object recognition through next view planning [57], [58] is also independent of the specific features used for modeling and view recognition. Appearance-based methods such as that of Borotschnig *et al.* [36] use pixel information from an entire image. Dickinson *et al.* [56] use volumetric primitives, which are associated with a high feature extraction cost. The same is true for the super-ellipsoids of Callari and Ferrie [71].

(2) *The system setup and viewing geometry*
Existing systems such as those of Hutchinson and Kak [55], Liu and Tsai [70], Callari and Ferrie [71], Dickinson *et al.* [56], Gremban and Ikeuchi [14], and Borotschnig *et al.* [36] assume that the object completely fits into the camera's field of view. Borotschnig *et al.* [36] assume a single degree of freedom (hereafter, DOF) between the object and the sensor. While Gremban and Ikeuchi [14] have experimented with such a case, they propose extensions to higher degrees of freedom. Most multiple view-based approaches using geometric features, implicitly or otherwise, assume the camera model to be orthographic. While our work on aspect graph-based modeling and recognition [57], [58] assumes a 1-DOF case and an orthographic camera, the work on part-based recognition of large 3-D objects considers the most general 6-DOF case, and a commonly used projective camera model. The latter does not assume that the object fits into the camera's field of view. Additionally, it is independent of the internal parameters of the camera.

(3) *Efficient representation of knowledge about object models*
The knowledge representation scheme should support an efficient mechanism to generate hypotheses on the basis of the evidence received. It should also play a role in optimally planning the next view. Dickinson *et al.* [56] use a hierarchical representation scheme based on volumetric primitives. Due to the non-hierarchical nature of Hutchinson and Kak's system [55], many redundant hypotheses are proposed, which have to be later removed through consistency checks. In our work on aspect graph-based modeling and recognition, the hierarchical knowledge representation scheme and probabilistic hypothesis generation mechanism itself refines feature evidence through different lev-

els – leading to simpler evidence propagation and less computational cost [57], [58]. Borotschnig *et al.* [36] use a parametric eigenspace-based representation, which is associated with a high storage and processing cost.

(4) *Speed and efficiency of algorithms for both hypothesis generation and next view planning*
Hypothesis generation should be fast, and incur minimal error. The next view planning strategy acts on basis of these hypotheses. In Hutchinson and Kak's system [55], the polynomial-time formulation overcomes the exponential time complexity associated with assigning beliefs to all possible hypotheses. However, their system still has the overhead of intersection computation in creating common frames of discernment. Consistency checks have to be used to remove the many redundant hypotheses produced earlier. Though Dickinson *et al.* [56] use Bayes nets for hypothesis generation, their system incurs the overhead of tracking the region of interest through successive frames.

(5) *Nature of the next view planning strategy*
The planning scheme should ensure adequate discriminatory ability between views common to more than one object in the model base. The cost incurred in this process should also be minimal. The system should, preferably be on-line and reactive – the past and present inputs should guide the planning mechanism at each stage. While schemes such as those of Borotschnig *et al.* [36] and our systems [57], [58], [59] are on-line, that of Gremban and Ikeuchi [14] is not. An off-line approach may not always be feasible, due to the combinatorial nature of the problem. An on-line scheme has the additional capability to react to unplanned situations, such as errors.

(6) *Uncertainty handling capability of the hypothesis generation mechanism*
Approaches such as those of Gremban and Ikeuchi [14], and Liu and Tsai [70] are essentially deterministic. An uncertainty-handling mechanism makes the system more robust and resistant to errors compared to a deterministic one. Dickinson *et al.* [56], Borotschnig *et al.* [36] and our systems [57], [58], [59] use Bayesian methods to handle uncertainty, while Hutchinson and Kak [55] use the Dempster-Shafer theory. In the work of Callari and Ferrie [71], the ambiguity in super ellipsoid-modeled objects is a function of the parameters estimated, on the basis of which the next move is determined. Schiele and Crowley [38] use a transinformation-based mechanism to propose the next move.

4 Active Scene Analysis Systems

The aims and domains of scene analysis systems are extremely diverse – even though active sensing and recognition usually form a common thread in each of them. Given their diverse natures, systems for scene analysis generally use specialized schemes for knowledge representation. They use these in conjunction with

the recognition and analysis strategies. In this section, we review some important classes of scene analysis systems, and their information representation and control schemes.

4.1 Next View Planning for Data Acquisition: Range Images

Maver and Bajcsy [73] present a strategy for next view planning which exploits occlusion information. The system exploits characteristics of the sensing process, to acquire yet-unknown 3-D information about the scene of interest. A related approach is that of Massios and Fisher [74]. The authors also use range images, and propose a quality criterion. This quality criterion aims at obtaining views that improve the overall range data quality of the imaged surfaces. Another recent approach is that of García, Velázquez and Sappa [75]. They present a two stage algorithm for determining the next view, using range images. The first stage applies a voting scheme that considers occlusion edges. Most of the surfaces of the scene are recovered this way. The second stage fills up the remaining holes through a scheme based on visibility analysis.

A related intensity image-based approach is one of recovering the surface shape with an active sensor. We discuss this in the following section.

4.2 Active Recovery of Surface Shape using Intensity Images

Kutulakos and Dyer [76] present an approach for recovering surface shape from an occluding contour of an object, using an active sensor. They use the fact that if the viewing direction is along a principal direction for a surface point whose projection is on the contour, it is possible to recover the surface shape (curvature).

4.3 Scene Geometry Interpretation and Exploration

Whaite and Ferrie [77] present a system for the interpretation of scene geometry in the form of parametric volumetric models. They describe ambiguity as a local probabilistic property of the misfit error surface in the parameter space of super-ellipsoid models. They propose a technique that uses this information to plan for the next view – which minimizes the ambiguity of subsequent interpretation. Marchand and Chaumette [78] present an autonomous active vision system for 3-D reconstruction of static scenes. They do not assume any prior knowledge of the number, localization, and the dimension of the different objects in the given scene. A controlled structure-from-motion method is used for reconstruction. This allows an optimal estimation of parameters of geometrical primitives. They present two

algorithms to ensure exploration of the scene. The first is an incremental reconstruction algorithm based on the use of a prediction/verification scheme involving decision theory and Bayes nets. The second algorithm is based on the computation of new viewpoints for the complete reconstruction of the 3-D scene.

