Active Object Recognition using Vocabulary Trees

Natasha Govender *
MIAS (CSIR)
South Africa
ngovender@csir.co.za

Jonathan Claassens
MIAS (CSIR)
South Africa
jclaassens@csir.co.za

Fred Nicolls
University of Cape Town
South Africa
fred.nicolls@uct.ac.za

Jonathan Warrell
University of Oxford Brookes
United Kingdom
jwarrell@brookes.ac.uk

Abstract

For mobile robots to perform certain tasks in human environments, fast and accurate object classification is essential. Actively exploring objects by changing viewpoints promises an increase in the accuracy of object classification. This paper presents an efficient feature-based active vision system for the recognition and verification of objects that are occluded, appear in cluttered scenes and may be visually similar to other objects present. This system is designed using a selector-observer framework where the selector is responsible for the automatic selection of the next best viewpoint and a Bayesian 'observer' updates the belief hypothesis and provides feedback. A new method for automatically selecting the 'next best viewpoint' is presented using vocabulary trees. It is used to calculate a weighting for each feature based on its perceived uniqueness, allowing the system to select the viewpoint with the greatest number of 'unique' features. The process is sped-up as new images are only captured at the 'next best viewpoint' and processed when the belief hypothesis of an object is below some pre-defined threshold. The system also provides a certainty measure for the objects identity. This system out performs randomly selecting a viewpoint as it processes far fewer viewpoints to recognise and verify objects in a scene.

1. Introduction

Reliable object classification is essential for robots to perform tasks in human environments. Many recognition systems operate on single views [1] [2]. In real-world situations a single viewpoint may be of poor quality and simply not contain sufficient information to reliably recognise or verify the object's identity unambiguously. This is especially true if they are occluded or appear in cluttered environments. There may also be a great variety of relevant objects with significant similarity. In such cases multiple viewpoints are necessary for recognition [3] [4]. Active vision allows a robot to actively search an environment to obtain more informative views to increase the accuracy of object identification and verification.

The two focus areas of active object recognition are: selecting the next best viewpoint and integration of relevant information. For selecting the next best viewpoint many systems simply use active vision to select the sequence in which a set of pre-captured images should be processed for recognition [5]. Often these sequences are of fixed length and optimization of appraisal time is not considered. To the authors' knowledge, no systems explicitly consider the possibility of occlusion or extremely cluttered environments.

We propose a unique framework for feature-based active object recognition and verification, which is comprised of an automatic viewpoint selector and an independent observer. The automatic viewpoint selector uses a vocabulary tree structure [6] to weight the uniqueness of each feature in a viewpoint. Every viewpoint of all objects in the database are then given a value which is obtained by summing the uniqueness measure of all its features. The higher the value, the more unique the viewpoint. This quantity is then used to select the subsequent view. The vocabulary tree also facilitates quick matching and provides a method to discretize the feature space to reduce feature dimensionality when considered in the observer component. Following the approach used in [7], the observer component updates an object belief probability with current view information in a recursive Bayesian manner using a prior determined from previous views. These two components are designed to be independent of each other. The advantage of this framework is that the algorithm for the next viewpoint selection can be altered or completely rewritten and it would not affect the observer component and vice-versa.

Interest points, which have the advantage that the representation is more robust to occlusions, clutter and noise,

*This work was funded by the Mobile Intelligent Autonomous Systems (MIAS) group in the Council for Scientific and Industrial Research (CSIR) South Africa.

have successfully been used in 3D object recognition [8] [9] [10]. Our system uses the Scale Invariant Feature Transform (SIFT) [11] detector and descriptor to extract relevant object features. SIFT is robust to changes in illumination and affine transformation. The structure of our system is, however, not SIFT dependent and thus any other descriptor or detector can be used for feature extraction.

The structure of the paper is as follows. Section 2 discusses related work and Section 3 elaborates on how the datasets for the experiment were collected. A complete description of the system's architecture is presented in Sections 4 and 5. Sections 6 and 7 present the experimental results and conclusions. Finally, possible future work is discussed in Section 8.

2. Related Work

Image processing methods used to create object models for classification include appearance-based methods [12], aspect graphs [13] [14] [15] [16], histogram of gradients [5] and neural networks [17]. Following [18] we use SIFT to model objects, which provides robustness to affine transformations and variable illumination.

Using multiple views for object recognition improves the accuracy of the recognition system [15]. The focus of active object recognition and verification is how to select the "next best viewpoint" which will provide the most amount of information to complete the task as quickly and as accurately as possible. Most active object recognition systems are based on selecting viewpoints that will minimise ambiguity using Shanon entropy [17] or Dempster-Shafner theory [13], minimise a weighted error [5] or maximise a defined activation function [18].

In our system, views are selected based on promised abundance and uniqueness of features. In contrast to existing approaches we rely on an efficient bag-of-words approach to organize the training feature database. This data structure is called a vocabulary tree and provides a measure of feature uniqueness per object and discrimination potential. The system also provides a confidence/certainty measure for the objects identity.

Vocabulary trees have been traditionally used in object recognition and Simultaneous Localization and Mapping (SLAM) approaches for matching similiar images and for loop closure [19]. Our application of the data structure differs in that we use it to calculate weightings for features to determine the next best viewpoint. We also use it to generate statistics to update the object belief. Following [12] [18] our system relies on a Bayesian framework for updating a belief function.

With the exception of [12], many of these systems use a pre-determined number of images and merely use active vision to select the sequence in which they should be used. Our system is different: it only captures a new image when required and thus optimises the number of views needed for reliable recognition or verification.

