Next Best View Planning for Object Recognition in Mobile Robotics

Christopher McGreavy, Lars Kunze and **Nick Hawes**
Intelligent Robotics Lab
School of Computer Science
University of Birmingham
United Kingdom
{cam586|l.kunze|n.a.hawes}@cs.bham.ac.uk

Abstract

Recognising objects in everyday human environments is a challenging task for autonomous mobile robots. However, actively planning the views from which an object might be perceived can significantly improve the overall task performance. In this paper we have designed, developed, and evaluated an approach for next best view planning. Our view planning approach is based on online aspect graphs and selects the next best view after having identified an initial object candidate. The approach has two steps. First, we analyse the visibility of the object candidate from a set of candidate views that are reachable by a robot. Secondly, we analyse the visibility of object features by projecting the model of the most likely object into the scene. Experimental results on a mobile robot platform show that our approach is (I) effective at finding a next view that leads to recognition of an object in 82.5% of cases, (II) able to account for visual occlusions in 85% of the trials, and (III) able to disambiguate between objects that share a similar set of features. Hence, overall, we believe that the proposed approach can provide a general methodology that is applicable to a range of tasks beyond object recognition such as inspection, reconstruction, and task outcome classification.

1 Introduction

Autonomous mobile robots that operate in real-world environments are often required to find and retrieve task-related objects to accomplish their tasks. However, perceiving and recognising objects in such environments poses several challenges. Firstly, object locations are continuously changing. That is, a mobile robot cannot simply rely on a fixed set of views from which it can observe an object. The robot needs to plan in a continuous space where to stand and where to look. That is, the robot has to select a view from the uncountably infinite set of possible views which allows it to observe the sought object. Matters are further complicated by dynamic obstacles which might hinder the robot to take a particular view. Or, when taking a view, relevant features of an object might be occluded by other objects and/or by the object itself (self-occlusion). Finally, other conditions such as lighting and/or sensor noise can influence the performance of object recognition tasks.

In previous work, we have enabled robots to find objects in particular rooms (Kunze et al. 2012) and in relation to other objects (Kunze, Doreswamy, and Hawes 2014). These approaches guide robots to locations from which they can potentially observe an object. In this work, however, we propose a complementary planning approach that selects the *next best view* after having identified a potential object candidate. Such local, incremental view planning is crucial for two reasons: (1) it allows robots to disambiguate between objects which share a similar set of features, and (2) it improves the overall performance of object recognition tasks as objects are observed and recognized from multiple views.

Our local view planning approach addresses all of these above mentioned challenges. It is based on using a realistic sensor model of an RGB-D camera to generate online aspect graphs (a set of poses around an object which describe object visibility at that point) and takes the kinematic constraints of a mobile robot platform into account. Hence, by changing the sensor and/or the kinematic model our approach can easily be transferred to robot platforms of different types. We further consider two environmental constraints when generating robot and camera poses: (1) *dynamic obstacles* which might hinder a robot to take particular views, and (2) *occlusions of object features* which might be hidden by other objects (or by the object itself). Finally, we consider learned object models in the planning process to predict the location and visibility of features of object candidates. An overview of the next best view planning approach is given in figure 1.

Experimental results show that our next best view planning approach, which takes multiple views of an object, allows us to improve the performance of recognition tasks when compared to single-view object recognition. Further, it enables robots to differentiate between objects that share a similar set of features.

To this end, this work contributes the following:

1. a method for analysing potential views using online aspect graphs by taking both into account: (1) occlusions, and (2) the visibility of features based on learned object models.
2. a next best view planner that selects a view based on the method above and an executive that accounts for dynamic obstacles during execution
3. a set of experiments which demonstrates (a) how robots can disambiguate objects which share similar feature sets,

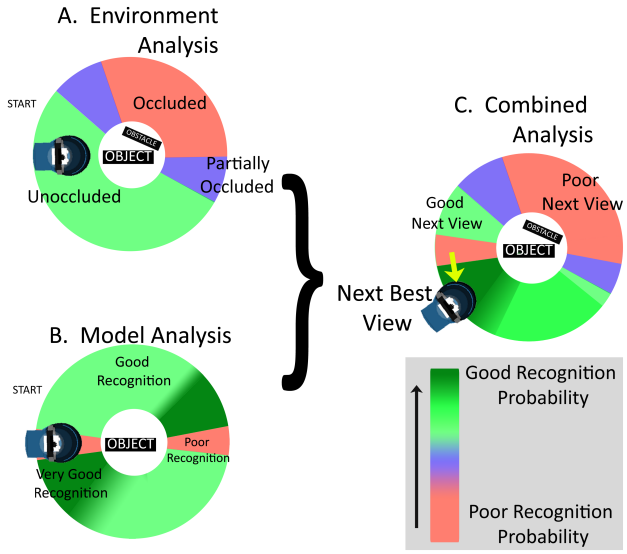

Figure 1: Next best view planning: Conceptual overview. The approach has two steps: (A) an environmental analysis and (B) a model analysis. The analysis of the environment reasons about the visibility of an object and is carried out based on a set of navigable poses. Poses in suitable areas are then carried on into the model analysis (B), in which the visibility of object features is evaluated. Conceptually, the resulting areas from (A) and (B) are combined to find the next best view (C).

and (b) how the performance of object recognition can be improved by taking multiple views.