4.4 Systems for 'Finding Waldo': Incorporating Colour and Other Cues

Grimson and co-workers [79] present an attentive active visual system which integrates visual cues to fixate candidate regions in which to recognize a target object. The system combines colour and stereo cues to perform figure/ground separation.

The following section describes an important paradigm in visual search namely, using intermediate objects.

4.5 Using Intermediate Objects to Enhance Visual Search

Wixson and Ballard [80] describe an active vision system that use intermediate objects to improve the efficiency of visual search. They show examples of trying to search for an object using an active camera, whose internal and external parameters can be varied, and which is also capable of foveal processing. They propose *indirect searches* to be more efficient as compared to direct searches for an object, in two cases. The first is when intermediate objects can be recognized at low resolutions and hence found with little extra overhead. The second is when they significantly restrict the area that must be searched for the target. Indirect searches use spatial relationships between objects to repeatedly look for intermediate objects, and look for the target object in the restricted region specified by these relationships. The authors present a mathematical model of search efficiency that identifies the factors affecting efficiency and can be used to predict their effects. They report that in typical situations, indirect search provides up to about an 8-fold increase in efficiency.

4.6 Selective Attention for Scene Analysis

Rimey and Brown [25] suggest the use of Bayes Nets for scene analysis through selective attention. They mention that the efficiency of a selective vision system comes from processing the scene only where necessary, to the level of detail necessary, and with only the necessary operators. Their system *TEA-1* uses not only the prior knowledge of a domain's abstract and geometrical structure, but is also reactive – it also uses information from a scene instance gathered during analysis. The knowledge representation is through 4 kinds of Bayes Nets, (the PART-OF net,

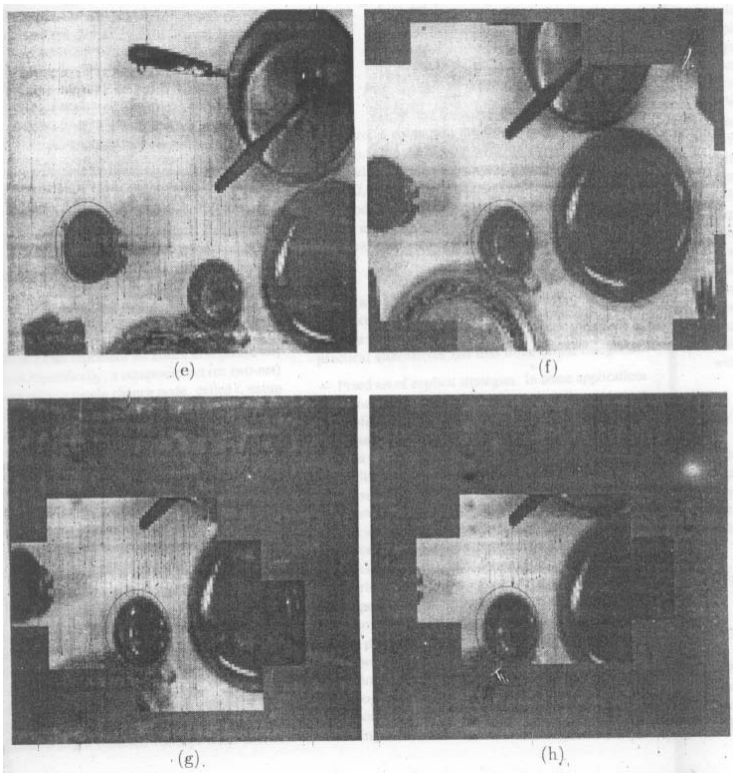


Fig. 8. An example of Rimey and Brown's system [25] trying to locate a cup - the change in its expected area as the system gathers evidence. This is Figure 9(e)–(h) in the paper, on page 187.

the expected area net, the IS-A tree, and the task net) which are used to store different kinds of domain knowledge. TEA-1 uses benefit-cost analysis for the control of visual and non-visual actions. The authors show the application of TEA-1 in analyzing dinner table scenes. Figure 8 shows an example of the system trying to locate a cup in a dinner table scene. Jensen, Christensen and Nielsen [81] adopt a similar approach. The conditional probabilities for their Bayesian network is obtained by subjective assessment. They show results on a network for discrimination between a British and a Continental breakfast table scene.

4.7 Dynamic Relevance in a Vision-Based Focus of Attention Strategy

Baluja and Pomerleau [82] use the concept of Dynamic relevance in their vision-based focus of attention strategy. The system ascertains the relevance of inputs by exploiting temporal coherence. In their system, relevance is a time-varying function of the previous and current inputs. It dynamically allocates relevance to inputs by using expectations of their future values. The expectation of what features will be there in the next frame decides which portion of the next visual scene will be focused on. The system uses a neural network with an input layer, a hidden layer and two sets of units in the output layer: one for the output, and one for the recon-

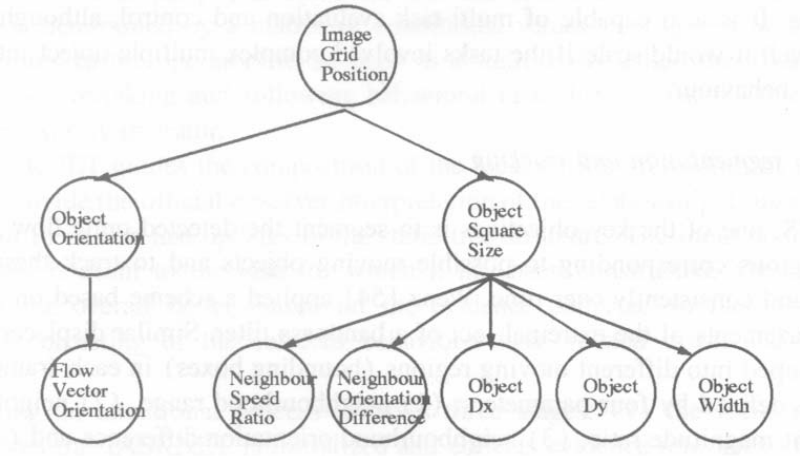


Fig. 9. A task-specific belief network in the system of Buxton and Gong [83] (Figure 6 on page 448). This belief network captures dependent relationships between the scene layout and relevant measures in motion segmentation and tracking, for their traffic surveillance application.

structed inputs. The weights between the input layer and the hidden layer, and those between the hidden layer and the outputs are trained to reduce the task error alone. The weights between the hidden layer and the reconstructed inputs are trained to reduce prediction error only. The architecture further has a feedback between the reconstructed 'next' inputs, and the input layer. The input layer actually uses the concept of a saliency map to make the system use filtered inputs. Thus, the information that is not relevant to the task will not be encoded in the hidden layer. The authors demonstrate the application of their ideas in various environments – vision-based autonomous control of a land vehicle, vision-based hand tracking in cluttered scenes, and the detection of faults in the plasma-etch step of semiconductor wafers.

