Active 3-D Object Recognition through Next View Planning

Sumantra Dutta Roy
Department of Electrical Engineering,
Indian Institute of Technology Bombay, Powai, Mumbai - 400 076, INDIA.
E-mail: sumantra@ee.iitb.ac.in
Web: http://www.ee.iitb.ac.in/∼sumantra

Abstract

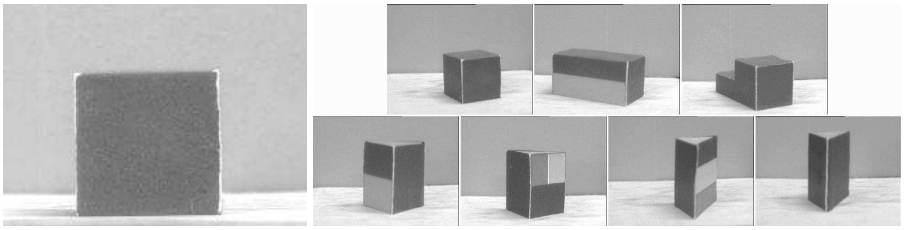
This article presents an overview of work on Active 3-D Object Recognition at the Indian Institute of Technology, Bombay and Indian Institute of Technology, Delhi. We have concentrated on the use of simple features and suitably planned multiple views to recognise a 3-D object with an uncalibrated camera. We use isolated planned views, without incurring the overhead of tracking the object of interest, across views. Our work has focussed primarily on two areas: aspect graph-based modelling and recognition using noisy sensors; and recognising *large* 3-D objects using Inner Camera Invariants. We have proposed new hierarchical knowledge representation schemes in both cases. A common thread in both is a novel robust probabilistic reasoning-based reactive object recognition strategy which is scalable to memory and processing constraints, if any. We present results of numerous experiments in support of our proposed strategies.

1 Introduction

3-D object recognition is a difficult task primarily because of the loss of information in the basic 3-D to 2-D imaging process. Most model-based 3-D object recognition systems consider features from a single image, using properties invariant to an object, and preferably, invariant to the viewpoint. We often need to recognise 3-D objects which because of their inherent asymmetry (in any set of features: geometric, photometric, or colour-based, for example), cannot be completely characterised by an invariant computed from a single view. In order to use multiple views for an object recognition task, one needs to maintain the relationship between different views of an object. In single-view recognition, systems often use complex feature sets, which are not easy to extract from images. In many cases, it may be possible to achieve unambiguous recognition using simple features, and suitably planned multiple views [6, 1].

A single view of a 3-D object often does not contain sufficient features to recognise it unambiguously. Objects which have two or more views in common with respect to a feature set, may be distinguished through a sequence of views. As a simple example [6,

1], let us consider the set of features to be the number of horizontal and vertical lines, and a model base of polyhedral objects. Fig. 1(a) shows a given view of an object. All



(a) (b)

Figure 1: (a) The given complete view of an object, and (b) the objects which this view could correspond to: This is Fig. 1 in [6], page 430.

objects in Fig. 1(b) have at least one view which correspond to two horizontal and two vertical lines.

A further complication arises if the given 3-D object does not fit inside the camera's field of view. Fig. 2 shows an example of such a case. The view in Fig. 3(a) could have