The remainder of the paper is structured as follows. We first discuss related work in Section 2, followed by a conceptual overview of our approach in Section 3 and a detailed description of the implementation in Section 4. In Section 5, we present and analyse experimental results and discussion, before we conclude with a summary and conclusion in Section 7.

2 Related Work

Hutchinson et al. (Hutchinson and Kak 1989) used aspect graphs to determine the pose in which most unseen object features were present when digitally reconstructing an object. By storing a geometric model of an object they were able to determine what features of the object could be seen from various poses around it. The sensor would then move to the pose in which the most features were available. However, geometric analysis accounted for minute edges and ridges which are not necessarily visible to the camera. Aspect graphs were computed offline and used as a lookup table for a mobile sensor. This work does not take into account camera limitations and thus may provide a view which is theoretically optimal, but practically unobtainable. To combat this, the work presented in this paper seeks to model the sensor used in task. Aspect graphs will also be built online to account for accessibility of the environment.

The approach used by Stampfer et al. (Stampfer, Lutz, and Schlegel 2012) bares most resemblance to the current work. By taking candidates from the initial scene, they selected next best view locations that maximised the probability of recognising objects in the scene from a local object database, which is analogous to this project. But instead of planning next best view locations based on geometric analysis they used photometric methods to locate specific visual features. A camera mounted on a manipulator arm is used to sequentially move to these feature rich locations. This however, does rely on the object to have colour/contrast differences, bar codes or text, which may not be true of all objects. Photometric based analysis has its merits, but geometic feature analysis has also been shown to be effective (Vázquez et al. 2001; Roberts and Marshall 1998). This also requires a robot with a manipulator to move around the object and does not account for visual occlusions that may hamper the line of sight of the camera. Although reliable this approach requires several sequential before recognition. In this project we aim to minimise the amount of views taken.

Early work into next view planning was based on detecting which parts of an object were causing self occlusion (Bajcsy 1988)(Connelly 1985)(Maver and Bajcsy 1993). These methods were effective at obtaining high levels of coverage of an object for digital reconstruction, but high amounts of new views were needed in order to achieve this. Methods on the current work are inspired by these concepts for detecting occlusions in the environment so as to avoid them in the next view.

Okamoto (Okamoto, Milanova, and Bueker 1998) used a model free approach to move to a precomputed best pose for recognition. A stream of video like images was taken en-route to this location and produced a recognition based on this stream. This method was able to disambiguate similar looking objects. However, this solution made no consideration for environmental obstacles and no backup if the optimal pose was unavailable. In our work, the next view planner will plan to avoid environmental obstacles and will not require the use of expensive visual processing methods but will still be able to differentiate between similar objects.

Callari (Callari and Ferrie 2001) sought to use contextual information from a robot's surroundings to identify an object. This is a useful metric, as contextual information has been shown to be useful in directing object search (Hanheide et al. 2010; Kunze et al. 2014). Although, in order for this solution to work, a great deal of prior information is required to influence identification and is unlikely to cope well in a novel environment. The current work does take in prior information, but this is limited to a snapshot of the current environment which is used to detect visual occlusions.

Wixson (Wixson 1994) proposed moving at set intervals around a target for high surface coverage of the object within a fixed number of movements with little computational cost. Though this naive approach offers very low computing costs with potentially high information gain, the cost of movement would potentially huge if recognition did not occur in the first couple of movements. The present work seeks to use current information to limit the number of movements required to identify an object.

Figure 2: Illustrates how the next best view planning in this paper fits within the perception-action cycle. Blue boxes represent action/perception stages. Green boxes represent planning stages.

Vasquez-Gomez and Stampfer (Vasquez-Gomez, Sucar, and Murrieta-Cid 2014; Stampfer, Lutz, and Schlegel 2012) considered some of the restricting effects of an uncertain environment using a mobile robot when digitally reconstructing an object with a camera attached to a multi-joint manipulator on a mobile base. The main contribution of this work to the current study is that not only did it consider the placement of the sensor but also of the mobile base which was always planned to be placed in open space. These considerations are extended in this project to account for agent placement and visual occlusions.

3 Next Best View Planning

This section provides a conceptual overview of the proposed view planning approach. Implementation details can be found in Section 4.

Figure 2 shows the next best view planner's place in the perception-action cycle of a robot. A hypothesis about an object's identity and its estimated pose are the inputs into the planner. As output, the planner provides the *next best view* from which it determines the robot has the best chance of identifying the object.