4.8 Visual Surveillance

Buxton and Gong [83] present a visual surveillance system for tracking moving objects and interpreting their patterns of behaviour. They use conceptual knowledge of both the scene and the visual task to provide constraints to the problem. The control of the system is through dynamic attention and selective processing. The authors use belief networks to model dynamic dependencies between parameters involved in visual interpretation. Figure 9 shows an example of a an application-specific belief network used in their system. They present experimental results on a traffic surveillance application, using a fixed pre-calibrated camera model and pre-computed ground plane geometry. To recognize different objects in the scene, they use volumetric models. The system tracks objects across frames.

Nashashibi and co-workers [84] describe a system for indoor scene terrain modeling using multiple range images. This relies on two grid-based representations: the local elevation map, and the local navigation map. The authors describe their interpolation method to build their grid-based representation of the terrain – the local elevation map. Elevation data are used to build a symbolic grid representation call the local navigation map. Here, each terrain patch is assigned to a pre-defined class of terrain. They do not assume any *a priori* world model or landmarks to be available. Lebègue and Aggarwal [85], [86] propose a scheme for the extraction an interpretation of of semantically significant line segments for a mobile robot. The detection and interpretation processes provide a 3-D orientation hypothesis for each 2-D segment. This is used to estimate the robot's pose, and delimit the free space visible in the image. A motion stereo algorithm uses the orientation data to fully estimate the 3-D Euclidean structure of the scene.

Faugeras, Ayache and Faverjon [87] also use visual cues for map-building. This paper proposes a method to build visual maps by combining noisy stereo measurements. The authors propose the idea of a Realistic Uncertain Description of the Environment (RUDE) which incorporates local information – it is attached to a specific reference frame, and incorporates both geometrical information, as well as the related uncertainty information. They relate this to pixel uncertainty, and show how the RUDE corresponding to different frames can be used to relate them by a rigid displacement, and a measure of its uncertainty. Finally, they use the relations between frames to update the associated RUDE and decrease the uncertainty. In a more recent work, Faugeras [88] describes deterministic computational geometry-based methods for map building. Tirumalai, Schunck and Jain [89] address the problem of building an environmental map utilizing sensory depth information from multiple viewpoints. They represent the environment in the form of a finite-resolution 3-D grid of voxels. The system uses the Dempster-Shafer theory for multi-sensory depth information assimilation.

Asada [90] extends the work of Elfes [91] (whose system uses sonar data) and proposes a method for building a 3-D world model for sensory data from from outdoor scenes. His system allows for different sources of input data, such as range and video data. Figure 10 shows the architecture of the system. First, a range image ('physical sensor map') is transformed to a height map ('virtual sensor map') relative to a mobile robot. the height map is segmented into unexplored, occluded, traversable and obstacle regions from the height information. The system classifies obstacle regions into artificial objects or natural objects according to their geometrical properties such as slope and curvature. Height maps are integrated into a local map by matching geometrical parameters and updating region labels.

Thrun [92] presents an approach that allows mobile robots to automatically se-

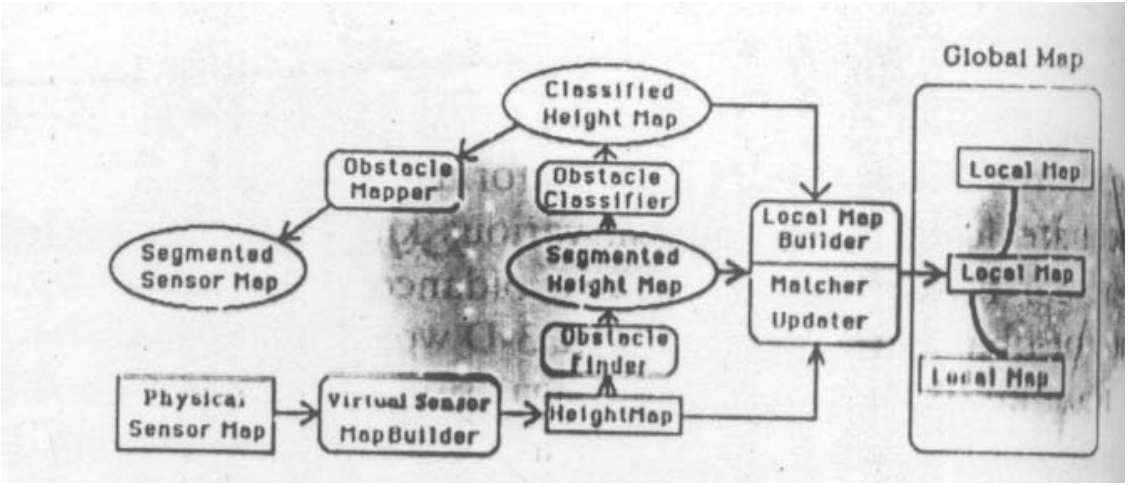


Fig. 10. Overview of the map building system in [90]: Figure 2, page 1328.

lect landmarks. Landmarks are chosen based on their utility for localization. He achieves this task by training landmark detectors so as to minimize the *a posteriori* localization error that the robot is expected to make after querying its sensors. The system trains a set of neural networks, each of which maps sensor input to a single value estimating the presence or absence of a particular landmark. He shows that using active perception helps in faster localization than with a static camera configuration. In [93], Thrun, Burgard and Fox address the problem of building large-scale geometric maps of indoor environments with mobile robots. In their experiments, they investigate a restricted version of the map-building problem, where a human operator tele-operates the robot through an environment. They pose the map-building problem as a constrained, probabilistic maximum-likelihood estimation problem. They demonstrate experimental results in cyclic environments of sizes up to 80 by 25 metres.