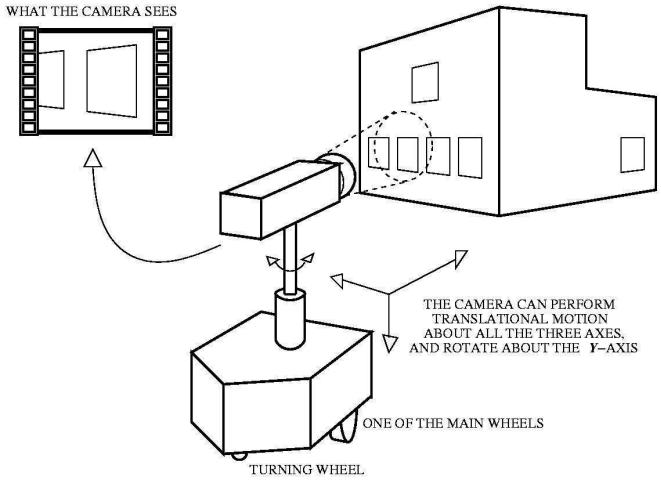
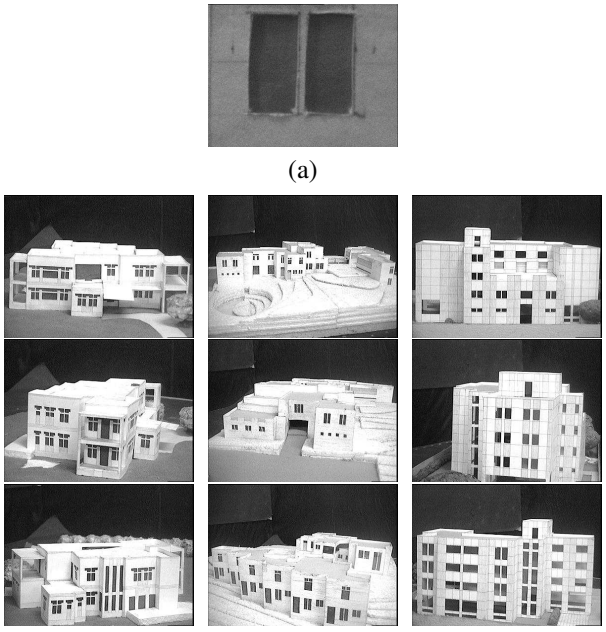


Figure 2: A robot with an attached camera, observing a building. The entire object does not fit in the camera's field of view. Not only is the identity of the object unknown, the robot also does not know its pose with respect to the object. This example shows 4 degrees of freedom (DOF) between the object and the camera. This is Fig. 2 in [7], page 282.

come from any of the objects in Fig. 3(b), (c) and (d). Even if the identity of the object were known, one may often like to know what part of the object the camera is looking at – the pose of the camera with respect to the object. Single-view recognition systems often use complex feature sets, which are associated with high feature extraction costs,

which in itself, may be noisy. A simple feature set is more applicable for a larger set of objects. In many cases, it may be possible to achieve recognition using a simpler feature set and suitably planned multiple observations.



(a) (b) (c) (d)

Figure 3: (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. This is Fig. 2 in [6], page 431.

An **Active Sensor** is one whose parameters can be varied in a purposive manner. For a camera, this implies a purposive control over the external parameters (the parameters $\mathbf{R}$ and $\mathbf{t}$ describing the 3-D Euclidean transformation between the camera coordinate system and a world coordinate system) and the internal parameters (given by the internal camera parameter matrix $\mathbf{A}$, composed of parameters: the focal lengths in the $x-$ and $y-$ image directions, the position of the principal point, and the skew factor). Our work in active 3-D object recognition has primarily considered two areas namely,

1. Aspect Graph-based Modelling and Recognition using Noisy Sensors
2. Recognition of *Large* 3-D objects through Next View Planning using Inner Camera Invariants

The following sections give an overview of our work in this area.

2 Aspect Graph-based Modelling and Recognition using Noisy Sensors

We consider the problem of recognising an isolated 3-D object using simple features, and suitably planned multiple views. We assume a single rotational degree of freedom (hereafter, DOF) between the object and the (uncalibrated) camera. We propose a novel hierarchical aspect graph-based knowledge representation scheme which encodes domain knowledge about different view of the objects in the model base. This plays an important role in calculating the probabilities of different entities as evidence comes in from a view, as well as planning an optimal next view, subject to memory and processing constraints. The planning process is reactive - the system uses both the past history and the current observation to plan the next view. This feature, coupled with the explicit modeling of uncertainty, allows a high degree of robustness to feature detection errors, and does not incur a high computational cost which would be associated with an off-line system. Information from each observation prunes the search for taking a view that disambiguates between the possible view interpretation hypotheses at any stage. To serve as a benchmark, we use a simple deterministic case to show that the number of views required to disambiguate between a set of n competing aspects corresponding to the first view, is $\mathcal{O}(\log n)$. An important feature of our system is that it does not incur the overhead of tracking the region of interest across views.