The following briefly describes the view planning process after a candidate identity is received: (I) potential viewing locations are checked for dynamic obstacles on the local cost map. (II) Views that are reachable by the robot are subjected to an *environmental analysis* to determine if any visual occlusions block the view of the candidate object. (III) Views which survive environmental analysis undergo *model analysis* to determine the amount of visible surface area visible from each view point.

Before describing the individual components of the view planning approach in detail, we motivate when *next best view planning* is initiated and why.

The Need for Local View Planning

In the context of object search tasks, a robot might seek objects in certain rooms (Kunze et al. 2012) or in proximity to other objects (Kunze, Doreswamy, and Hawes 2014). However, objects cannot always be recognized with high confidence from a first view. To verify the identity of an object the robot may have to take an additional view. We have identified the following situations in which one or more additional views would be beneficial:

1. When the recognition service provides a low confidence estimate of an object's identity. In this situation, another view would be required to confirm or deny this hypothesis. By moving the camera to a location where more of the object is visible there is a higher chance of obtaining a high confidence identification.
2. When a identification is returned with high confidence but not high enough to meet other task requirements; another view could lead to a higher confidence. This may be useful in identifying high priority items in a service environment, such as looking for the correct medicine.
3. In the event of more than one candidate identities being returned for the same object. This can occur when the visible features match more than one modelled object. Further views of the object can lead to disambiguation.

When any of these conditions are met, next view planning is the initiated. The components of the approach are described in the following sections.

Perception

Streams of images received from the robot's camera are processed to detect candidate objects. Bottom-up perception algorithms used - making estimates of object identities based on information available in the scene and no contextual information. Segmented sections of the scene are compared with a model database; any matches between the two are returned with a confidence measure as estimated object identities.

Online Aspect Graph Building

An object cannot be seen in its entirety from one viewpoint and different viewpoints prnt different features to a sensor. Aspect graphs are a method of simulating which features may be visible at different viewpoints around an object. A typical aspect graph for next view planning consists of a sphere of poses around a model and geometric analysis determines which features are visible from each point. In past work, aspect graphs have been computed offline (Cyr and Kimia 2004; Maver and Bajcsy 1993; Hutchinson and Kak 1989) which produces a set of coordinates for a sensor and an associated value to denote the number of visible features at each pose. This is used as a lookup table and requires knowledge of the object being viewed and its pose.

This project will instead build aspect graphs online which will be built in two stages. By moving online, we can account for real-time information about environmental obstructions and their effect on potential new poses. Aspect

graphs are generated for both *Environmental Analysis* and *Model Analysis*, which will be discussed next. The shape of aspect graphs in this project are governed by the robot's degrees of freedom. The robot used in this paper had 3-DOF (base movement and pan/tilt unit) and so graph nodes were arranged in a disc surrounding the candidate object.

Environmental Analysis After receiving a candidate identity we need to decide which available pose offers the next best view. The first step is to determine if any part of the environment lies between the sensor and the candidate object, thus creating a visual occlusion. To achieve this, a snapshot of the current environment is taken and converted into a volumetric representation. From various points around the candidate object the sensor (oriented towards the object) is modelled. Within this model, any part of the sensor's field of view which does not reach the estimated position of the object is discarded. This leaves a circle of poses around the object, each containing its respective remaining field of view which allows unobstructed line-of-sight to the object. These remaining parts of the field of view are then carried forward to model analysis.

Model Analysis Surviving sections of the modelled camera at each pose do not necessarily represent the visibility of the object at that location. In order to establish which view provides the most information about the object we use the model of the object from the object database along with surviving sections of the modelled cameras at each pose from the previous step.

The object database contains a manipulable model of the target object. This model is rotated to match the estimated pose of the candidate object. From here we again simulate each camera pose from the previous step, using only the surviving fields of view. Each modelled camera at each pose will be oriented towards the object model; the proportion of the field of view of each camera which is filled by the object is then saved and represents the visibility of the object from each pose around it.

Next Best View Selection

After environmental and model analysis, each pose is matched with a score which represents its visual coverage of the candidate object. The pose with the highest score is determined to be the next best view; this is sent to the robot's navigation component. Once at the new location the robot will either accept or reject the initial identity estimate or begin to determine another next best view if the first did not lead to recognition.

4 Implementation

In this section we describe how the view planning approach is integrated with the data structures and algorithms of the robot's perception and control components. Figure 3 provides an overview of the different components and explains the view planning process step-by-step.

Object Recognition

In this work, we build on a state-of-the-art object modelling and object recognition framework (Aldoma et al. 2013; Prankl et al. 2015). Our implementation is based on ROS[1], a message based operating system used for robot control. For object instance recognition, we use an adapted version of a ROS-based recognition service of the above mentioned framework. The service takes an RGB point cloud as input and returns one of following:

(1) list of candidate object hypotheses identities In case an object's identity cannot be verified a list of hypotheses is returned. The hypotheses include the object's potential identity and a pose estimate in the form of a 4×4 transformation matrix which aligns the object to the point of view of the camera.