Map building strategies use two major paradigms to represent the environment – grid-based, and topological. While grid-based methods produce accurate metric maps, their complexity often prohibits efficient path planning. (Schiele and Crowley [94] examine the problem of pose estimation using occupancy grids.) Topological maps do not suffer from this problem. However, accurate and consistent topological maps are often difficult to learn and maintain in large-scale environments. Thrun [95] proposes an approach that integrates both paradigms. The approach learns grid-based maps using artificial neural networks and Bayesian integration of sensor output. Topological maps are generated on top of the grid-based maps by partitioning the latter into coherent regions. The paper presents results for autonomous exploration, mapping and operation of a mobile robot in populated multi-room environments.

Basri and Rivlin [96] present a method of representation that may be useful for a reactive vision-based navigating robot. The authors extend the work of Ullman and Basri [97] on recognition by a linear combination of models. They analyze three basic tasks in autonomous robot navigation namely, localization, positioning and homing. They define localization as the act of recognizing the environment *i.e.*, assigning consistent labels to different locations. Positioning is the act of computing the coordinates of the robot in the environment. Homing is the task of returning to a previously visited position. The authors represent a scene as a set of 2-D views and predict the appearances of novel views by linear combinations of the model views. They assume weak perspective projection. For the case when the weak perspective assumption is invalid, they propose using either a larger number of models, or an iterative solution for perspective distortions. They present a method for localization from only a single 2-D view without calibration. They have a similar method for positioning, and a simple qualitative algorithm for homing.

Kosaka and Kak [98] present a fast vision-guided robot navigation system FINALE using model-based reasoning and the prediction of uncertainties. Although this system is primarily meant for a path planning task, many ideas presented here are relevant for scene interpretation as well. The vision system maintains a model of uncertainty and keeps track of the growth of uncertainty as the robot travels towards the goal position. For example, the uncertainty with respect to a line is modeled as the convex hull for the two ellipses of uncertainty at the end-points of the line. These ellipses of uncertainty depend on the mean vector and covariances matrices of the uncertainty in position associated with the end points of the line. The system uses these uncertainty estimates to predict bounds on the locations and orientations of landmarks expected to be seen in a monocular image. There is a sequential reduction in uncertainty as each image feature is matched successfully with a landmark, allowing subsequent features to be matched more easily.

Fennema *et al.* [99] describe an autonomous robot navigation system at the University of Massachusetts, Amherst. Model-based processing of the visual sensory data is the primary mechanism used for controlling movement through the environment, measuring progress towards a given goal, and avoiding obstacles. They assume a partially modeled unchanging environment containing no unmodeled obstacles. the system integrated perception, planning and execution of actions. The system models the environment using a CAD modeler. The system uses reactive planning processes that reason about landmarks that should be perceived at various stages of task execution. The correspondence information between image features and expected landmark locations (the system uses line features) is used at several abstraction levels to ensure proper plan execution. For example, when the image of a landmark moves differently from what is expected, the system makes corrections to the motor system. The system proposes partially-developed tentative plans

about what action to take next. These are developed depth-first with less developed details away from the current location. Failures trigger changes in plans at various levels. Landmarks selected from the model are used to steer the robot. Chenavier and Crowley [100] describe a method for position estimation for a mobile robot, using vision-based and odometric information. The system uses landmarks for correcting the position and orientation of a robot vehicle. There are numerous other examples of landmark-based navigation strategies *e.g.*, Levitt and Lawton [101], Onoguchi *et al.* [102],

Burgard *et al.* [103] present a modular and distributed software architecture of an autonomous interactive tour-guide robot. The architecture integrates localization, mapping, collision avoidance, planning and various modules concerned with user interaction and Web-based tele-presence. The authors demonstrate results of the deployment of their system in a densely populated museum for a period of six days. Chen and Tsai [104] present a incremental-learning-by-navigation approach to vision-based autonomous land vehicle (ALV) guidance in indoor environments. The approach consists of three stages – initial (manual) learning, navigation and model updating. In the navigation stage, the ALV moves along the learned environment automatically. It locates itself by model matching, and records necessary information for model updating. The approach uses information about vertical lines. In the model-updating stage, the system refines the learned model off-line. A more precise model is obtained after each navigation-and-update iteration. The authors show results on a real ALV in indoor corridors.

5 An Analysis of Scene Interpretation Systems

Similar to our analysis of object recognition schemes, we analyze different scene analysis systems on the basis of the following issues:

(1) *Features used for modeling and view recognition*
Existing scene analysis systems primarily work with geometric features, irrespective of whether they are obtained from a vision-based sensor, a range sensor, a haptic sensor, or ultrasonic sensors. Systems such as that of Grimson *et al.* [79] additionally use colour information.

(2) *The system setup and viewing geometry*
Object data acquisition systems, and systems for recovering surface shape, both assume that the object completely fits into the sensor's field of view. For the other application areas, the entire scene may not fall within the sensor's field of view. The aim of these systems is to use the sensor in a purposive manner, to fulfill its task. The sensors for scene analysis applications typically have three translational and one rotational degree of freedom (*e.g.*, navigational applications as in the system of Kosaka and Kak [98]). Some systems such as those of Rimey and Brown [25] do not make any explicit assumptions

about the viewing geometry. Systems such as that of Basri and Rivlin [96] explicitly assume weak perspective projection, while those of Lebègue and Aggarwal [85], [86] assume perspective projection.

(3) *Representation of domain knowledge*
Different scene analysis applications need different representation schemes to fulfill their requirements. Rimey and Brown [25] use Bayes nets to represent domain knowledge, and encode task specifications. In their system for 3-D reconstruction of static scene, Marchand and Chaumette [78] propose a prediction/verification scheme using decision theory and Bayes nets. The visual surveillance system of Buxton and Gong [83] uses many different representations for its components, such as Bayes nets and ground plane maps. Artificial neural networks form the architecture of systems that use some form of learning, such as those of Baluja and Pomerleau [82], and Thrun [92].