We present details of this system in [2]. We have experimented with databases of reasonably complex shapes, which have a large degree of interpretation ambiguity corresponding to a view. Fig.4 shows the objects in the aircraft model base, and results of some experiments with the model base. For the experiments in Fig. 4, we use a very simple set of features: the number of horizontal and vertical lines in an image of the object, and the number of circles. In the bottom part of this figure, the initial view in each of these cases has the same features: three horizontal and vertical lines each, and two circles.

In [5], we present a characterisation of errors in aspect graphs, as well an algorithm for estimating aspect graphs, given noisy sensor data. The algorithm has low-order polynomial time complexity, in the size of the tessellated viewpoint space. We also propose a function to evaluate the output of aspect graph construction algorithms. We have examined the both a single rotational DOF case, as well as a 3-DOF case for rotations.

For the experiments in Fig. 4, the top row right set shows an example of using our aspect graph construction strategy which explicitly models feature detection errors. Due to the shadow of the left wing on the fuselage of the aircraft, the feature detector detects four vertical lines instead of three, the correct number. Our error modelling scheme in the aspect graph construction, and error handler in the object recognition system enable it to recover from this feature detection error. [3] presents a complete overview of the entire system, the aspect graph construction and the object recognition part. The papers also enumerate the three cases when our algorithm is not guaranteed to succeed.

In [8], we propose the idea of an *appearance-based aspect graph*. This is an alternative to feature-based methods, since eigenspace information captures all information

OBJECTS IN THE AIRCRAFT MODEL BASE

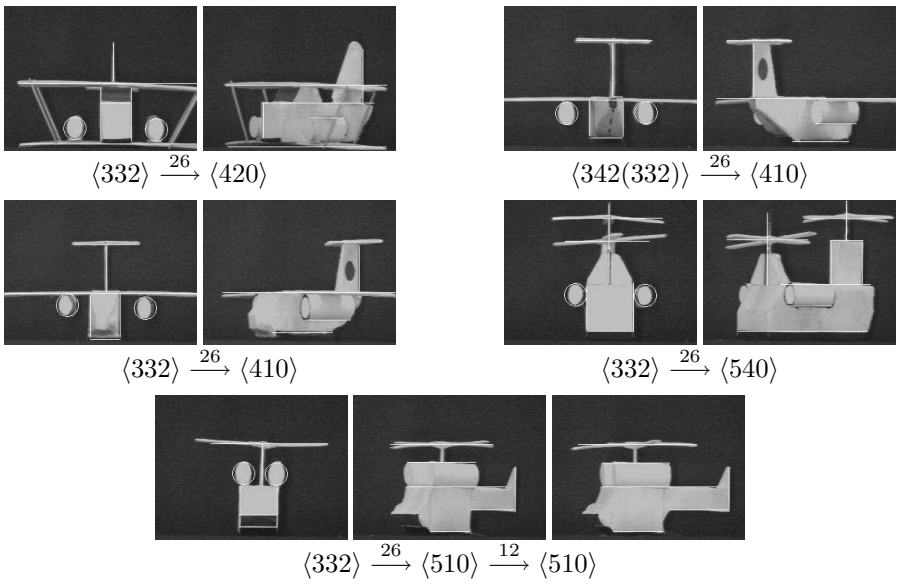


Figure 4: The set of models in our aircraft model base (top), and some experiments with our first system on the set, with the same initial view with respect to the feature set used. The numbers above the arrows denote the number of turntable steps. (The figure in parenthesis shows an example of recovery from feature detection errors). These are Fig. 5 and Fig. 6 in [6], page 436.

about the view of an object that can be obtained from an image. In this case as well, we propose a hierarchical knowledge representation scheme, and a probabilistic reasoning framework which helps in both generating hypothesis about a given view of an object, and planning the next view, if required. The scheme is robust to problems which are usually associated with appearance-based methods namely, background clutter, and size and position changes in a view of an object. The paper [8] represents work in progress, and presents some very preliminary work in this area.

3 Recognition of *Large* 3-D objects through Next View Planning using Inner Camera Invariants

Our second scheme proposes a new on-line method for recognising *large* 3-D objects, which may not fit in a camera's field of view. Unlike the previous case, we consider a

projective camera model, and consider the case when the internal camera parameters may vary - either accidentally, or on purpose (*e.g.*, a zoom-in operation to get details of a particular portion of the object, or a zoom-out operation to get a wider field of view).