(2) a verified object identity If an object is identified with high confidence, the object's identity, pose and the confidence level are returned.

In Figure 3, the input to the object recognition service, the service itself, and a visualisation of the output is depicted in Step A, B, and C respectively. In this case, the output is a hypothesis for a candidate object identity (here: a book). The object recognition service is used after moving to the next best view (see Step H, I, and J) this time, the outcome is a verified identification of the object.

View Generation

After a candidate identity is received, a series of poses are generated around it. These poses are generated in two uniformly distributed rings of robot poses around the estimated location of the object, each oriented directly towards the location of the object.

The accessibility of each of these poses is assessed by collision checking the views using the local cost map. Any views in which the robot would collide with environmental obstructions are discarded. This is seen in Figure 3 (Step D), where views that make contact with the supporting surface of the object are eliminated. The remaining views are carried forward to be assessed for visual occlusions.

Environmental Analysis

After a set of accessible views has been generated and tested, environmental analysis is performed on the initial point cloud to assess whether any regions of the environment might occlude the view of the object. In order to do this, the current point cloud is converted into an octree representation (Figure 3 Step E) (using Octomap (Hornung et al. 2013)) an RGB-D camera is then modelled at every view. The sections of each modelled camera that do not make contact with the bounding box of the candidate object are eliminated from future analysis. The sections of each camera model that allows an unobstructed of the candidate object are then carried forward to the next stage, all other sections are discarded.

Model Analysis

Figure 3 (Step F), shows the model analysis step. By completing the previous two steps, the amount of possible view locations and camera fields of view have been reduced. Each

[1] `http://www.ros.org/`

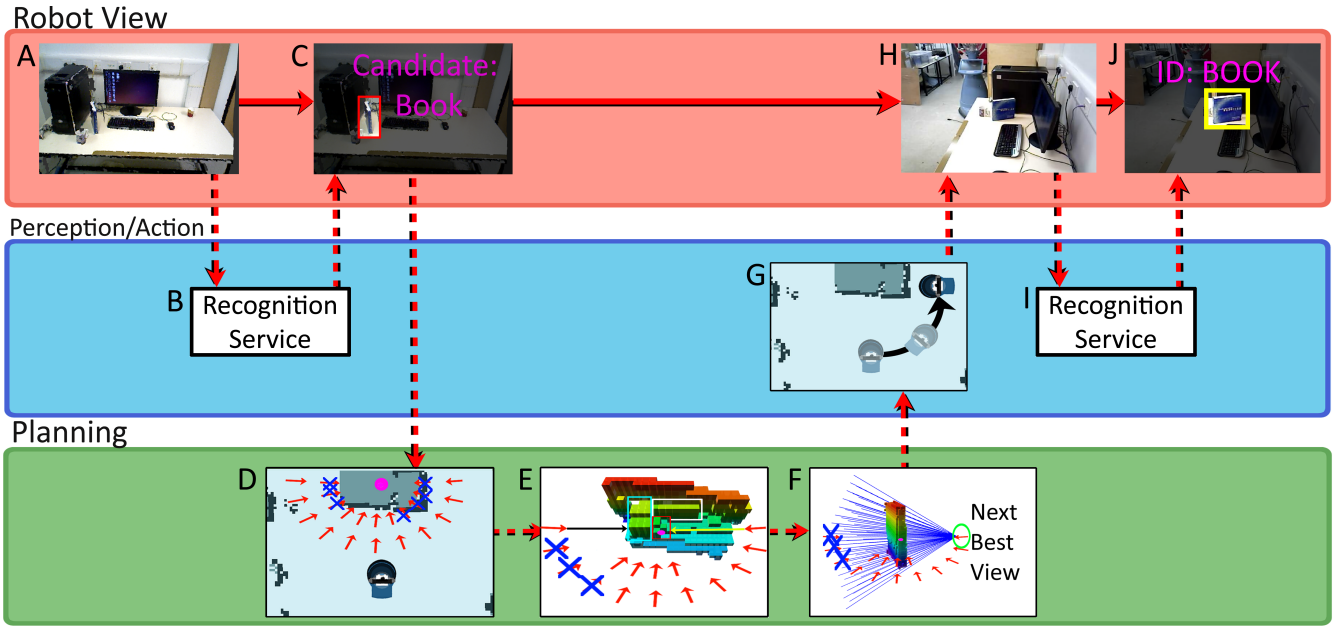

Figure 3: The next best view planning process step-by-step: **A:** Receive initial view. **B:** Recognition service processes point cloud. **C:** Candidate object and pose identified. **D:** Collision detection around object location. **E:** Environmental analysis for occlusions (candidate object in pink, occlusions in white and blue). **F:** Model analysis with remaining rays. **G:** Move to best view pose. **H:** New point cloud sent to recognition service and object recognised. **I:** Point cloud from second view is sent to recognition service. **J:** Visualisation of recognition service correctly and fully recognising the target object.

surviving part of the modelled camera is simulated and directed towards a model of the candidate object. The remaining sections of each camera are simulated using ray-casting; this computes a line from the origin of the camera out in the direction of the field of view. If the line makes contact with the model it is considered that the part of the object with which it made contact would be visible to the camera from that viewpoint. After the camera is modelled at each view, we are left with a measure of the amount of object visible at each location. The view which enables the highest visibility is then considered the next best view.