As mentioned in Section 4, Active map-building strategies primarily consider grid-based maps (*e.g.*, Nashashibi *et al.* [84], Elfes [91] (using sonar data)) as against topological maps (*e.g.*, [105] (using sonar and laser range data)). Thrun [95] proposes an approach that integrates both paradigms. Basri and Rivlin [96] represent a scene in terms of 2-D views as against the representation of Marchand and Chaumette [78], who use explicit 3-D geometric models.

(4) *Algorithms for hypothesis generation and next view planning*
Algorithms vary according to the nature of the application. Systems may use explicit scene information to compute the next view. The approach of Maver and Bajcsy [73] uses occlusion information, while that of Kutulakos and Dyer (*e.g.*, [76] uses curvature measurements on the occluding contour. The strategy may be based on minimizing an uncertainty function as in [77]. Grimson and co-workers [79] use colour and stereo features in their multi-stage algorithm. Rimey and Brown [25] use a benefit-cost analysis to plan actions. There may be a high-level general control paradigm, as in the approach of Wixson and Ballard [80]. Map-building algorithms primarily focus on algorithms for integrating evidences taken at different points in space and time, such as that in [93]. Reactive navigation strategies primarily focus on reaching a goal, subject to positional uncertainty and navigational obstacles.

(5) *Nature of the next view planning strategy*
All systems described in Section 4 have an on-line component. The on-line nature of such systems illustrates their reactive property – an essential requirement of an active scene analysis system.

(6) *Uncertainty handling capability*
Some approaches such as that of Maver and Bajcsy [73] are deterministic. Most systems handle uncertainty explicitly. Uncertainty representation schemes include probability theory (as in the work of Marchand and Chaumette [78], Rimey and Brown [25] and Thrun *et al.* [92], [93]), Dempster-Shafer theory (as in the system of Tirumalai, Schunck and Jain [89]), and Fuzzy logic (*e.g.* the real-time map building and navigation system of Oriolo *et al.* [69] which does not use vision sensors)

6 Conclusions

Sections 2 and 3 survey and analyze different active 3-D object recognition systems. We repeat the process for different scene analysis systems (Sections 4 and 5) due to the commonality of many issues in the two problems. Based on this survey and analysis, we draw the following conclusions:

- Geometric features are useful in a recognition task. We may supplement them with other features such as colour and photometric information. Some recognition systems are tightly coupled with the properties of the particular features they use. However in some cases, we may have a system that is not explicitly based on any particular set of features.
- The 1-DOF (rotational) case between the object and an orthographic camera is an important and fairly complex problem. The complexity of the recognition task increases with the number of degrees of freedom between the object and the camera, and the increasing generality of the camera model – from orthographic to projective.
- The knowledge representation scheme plays an important role in both generating hypotheses corresponding to a given view, as well as in planning the next view.
- Noise may corrupt the output of feature detectors used for analyzing a given view. An important issue is accounting for noise at both the model-building stage, as well as in the recognition phase.
- A system with uncertainty handling capability gives it an edge over one that uses a pure deterministic strategy – the former is more robust to errors.
- It is often desirable to use an uncalibrated camera for recogntion. An active vision system may *purposively* change either the external parameters of the camera (*e.g.*, the 3-D position), or the internal parameters (*e.g.*, zoom-in/zoom-out). The planning scheme should take these factors into account.
- The domain of application influences the design of the recognition algorithm. In general, the system should plan a minimal number of steps (each step corresponds to an epoch where sensor data is processed) in order to achieve its goal. Such a process is subject to memory and processing limitations, if any.
- The next view planning strategy should preferably be on-line. The system should balance plan-based schemes and pure reactive behaviour. A pure reactive behaviour may veer a system away from its goal. On the other hand, the reactive nature of a system allows it to handle unplanned situations.

References

[1] P. J. Besl, R. C. Jain, Three-Dimensional Object Recognition, ACM Computing Surveys 17 (1) (1985) 76 – 145.

[2] R. T. Chin, C. R. Dyer, Model Based Recognition in Robot Vision, ACM Computing

Surveys 18 (1) (1986) 67 – 108.

[3] A. Zisserman, D. Forsyth, J. Mundy, C. Rothwell, J. Liu, N. Pillow, 3D Object Recognition using Invariance, Artificial Intelligence 78 (1995) 239 – 288.

[4] D. P. Mukherjee, D. Dutta Majumder, On Shape from Symmetry, Proc. Indian National Science Academy 62, A (5) (1996) 415 – 428.

[5] B. V. Funt, G. D. Finlayson, Color Constant Color Indexing, IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (1995) 522 – 529.

[6] S. K. Nayar, R. M. Bolle, Reflectance Based Object Recognition, International Journal of Computer Vision 17 (3) (1996) 219 – 240.

[7] J. B. Burns, R. S. Weiss, E. M. Riseman, The Non-Existence of General-Case View-Invariants, in: A. Zisserman, J. Mundy (Eds.), Geometric Invariance in Computer Vision, MIT Press, 1992.

[8] D. P. Mukherjee, A. Zisserman, J. M. Brady, Shape from Symmetry: Detecting and Exploiting Symmetry in Affine Images, Phil. Trans. R. Soc. London. A 351 (1995) 77 – 106.

[9] N. Pillow, Recognition of Generalized Cylinders using Geometric Invariance, Ph.D. thesis, University of Oxford (1996).

[10] N. Pillow, S. Utcke, A. Zisserman, Viewpoint Invariant Representation of Generalized Cylinders uing the Symmetry Set, Image and Vision Computing 13 (5) (1995) 355 – 365.

[11] L. Van Gool, T. Moons, D. Ungureanu, E. Pauwels, Symmetry from Shape and Shape from Symmetry, International Journal of Robotics Research 14 (5) (1995) 407 – 424.

[12] R. Choudhury, S. Chaudhury, J. B. Srivastava, A Framework for Reconstruction based Recognition of Partially Occluded Repeated Objects, Journal of Mathematical Imaging and Vision 14 (1) (2001) 5 – 20.