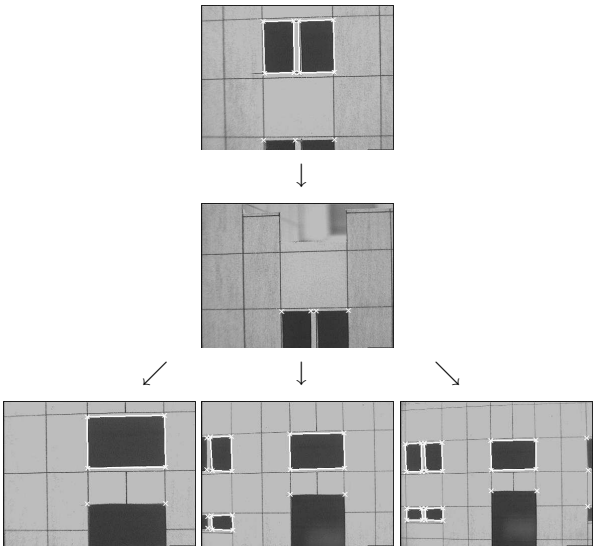
In [9, 10], we propose a new class of invariants for complete 3-D Euclidean pose estimation using an uncalibrated camera - *Inner Camera Invariants*. These are image-computable functions, independent of the internal parameters of the camera. We use these to advantage in our recognition scheme for large 3-D objects (Details in [4, 7]). We propose a new part-based recognition knowledge representation scheme. We consider a very general definition of the word 'part': in our formulation, an object is composed of parts, but is not partitioned into a set of parts. 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 do not have features detectable using the given set of feature detectors). This framework also uses a probabilistic reasoning framework for recognition and planning the next view. The part pose estimation itself helps in a first-level pruning of the list of competing view interpretation hypotheses. The system is fairly robust to small movement errors, presence of clutter in an image, and cases of non-detection of parts, in addition to be independent of changes in camera internal parameters - zoom-in and zoom-out operations, for example [4, 7]. We show results of successful recognition and pose estimation for a large uncertainty corresponding to the interpretation of a given view.

We have experimented with a set of architectural models (Fig. 3), and a set of buildings in the I.I.T. Bombay academic area. While our formulation is for a general 6-DOF case, we have experimented with a 4-DOF setup, similar to the one depicted in Fig. 2. For the architectural models, we show results of successful recognition and pose estimation even in cases of a high degree of interpretation ambiguity associated with a view. Fig. 3 shows such an example. Such a view could have come from any of the three models, different views of which are shown in Fig. 3(b), (c) and (d), respectively. Fig. 5 shows an example of the system's resilience to changes in the internal parameters of the camera. For the same two initial views, we progressively zoom the camera out at the third view. The system correctly recognises the object in each case, and estimates the pose accurately in each case ($\langle\ 9.425^\circ,\ -22.000mm,\ -9.999mm,\ 150.000mm\ \rangle$, $\langle\ 9.888^\circ,\ -22.000mm,\ -9.999mm,\ 150.000mm\ \rangle$, and $\langle\ 9.896^\circ,\ -22.000mm,\ -9.999mm,\ 150.000mm\ \rangle$, respectively). While [4] presents a preliminary description of our system, [7] describes the system in detail.

We have also experimented with an extremely difficult operating environment - buildings in the I.I.T. Bombay academic area. There are numerous trees and other unmodelled objects. Additionally, occlusions and improper lighting conditions also adversely affect the performance of the system. Fig. 6 and Fig. 7 show examples of experiments with the real buildings. We describe these in detail in [7], examine robustness issues, and state the limitations of the proposed method.