Multi-object Disambiguation Note, if the recognition service returns more than one candidate hypothesis, aspect graph building is performed for every object, each with its own pose transformation. After this, each is analysed to find the best compromise view—one which gives the best chance for recognition of each model. A solution to this is, after calculating the sum of visible surface area for each view, the difference between them is then subtracted. This ensures no high surface area visibility from one pose dominates over a low score for the same pose on another object.

View Execution: Navigation & Recognition

When the next best view has been found, the pose associated with it is sent to the robot's navigational packages, which manoeuvres the robot to that view (Figure 3, Step G). In this case, the pose consists of movement by the base of the robot and angular movement by pan/tilt unit to centre the camera on the object's location. Once the goal is reached, input resumes to the recognition service (Figure 3, Step I); The response of the recognition service will denote different things:

1. **Verification**: If a high confidence estimate of the object's identity is returned, the next view is considered successful and the object considered identified.
2. **No candidate**: If the recognition service returns no high confidence identity after the next view it should either be considered successful in dispelling a false hypothesis from the first view (true negative) or unsuccessful, as it has lowered the amount of information available to the recognition service to deny further identification.

Depending on the result of the movement: the object search will end, another view is taken or the candidate will be discarded and the search continued. This gives a detailed overview of the process this next view selection undergoes in order to produce a next best view pose. The components of this system will now be assessed and results examined.

5 Experimentation and Evaluation

Experiments were conducted to test the capabilities of this next best view planner. All experiments were carried out on a Scitos G5 robot equipped with pan/tilt unit. In each experiment the robot was located in an open area with few obstacles. The centre of this area contained a tall, thin plinth, the plateau of which was just below the robot's camera height. Test objects and obstacles were placed on this supporting table and the robot was able to move around to reach a new

Table 1: Results of trials when presented with a single object and no obstruction (Experiment 1)

Result	# Trials	Percentage %
Moved. Recognised	33	82.5%
Moved. No Recognition	4	10.0%
No Initial Candidate	3	7.5%
Total	40	100.0%

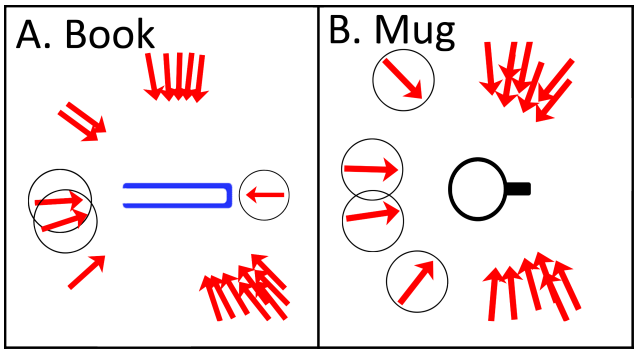

Figure 4: Graphical representation of selected next best view poses for a book and mug in experiment 1. Poses surrounded by black rings did not lead to recognition.

view. Each experiment used its own specific of target objects. Details of the set-up of individual experiments are given below.

Experiment 1: Non-obstructed Next View Selection

Set-up The primary function of this planner is to take in hypotheses about potential objects and select the best pose to enable their accurate confirmation or rejection. This function was tested during this experiment. No obstacles or occlusions were used apart from the object's supporting plane. 40 trials were carried out in total, 20 each on two different objects: a large book and a standard mug. In all cases the robot was initially positioned with a view of a the object which gave a low chance of recognition. Success in these trials were defined by the ability of the recognition service to make a high confidence identification of the target object after the next best view planner selected a new pose and the robot had moved. In each trial up to two next best view locations were permitted.

Results Table 1 shows the results of this experiment. The table shows a successful verification rate of 82.5%, meaning that in most cases the planner was able to take an uncertain hypothesis and verify it through moving to a new location. In 10% of trials, two next view poses were taken, but no recognition was achieved.

Discussion Failure to provide a verified hypothesis in 10% of cases can be attributed to the pose estimation provided by the recognition service. If the view of the candidate in the initial view is very low quality, pose estimation can be inaccurate; this leads to poorly aligned candidate poses and thus inaccurate movement. In some cases this was be compensated for: of the 33 successful trials, 7 required two views before high confidence recognition was achieved. The initial inaccurate pose estimation led to poor next view selection; however, the new view presented better quality information about the object to a point where the pose estimation improved and the subsequent view resulted in recognition. The nature of taking pose estimates from low confidence identity hypotheses does inherently hold the risk of inaccurate pose estimation, so this is to be expected.