[13] R. Choudhury, J. B. Srivastava, S. Chaudhury, Reconstruction based Recognition of Scenes with Translationally Repeated Quadrics, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (6) (2001) 617 – 632.

[14] K. D. Gremban, K. Ikeuchi, Planning Multiple Observations for Object Recognition, International Journal of Computer Vision 12 (2/3) (1994) 137 – 172, Special Issue on Active Vision II.

[15] Y. Aloimonos, I. Weiss, A. Bandopadhyay, Active Vision, International Journal of Computer Vision 1 (4) (1987) 333 – 356.

[16] R. Bajcsy, M. Campos, Active and Exploratory Perception, Computer Vision, Graphics and Image Processing 56 (1) (1985) 31 – 40.

[17] R. Bajcsy, Active Perception, Proceedings of the IEEE 76 (8) (1988) 996 – 1005.

[18] D. H. Ballard, Animate Vision, Artificial Intelligence 48 (1) (1991) 57 – 86.

[19] D. H. Ballard, C. M. Brown, Principles of Animate Vision, Computer Vision, Graphics and Image Processing: Image Understanding 56 (1) (1992) 3 – 21.

[20] J. L. Crowley, ECVNet Tutorial on Active Computer Vision, http://wwwprima.imag.fr/ECVNet/Tutorial/av_tutorial.html.

[21] A. Blake, A. Yuille (Eds.), Active Vision, The MIT Press, 1992.

[22] M. J. Swain, M. A. Stricker, Promising Directions in Active Vision, International Journal of Computer Vision 11 (2) (1993) 109 – 126.

[23] D. W. Murray, P. F. McLauchlan, I. D. Reid, P. M. Sharkey, Reactions to Peripheral Image Motion using a Head/Eye Platform, in: Proc. IEEE International Conference on Computer Vision (ICCV), 1993, pp. 403 – 411.

[24] J. L. Crowley, J. M. Bedrune, M. Bekker, M. Schneider, Integration and Control of Reactive Visual Processes, in: Proc. European Conference on Computer Vision (ECCV), 1994, pp. II:47 – 58.

[25] R. D. Rimey, C. M. Brown, Control of Selective Perception using Bayes Nets and Decision Theory, International Journal of Computer Vision 12 (2/3) (1994) 173 – 207, Special Issue on Active Vision II.

[26] C. Colombo, J. L. Crowley, Uncalibrated Visual Tasks via Linear Interaction, in: Proc. European Conference on Computer Vision (ECCV), 1996, pp. II:583 – 592.

[27] K. A. Tarabanis, P. K. Allen, R. Y. Tsai, A Survey of Sensor Planning in Computer Vision, IEEE Transactions on Robotics and Automation 11 (1) (1995) 86 – 104.

[28] K. A. Tarabanis, R. Y. Tsai, P. K. Allen, The MVP Sensor Planning System for Robotic Vision Tasks, IEEE Transactions on Robotics and Automation 11 (1) (1995) 72 – 85.

[29] D. R. Roberts, A. D. Marshall, Viewpoint Selection for Complete Surface Coverage of Three Dimensional Objects, in: Proc. British Machine Vision Conference (BMVC), 1998, pp. 740 – 750.

[30] C. K. Cowan, P. D. Kovesi, Automatic Sensor Placement from Vision Task Requirements, IEEE Transactions on Pattern Analysis and Machine Intelligence 10 (3) (1988) 407 – 416.

[31] G. H. Tarbox, S. N. Gottschlich, Planning for Complete Sensor Coverage in Inspection, Computer Vision and Image Understanding 61 (1) (1995) 84 – 111.

[32] S. O. Mason, A. Grun, Automatic Sensor Placement for Accurate Dimensional Inspection, Computer Vision and Image Understanding 61 (1995) 454 – 467.

[33] R. Kasturi, R. C. Jain, Computer Vision: Principles, IEEE Computer Society Press Tutorial, 1991, Ch. 5: Three-Dimensional Object Recognition.

[34] H. Murase, S. K. Nayar, Visual Learning and Recognition of 3-D Objects from Appearance, International Journal of Computer Vision 14 (1995) 5 – 24.

[35] T. M. Breuel, View-Based Recogntion, in: Proc. IAPR Workshop on Machine Vision Applications, 1992.

[36] H. Borotschnig, L. Paletta, M. Prantl, A. Pinz, Active Object Recognition in Parametric Eigenspace, in: Proc. British Machine Vision Conference (BMVC), 1998, pp. 629 – 638.

[37] B. Schiele, J. L. Crowley, Probabilistic Object Recognition using Multidimensional Receptive Field Histograms, in: Proc. International Conference on Pattern Recognition (ICPR), 1996.

[38] B. Schiele, J. L. Crowley, Transinformation for Active Object Recognition, in: Proc. IEEE International Conference on Computer Vision (ICCV), 1998, pp. 249 – 254.

[39] J. J. Koenderink, A. J. van Doorn, The Internal Representation of Solid Shape with Respect to Vision, Biological Cybernetics 32 (1979) 211 – 216.

[40] I. Chakravarty, H. Freeman, Characteristic Views as a Basis for Three Dimensional Object Recognition, in: Proc. SPIE Conference on Robot Vision, Vol. 336, 1982, pp. 37 – 45.

[41] K. D. Gremban, K. Ikeuchi, Appearance-Based Vision and the Automatic Generation of Object Recognition Programs, in: A. K. Jain, P. J. Flynn (Eds.), Three-Dimensional Object Recognition Systems, Elsevier-Science Publishers, 1993, pp. 229 – 258.

[42] D. W. Eggert, K. W. Bowyer, C. R. Dyer, H. I. Christensen, D. B. Goldgof, The Scale Space Aspect Graph, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (11) (1993) 1114 – 1130.

[43] P. J. Flynn, A. K. Jain, BONSAI: 3-D Object Recognition Using Constrained Search, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (10) (1991) 1066 – 1075, Special Issue on Interpretation of 3-D Scenes - Part I.

[44] S. J. Dickinson, A. P. Pentland, A. Rosenfield, Qualitative 3D Shape Reconstruction using Distributed Aspect Graph Matching, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (2) (1992) 174 – 198.