Contact

- **Sumantra Dutta Roy** completed his B. E. (Computer Engineering) from D.I.T. (currently N.S.I.T.), Delhi in 1993, and the M. Tech. and Ph. D. (Computer Science and Engineering) from I.I.T. Delhi in 1995 and 2001, respectively. He has



(a) ⟶ (b) ⟶ (c)
The camera progressively zooms out

Figure 5: 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 estimates camera pose accurately in each case. This is Fig. 10 in [7].

been an Assistant Professor in the Department of Electrical Engineering at I.I.T. Bombay since 2001. He is a recipient of the Young Engineer Award of the Indian National Academy of Engineering for the year 2004, and the BOYSCAST Fellowship of the Department of Science and Technology, Government of India for the year 2004 - 2005. His research interests include Computer Vision and Image Processing (Active 3-D Object Recognition, Tracking and Gesture Analysis, Mosaicing, Biometrics: Fingerprint Analysis), and Music Information Retrieval. For more information about his research activities, please visit `http://www.ee.iitb.ac.in/~sumantra`

- **Santanu Chaudhury** did his B. Tech. (1984) in Electronics and Electrical Communication Engineering and Ph. D. (1989) in Computer Science and Engineering from I.I.T. Kharagpur, India. Currently, he is a professor in the Department of Electrical Engineering at I.I.T. Delhi. He was awarded the Young Scientist Medal of the Indian National Science Academy in 1993. His research interests are in the areas of Computer Vision, Artificial Intelligence and Multimedia Systems.

- **Subhashis Banerjee** did his B. E. (Electrical Engineering) from Jadavpur University, Calcutta in 1982, and M. E. (Electrical Engineering) and Ph. D. from the Indian Institute of Science, Bangalore, in 1984 and 1989 respectively. Since 1989 he has been on the faculty of the Department of Computer Science and En-

Figure 6: Experiment 6 (I.I.T. Bombay Buildings): Backtracking on reaching a view without any part, and successful final recognition. This is Fig. 12 in [7] (details in paper).

Figure 7: Experiment 7 (I.I.T. Bombay buildings): Catastrophic failure - the effect of an occlusion (left), and reflection on the window panes and tree (centre). This is Fig. 13 in [7] (details in paper).

gineering at I.I.T. Delhi where he is currently a Professor. His research interests include Computer Vision and Real-time Embedded Systems. For more information about his research activities, please visit
`http://www.cse.iitd.ac.in/~suban`

References

[1] S. Dutta Roy. *Active Object Recognition through Next View Planning*. PhD thesis, Department of Computer Science and Engineering, Indian Institute of Technology, Delhi, 2001.

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

[3] S. Dutta Roy, S. Chaudhury, and S. Banerjee. Aspect Graph Based Modeling and Recognition with an Active Sensor: A Robust Approach. *Proc. Indian National Science Academy, Part A*, 67(2):187 – 206, March 2001. Special Issue on Image Processing, Vision and Pattern Recognition.

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

[5] S. Dutta Roy, S. Chaudhury, and S. Banerjee. Aspect Graph Construction with Noisy Feature Detectors. *IEEE Trans. on Systems, Man and Cybernetics - Part B: Cybernetics*, 33(2):340 – 351, April 2003.

[6] S. Dutta Roy, S. Chaudhury, and S. Banerjee. Active Recognition through Next View Planning: A Survey. *Pattern Recognition*, 37(3):429 – 446, March 2004.

[7] S. Dutta Roy, S. Chaudhury, and S. Banerjee. Recognizing Large Isolated 3-D Objects through Next View Planning using Inner Camera Invariants. *IEEE Trans. on Systems, Man and Cybernetics - Part B: Cybernetics*, 35(2):282 – 292, April 2005.

[8] S. Dutta Roy and N. Kulkarni. Active 3-D Object Recognition using Appearance-Based Aspect Graphs. In *Proc. Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pages 40 – 45, 2004.

[9] M. Werman, S. Banerjee, S. Dutta Roy, and M. Qiu. Robot Localization Using Uncalibrated Camera Invariants. In *Proc. IEEE International Conference on Computer Vision and Pattern Recogntion (CVPR)*, pages II: 353 – 359, 1999.

[10] M. Werman, M. Qiu, S. Banerjee, and S. Dutta Roy. Inner Camera Invariants and their Applications. Technical report, Department of Computer Science and Engineering, I.I.T. Delhi, August 2001. `http://www.cse.iitd.ac.in/~suban/papers/inner07.pdf`.