Circled poses in figure 4 show the final positions of 10% of the trials in which no recognition could be made after two views, which can be attributed to a succession of inaccurate pose estimations. The red pose arrows in figure 4 tend to cluster in areas with a view of large surface area of the object. When using the book this was a view from which both the spine and cover were visible and where the handle and body were visible on the mug. This demonstrates that given a reasonably accurate initial pose estimation of the object, the next best view planner is able to locate the view which presents one of the largest faces of the object to the camera and that doing this leads to reliable recognition. This shows that aspect graphs can be computed online to lead to high recognition accuracy and do not require excessive analysis.

Experiment 2: Confirmation Views

Set-up After making a high confidence identification of an object, it may be useful to take another from a different position, as two high confidence identifications provide more certainty than one. This experiment followed the same procedure as experiment 1, except for the starting position of the robot which was positioned to allow a high confidence identification. Success in this experiment was measured firstly on whether the next view led to another high confidence recognition and secondly on how much identification confidence increased from the first pose to the next.

Results Table 2 shows the descriptive statistics for the second experiment. On average, moving from one verified viewpoint to another resulted in an increase in confidence in 60% of cases, with an average increase in confidence of 3.65%. The size of increase/decrease fluctuations was quite large, but the average change in confidence shows an upward trend over these 20 trials.

Discussion Increases in confidence were largest when the start position did not align with the largest face of the object; next view selection was then able to find the largest face in the subsequent view. On the contrary, in situations where confidence decreased, the initial view coincided with the largest face of the object so the subsequent view moved away from this largest face. However, instances of reduced confidence should are not necessarily failures, as this still provided two high confidence identifications of the object. This block of experiments shows the planner is able to select a view to verify an object hypotheses with a good level of reliability.

Table 2: Descriptive statistics of the results of the confirmation views experiment (Experiment 2)

Measure	Result
Attempts	20
Found New Confirmation	20
Increased	12 (60%)
Decreased	8 (40%)
Average Change	+ 3.65%
Standard Deviation	6.14%
Largest Decrease	-9.67%
Largest Increase	+15.02%
Mean Initial Confidence	62.35%
Mean Final Confidence	64.53%

Experiment 3: Visually Obstructed Views

Set-up Environmental obstacles potentially act as visual occlusions when selecting a view. Next view planning must be able to recognise these in one view and assess their impact on the next. Without this ability the next view planner can select a view that would theoretically lead to a recognition confidence of 100%, but but this view may be blocked by another object, thus the actual recognition confidence is closer to 0%. To test this functionality the robot is presented with a scene which contained a target object and potential occlusion to block all or large parts of the object. Over 20 trials the robot was provided with a initial view of the object which allowed low confidence recognition. Each trial would be deemed a success if the object was recognised with a high level of confidence after the first movement.

Results Results showed that in 17 of the 20 cases (85%), the planner was able to select a pose which both avoided the occlusion and lead to recognition. Of the remaining trials, the selected pose allowed a view of the target object but the view was incomplete - being partially occluded by the obstacle.

Discussion In most cases the planner was able to account for an environmental occlusion and choose a best view pose that avoided it. In the remainder of cases the next view pose lead to a partially obscured view of the object. This is due to occlusion modelling being based on the information gained from a single frame during the initial view. If the potentially occluding object is occluded by the target object then the environmental analysis is detrimentally affected by partial observability of the obstacle.

Experiment 4: Ambiguous Objects

Set-up A new view of an object can be taken to differentiate between objects that share similar features, which was the basis of this experiment. When presented with a view of an object that could belong to a target object or unrelated object, it would be best to disambiguate this view to decide if the target object has been found. To test this, two custom objects were used. Figure 5 shows the two modelled objects share a face from which the are almost indistinguishable, but are structurally different from other angles. In this experiment the robot will be placed with a view of the common face of these two objects and is expected to decide on a pose which increases the difference between the number of clusters recognised from each object. In all experiments the cuboid object(figure 5) was the target object. The robot was required to select a new pose to increase the strength of one correct hypothesis and decrease that of the incorrect one in 20 trials. The number of visible features that the recognition service matched to each candidate identity is a measure of the strength of that hypothesis.