[45] S. J. Dickinson, A. P. Pentland, A. Rosenfield, From Volumes to Views: An Approach to 3D Object Recognition, Computer Vision, Graphics and Image Processing: Image Understanding 55 (2) (1992) 198 – 211.

[46] K. Sato, K. Ikeuchi, T. Kanade, Model Based Recognition of Specular Objects Using Sensor Models, Computer Vision, Graphics and Image Processing: Image Understanding 55 (2) (1992) 119 – 129.

[47] K. Ikeuchi, T. Kanade, Automatic Generation of Object Recognition Programs, Proceedings of the IEEE 76 (8) (1988) 1016 – 1035.

[48] K. Ikeuchi, T. Kanade, Modeling Sensors: Towards Automatic Generation of Object Recognition Programs, Computer Vision, Graphics and Image Processing 48 (1989) 50 – 79.

[49] M. Robey, G. West, S. Venkatesh, An Investigation into the Use of Physical Modeling for the Prediction of Various Feature Types Visible from Different Viewpoints, Computer Vision and Image Understanding 61 (3) (1995) 417 –429.

[50] H. Lu, L. G. Shapiro, A Relational Pyramid Approach to View Class Determination, in: Proc. International Conference on Pattern Recognition (ICPR), 1988, pp. 379 – 381.

[51] H. Lu, L. G. Shapiro, O. I. Camps, A Relational Pyramid Approach to View Class Determination, in: Proc. IEEE Workshop on Interpretation of 3D Scenes, 1989, pp. 177 – 183.

[52] O. I. Camps, L. G. Shapiro, R. M. Haralick, PREMIO: An Overview, in: Proc. IEEE International Workshop on Directions in Automated CAD Based Vision, 1991, pp. 11 – 21.

[53] O. Munkelt, Aspect-Trees: Generation and Implementation, Computer Vision and Image Understanding 61 (3) (1995) 365 – 386.

[54] S. Dutta Roy, S. Chaudhury, S. Banerjee, Aspect Graph Construction with Noisy Feature Detectors, IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics (Accepted for Publication).

[55] S. A. Hutchinson, A. C. Kak, Planning Sensing Strategies in a Robot Work Cell with Multi-Sensor Capabilities, IEEE Transactions on Robotics and Automation 5 (6) (1989) 765 – 783.

[56] S. J. Dickinson, H. I. Christensen, J. Tsotsos, G. Olofsson, Active Object Recognition Integrating Attention and View Point Control, Computer Vision and Image Understanding 67 (3) (1997) 239 – 260.

[57] S. Dutta Roy, S. Chaudhury, S. Banerjee, Isolated 3-D Object Recognition through Next View Planning, IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans 30 (1) (2000) 67 – 76.

[58] S. Dutta Roy, S. Chaudhury, S. Banerjee, Aspect Graph Based Modeling and Recognition with an Active Sensor: A Robust Approach, Proc. Indian National Science Academy, Part A 67 (2) (2001) 187 – 206.

[59] S. Dutta Roy, S. Chaudhury, S. Banerjee, Recognizing Large 3-D Objects through Next View Planning using an Uncalibrated Camera, in: Proc. IEEE International Conference on Computer Vision (ICCV), 2001, pp. II: 276 – 281.

[60] I. Biederman, Human Image Understanding: Recent Research and a Theory, Computer Vision, Graphics and Image Processing 32 (1985) 29 – 73.

[61] R. Bergevin, M. Levine, Part Decomposition of Objects from Single View Line Drawings, Computer Vision, Graphics and Image Processing: Image Understanding 55 (1) (1992) 73 – 83.

[62] R. Bergevin, M. Levine, Generic Object Recognition: Building and Matching Coarse Descriptions from Line Drawings, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (1) (1993) 19 – 36.

[63] C. Y. Huang, O. I. Camps, T. Kanungo, Object Recognition Using Appearance-Based Parts and Relations, in: Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 1997, pp. 877 – 883.

[64] O. I. Camps, C. Y. Huang, T. Kanungo, Hierarchical Organization of Appearance-Based Parts and Relations for Object Recognition, in: Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 1998, pp. 685 – 691.

[65] A. P. Dempster, A Generalization of Bayesian Inference, Journal of the Royal Statistical Society 30 (Series B) (1968) 205 – 247.

[66] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, Princeton, N. J., 1976.

[67] L. A. Zadeh, Fuzzy Sets, Information and Control 8 (1965) 338 – 353.

[68] J. Pearl, Fusion, Propagation and Structuring in Belief Networks, Artificial Intelligence 29 (1986) 241 – 288.

[69] G. Oriolo, G. Ulivi, M. Vendittelli, Real-Time Map Building and Navigaiton for Autonomous Robots in Unknown Environments, IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics 28 (3) (1998) 316 –333.

[70] C.-H. Liu, W.-H. Tsai, 3D Curved Object Recognition from Multiple 2D Camera Views, Computer Vision, Graphics and Image Processing 50 (1990) 177 – 187.

[71] F. G. Callari, F. P. Ferrie, Active Recognition: Using Uncertainty to Reduce Ambiguity, in: Proc. International Conference on Pattern Recognition (ICPR), 1996, pp. 925 – 929.

[72] M. Werman, S. Banerjee, S. Dutta Roy, M. Qiu, Robot Localization Using Uncalibrated Camera Invariants, in: Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 1999, pp. II: 353 – 359.

[73] J. Maver, R. Bajcsy, Occlusions as a Guide for Planning the Next View, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (5) (1993) 76 – 145.

[74] N. A. Massios, R. B. Fisher, A Best Next View Selection Algorithm Incorporating a Quality Criterion, in: Proc. British Machine Vision Conference (BMVC), 1998, pp. 780 – 789.

[75] M. A. García, S. Velázquez, A. D. Sappa, A Two-Stage Algorithm for Planning the Next View From Range Images, in: Proc. British Machine Vision Conference (BMVC), 1998, pp. 720 – 729.

[76] K. N. Kutulakos, C. R. Dyer, Recovering Shape by Purposive Viewpoint Adjustment, International Journal of Computer Vision 12 (2/3) (1994) 113 – 136, Special Issue on Active Vision II.