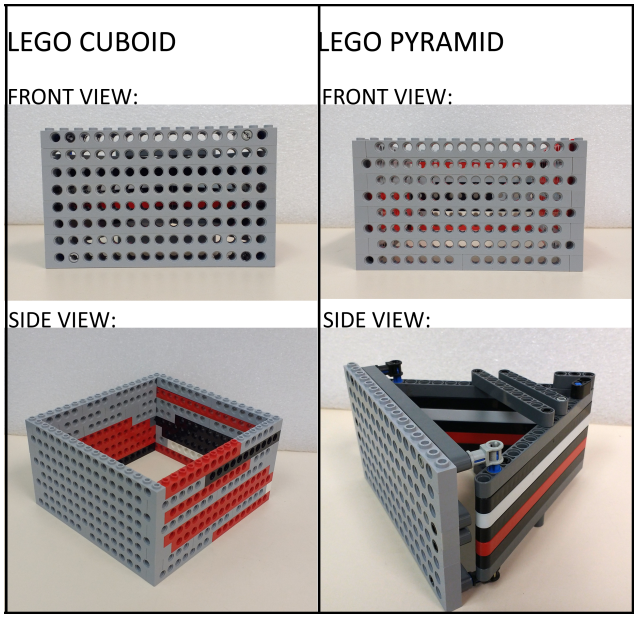

Figure 5: Objects which share common features. They appear identical from a front view, but are distinguishable from other angles (Experiment 4).

Results Figure 6 shows that when presented with a scene in which one of two objects is present, the next best view planner can strengthen the hypothesis of the correct object and weaken that of the incorrect identity. Of the clusters recognised in the initial view, an average of 367.4 belonged to the correct object and 210.65 to the incorrect object. After one movement based on the selection of the next best view selection, the average available clusters for the correct object rose to 456.4 and 152.8 for the incorrect object.

Discussion This shows that selecting one new view can increase the differentiation between two ambiguous objects and lead to a reliable identification. This suggests much simpler and less computationally expensive methods of hypothesis differentiation that in Okamoto (Okamoto, Milanova, and Bueker 1998) and that by only taking one view rather than a constant stream, the process is also much simpler.

6 Operation Time

The time for making one movement: from receiving a candidate identity and pose estimation to arriving at the next location is 1 minute 44 seconds. For two cycles the completion

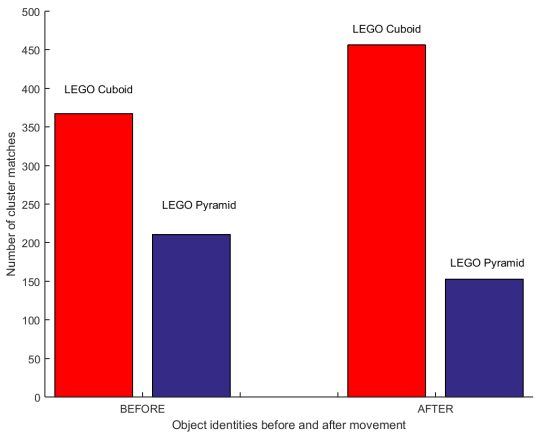

Figure 6: Results for experiment 4. Percentage of available clusters for two ambiguous objects before and after movement.

time jumps to 4 minutes 58. This is due to the large amount of data needed to compute each camera model at each pose. This is clearly a number that needs to be reduced and can be a subject for future work.

General Discussion

Experimental results show this is a strong next view algorithm for object recognition that can work reliably in cluttered, unpredictable environments.

In order to improve this solution further some areas can be enhanced to make it more robust and generalisable. Potential next view locations are currently set at a fixed distance from the candidate object; this can be a hindrance in certain topological layouts. Developing adaptable next view locations which, rather than test which of a fixed set of locations are in free space and therefore available, the potential locations should be instead generated in exclusively free space and then environmental and model analysis take place from there.

In adopting a greedy approach, this work selected only poses with the highest visible portion of the object; future work should focus on including a cost function to form a utility between movement distance and amount of the model which is visible.

7 Summary & Conclusion

A summary of the contributions of this project are shown below. The results shown in the previous section will be presented in support of these.

In summary, the aims of this study were to show that:

1. a method for analysing potential views using online aspect graphs by taking both into account: (1) occlusions, and (2) the visibility of features based on learned object models.
2. a next best view planner that selects a view based on the method above and an executive that accounts for dynamic obstacles during execution
3. a set of experiments which demonstrates (a) how robots can disambiguate objects which share a similar set of features, and (b) how the performance of object recognition can be improved by taking multiple views.

Results of experiment 1 show that the online aspect graphs analysis is able to verify candidates put forward by the recognition service with an accuracy of 82.5%; however this also showed that the next best view planner is also highly dependent on the accuracy of the pose estimation provided to it. Experimentation also showed that dynamic collision detection was able to eliminate unavailable poses; removing around 17 of 38 poses on average during every trial. We can also show that in 85% of cases, the planner was able to avoid visual occlusions in the environment, but this was heavily dependent on the visibility of the obstruction during the initial view. This was dually confirmed when two identical starting poses in experiment 1 & 3 both arrived at different final poses, as the best view when no occlusions are present was unavailable when clutter was introduced. Finally, we showed that the planner was able to decrease ambiguity in objects that have identical faces.