[77] P. Swain, F. P. Ferrie, From Uncertainity to Visual Exploration, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (10) (1991) 1038 – 1049, Special Issue on Interpretation of 3-D Scenes - Part I.

[78] E. Marchand, F. Chaumette, An Autonomous Active Vision System for Complete and Accurate 3D Scene Reconstruction, International Journal of Computer Vision 32 (3) (1999) 171 – 194.

[79] W. E. L. Grimson, A. Lakshmi Ratan, P. A. O'Donnel, G. Klanderman, An Active Vision System to "Play Where's Waldo", in: Proc. DARPA Conference on Image Understanding, 1994.

[80] L. E. Wixson, D. H. Ballard, Using Intermediate Objects to Improve the Efficiency of Visual Search, International Journal of Computer Vision 12 (2/3) (1994) 209 – 230, Special Issue on Active Vision II.

[81] F. Jensen, H. Christensen, J. Nielsen, Bayesian Methods for Interpretaiton and Control in Multiagent Vision Systems, in: K. W. Bowyer (Ed.), SPIE Applications of AI X: Machine Vision and Robotics, Vol. 1708, 1992, pp. 536 – 548.

[82] S. Baluja, D. Pomerleau, Dynamic Relevance: Vision-Based Focus of Attention using Artificial Neural Networks, Artificial Intelligence 97 (1-2) (1997) 381 – 395, Special Issue on Relevance.

[83] H. Buxton, S. Gong, Visual Surveillance in a Dynamic and Uncertain World, Artificial Intelligence 78 (1995) 431 – 459.

[84] F. Nashashibi, M. Davy, P. Fillatreau, Indoor Scene Terrain Modeling using Multiple Range Images for Autonomous Mobile Robots, in: Proc. IEEE International Conference on Robotics and Automation (ICRA), 1992, pp. I:40 – 46.

[85] X. Lebègue, J. K. Aggarwal, Extraction and Interpretation of Semantically Significant Line Segments for a Mobile Robot, in: Proc. IEEE International Conference on Robotics and Automation (ICRA), 1992, pp. I:1778 – 1785.

[86] X. Lebègue, J. K. Aggarwal, Generation of Architectural CAD Models Using a Mobile Robot, in: Proc. IEEE International Conference on Robotics and Automation (ICRA), 1994, pp. I:711 – 717.

[87] O. D. Faugeras, N. Ayache, B. Faverjon, Building Visual Maps by Combining Noisy Stereo Measurements, in: Proc. IEEE International Conference on Robotics and Automation (ICRA), 1986, pp. III:1433 – 1438.

[88] O. Faugeras, Three-Dimensional Computer Vision: A Geometric Viewpoint, The MIT Press, 1996.

[89] A. P. Tirumalai, B. G. Schunck, R. C. Jain, Evidential Reasoning for Building Environmental Maps, IEEE Transactions on Systems, Man and Cybernetics 25 (1) (1995) 10 – 20.

[90] M. Asada, Map Building for a Mobile Robot from Sensory Data, IEEE Transactions on Systems, Man and Cybernetics 37 (6) (1990) 1326 – 1336.

[91] A. Elfes, Sonar-Based Real-World Mapping and Navigation, IEEE Journal of Robotics and Automation RA-3 (3) (1987) 249 – 265.

[92] S. Thrun, A Bayesian Approach to Landmark Discovery and Active Perception in Mobile Robot Navigation, Tech. Rep. CMU-CS-96-122, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 (May 1996).

[93] S. Thrun, W. Burgard, D. Fox, A Probabilistic Approach to Concurrent Mapping and Localization for Mobile Robots, Machine Learning 31 (1998) 29 – 53.

[94] B. Schiele, J. L. Crowley, A Comparison of Position Estimation Techniques using Occupancy Grids, in: Proc. IEEE International Conference on Robotics and Automation (ICRA), 1994, pp. II:1628 – 1634.

[95] S. Thrun, Learning Metric Topological Maps for Indoor Mobile Robot Navigation, Artificial Intelligence 99 (1) (1998) 21 – 71.

[96] R. Basri, E. Rivlin, Localization and Homing using Combinations of Model Views, Artificial Intelligence 78 (1-2) (1995) 327 – 354.

[97] S. Ullman, R. Basri, Recognition by Linear Combination of Models, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (1991) 992 – 1006.

[98] A. Kosaka, A. C. Kak, Fast Vision-Guided Mobile Robot Navigaiton Using Model-Based Reasoning and Prediction of Uncertainties, Computer Vision, Graphics and Image Processing: Image Understanding 56 (3) (1992) 271 – 329.

[99] C. Fennema, A. Hanson, E. Riseman, J. Ross Beveridge, R. Kumar, Model-Directed Mobile Robot Navigation, IEEE Transactions on Systems, Man and Cybernetics 20 (6) (1990) 1352 – 1369.

[100] F. Chenavier, J. L. Crowley, Position Estimation for a Mobile Robot using Vision and Odometry, in: Proc. IEEE International Conference on Robotics and Automation (ICRA), 1992, pp. III:2588 – 2593.

[101] T. S. Levitt, D. T. Lawton, Qualitative Navigation for Mobile Robots, Artificial Intelligence 44 (1990) 305 – 360.

[102] K. Onoguchi, M. Watanabe, Y. Okamoto, Y. Kuno, H. Asada, A Visual Navigation System Using a Multi-Information Local Map, in: Proc. IEEE International Conference on Robotics and Automation (ICRA), 1990, pp. II:767 – 774.

[103] W. Burgard, A. B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, S. Thrun, Experiences with an Interactive Museum Tour-Guide Robot, Artificial Intelligence 114 (1-2) (1999) 3 – 55.

[104] G. Y. Chen, W. H. Tsai, An Incremental-Learning-by-Navigation Approach to Vision-Based Autonomous Land Vehicle Guidance in Indoor Environments Using Vertical Line Information and Multiweighted Generalized Hough Transform Technique, IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics 28 (5) (1998) 740 – 748.

[105] B. Yamauchi, R. Beer, Spatial Learning for Navigation in Dynamic Environments, IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics 26 (3) (1996) 496 – 505, Special Issue on Learning Autonomous Robots.