To achieve these aims we used online aspect graph building and octree based visual occlusion detection. These were new ways of approaching next best view planning and showed that online aspect graph analysis for view planning was possible and unlike offline examples (Hutchinson and Kak 1989) could account for full or partial occlusions in the environment and thus avoid these when planning the next best view. Also, online aspect graph building allows models to be added during autonomous patrol and immediately available for recognition, whereas offline building would require a period of down-time. By decreasing the ambiguity between two identical looking objects, we showed that expensive image streaming methods (Okamoto, Milanova, and Bueker 1998) are not necessary and a more intelligent approach that fixed angle movements (Wixson 1994) was possible, using no more than two views, with an identification rate of 82.5%.

The work presented in this paper was successful in its aims. From online aspect graph building and collision detection to camera modelling and near real time occlusion analysis, the way this planner was designed allows it to be plugged into any robot using any model based recognition system, meaning this planner is available for a variety of robots that conduct object search in cluttered environments.

Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No 600623, STRANDS.

References

Aldoma, A.; Tombari, F.; Prankl, J.; Richtsfeld, A.; Di Stefano, L.; and Vincze, M. 2013. Multimodal cue integration through hypotheses verification for rgb-d object recog-

nition and 6dof pose estimation. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, 2104–2111. IEEE.

Bajcsy, R. 1988. Active perception. *Proceedings of the IEEE* 76(8):966–1005.

Callari, F. G., and Ferrie, F. P. 2001. Active object recognition: Looking for differences. *International Journal of Computer Vision* 43(3):189–204.

Connelly. 1985. The Determination of next best view. *IEEE* 432–435.

Cyr, C. M., and Kimia, B. B. 2004. A similarity-based aspect-graph approach to 3d object recognition. *International Journal of Computer Vision* 57(1):5–22.

Hanheide, M.; Hawes, N.; Wyatt, J.; Göbelbecker, M.; Brenner, M.; Sjöö, K.; Aydemir, A.; Jensfelt, P.; Zender, H.; and Kruijff, G.-J. 2010. A framework for goal generation and management. In *Proceedings of the AAAI workshop on goal-directed autonomy*.

Hornung, A.; Wurm, K. M.; Bennewitz, M.; Stachniss, C.; and Burgard, W. 2013. Octomap: An efficient probabilistic 3d mapping framework based on octrees. *Autonomous Robots* 34(3):189–206.

Hutchinson, S. a., and Kak, A. C. 1989. Planning sensing strategies in a robot work cell with multi-sensor capabilities. *IEEE Transactions on Robotics and Automation* 5(6):765–783.

Kunze, L.; Beetz, M.; Saito, M.; Azuma, H.; Okada, K.; and Inaba, M. 2012. Searching objects in large-scale indoor environments: A decision-theoretic approach. In *2012 IEEE International Conference on Robotics and Automation (ICRA)*, 4385–4390. IEEE.

Kunze, L.; Burbridge, C.; Alberti, M.; Thippur, A.; Folkesson, J.; Jensfelt, P.; and Hawes, N. 2014. Combining top-down spatial reasoning and bottom-up object class recognition for scene understanding. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2910–2915. IEEE.

Kunze, L.; Doreswamy, K. K.; and Hawes, N. 2014. Using qualitative spatial relations for indirect object search. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 163–168. IEEE.

Maver, J., and Bajcsy, R. 1993. Occlusions as a Guide for Planning the Next View. *IEEE transactions on pattern analysis and machine intelligence* 15(5):417–433.

Okamoto, J.; Milanova, M.; and Bueker, U. 1998. Active perception system for recognition of 3D objects in image sequences. *Advanced Motion Control, 1998. AMC '98-Coimbra., 1998 5th International Workshop on* 700–705.

Prankl, J.; Aldoma, A.; Svejda, A.; and Vincze, M. 2015. Rgb-d object modelling for object recognition and tracking. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, 96–103. IEEE.

Roberts, D., and Marshall, A. D. 1998. Viewpoint selection for complete surface coverage of three dimensional objects. In *BMVC*, 1–11.

Stampfer, D.; Lutz, M.; and Schlegel, C. 2012. Information driven sensor placement for robust active object recognition based on multiple views. In *2012 IEEE International Conference on Technologies for Practical Robot Applications (TePRA)*, 133–138. IEEE.

Vasquez-Gomez, J. I.; Sucar, L. E.; and Murrieta-Cid, R. 2014. View planning for 3D object reconstruction with a mobile manipulator robot. *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems* (Iros):4227–4233.

Vázquez, P.-P.; Feixas, M.; Sbert, M.; and Heidrich, W. 2001. Viewpoint selection using viewpoint entropy. In *VMV*, volume 1, 273–280.

Wixson, L. 1994. Viewpoint selection for visual search. *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on* 800–805.