Figure 1. An example of two different objects used in the database that share similar views

When classifying objects, all of the above systems, except for [18], consider scenes with a single object. In [18] the target object is placed in the centre of the image with no occlusions or clutter. Our system recognizes and verifies objects which not only occur in cluttered environments but are also occluded. Few systems in the literature consider datasets with objects that share many visual similarities. Exceptions include [12] [20]. Our database contains a number of visually similar objects which can only be differentiated by appraising specific viewpoints.

3. Data Collection

The training database used consists of twenty everyday objects. This is much larger than other databases used for active vision experiments. To assemble the training set for the vocabulary tree, images were captured every 20 degrees against a plain background on a turntable using a Prosilica GE1900C camera. Verifying or recognising objects tends to become more complicated if two or more objects have views in common with respect to a feature set. These types of objects may be distinguished only through a sequence of images which the viewpoint selection algorithm is required to determine. For this reason, objects that share a number of similar views were included in the dataset, as shown in Figure 1. The database used is available on request.

For the test set, the objects used in the training data were captured at every 20 degrees in a cluttered environment with significant occlusion. In all the presented experiments, images are captured around the y-axis, which represents 1 degree-of-freedom (DoF). This is not a limitation of our proposed system. Our viewpoint selection system can easily be applied to several degrees-of-freedom with a modest increase in required computation.

4. Active Viewpoint Selection

The aim of the automatic view selection algorithm is to select the 'next best viewpoint' for object recognition and

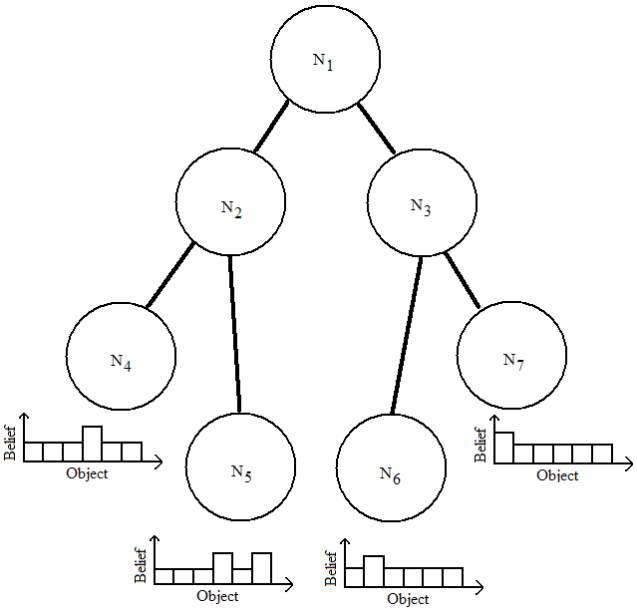

Figure 2. A schematic of the modified vocabulary tree

verification i.e. the viewpoint which will provide the most amount of useful information to optimally complete the process. The proposed scheme uses a vocabulary tree [6]. This structure is typically used in bag-of-words object recognition and visual loop closure approaches as an efficient alternative to Sivic and Zisserman's Video Google [21]. The idea is to gather all features in the training set, cluster them hierarchically and calculate a uniqueness weighting for each feature. The vocabulary tree data structure was designed for large volumes of data and thus will scale easily if more objects are to be added to the database.

The vocabulary tree is constructed using hierarchical k-means clustering where similar features are clustered together. k defines the number of children of each node of the tree. Initially, for the root of the tree, all the training data is grouped into k clusters. The training data is then used to construct k groups, where each group consists of SIFT descriptors closest to a particular cluster centre. This process is recursively applied to each group up to some depth D. This process is illustrated in Figure 2.

For each node in the tree a TFIDF-like (Term Frequency Inverse Document Frequency) metric is calculated to capture the node's uniqueness:

$$w_i = \ln \frac{M}{M_i} \tag{1}$$

where *M* is the total number of images in the database and M_i is the number images in the database with at least one feature that passes through node i.

Using this quantity, a feature's uniqueness may be calculated. This is done in the following way. The feature's

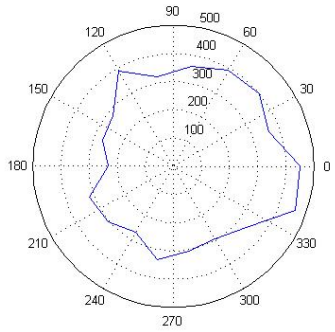

Figure 3. Viewpoint weightings for a spice bottle object in the database.

path through the vocabulary tree is determined by evaluating the closest cluster centers at each level. A measure of uniqueness is given by the sum of all the TFIDF numbers, or weights, of the nodes it passes through. The higher the weighting, the more unique the feature. The uniqueness of the viewpoint may then be given by summing these totals for all the SIFT features extracted from that viewpoint. We term this metric the viewpoint weighting. This calculation is performed for every viewpoint in the dataset.

It is important to note that SIFT features detected on the background will not negatively effect the weighting since all images were captured using the same background and their uniqueness weighting will be extremely low. Figure 3 is an example polar plot of viewpoint weightings for a spice bottle object in the database. The polar plot indicates that the most distinctive viewpoints (highest weightings) are at 340 degrees and at 0 degrees and the most indistinguishable (lowest weighting) viewpoint is at 180 degrees.

The view selection component of the proposed active vision system functions as follows. For object verification, an input image is provided to the system with the necessary object hypothesis. The input image is matched using standard SIFT matching and a hough transform to the hypothesized object's training images to determine the closest training image which provides the initial pose estimate of the object. Relative to the pose estimate the view selection component selects a view that it has not previously visited and has the largest uniqueness weighting for that object.

For object recognition, no object hypothesis is given to the system. The criteria for selecting the next best viewpoint is based on the viewpoint which has the highest combined weighting across all objects in the database and has not been previously visited. In the experiments section, we will show that both these selection methods significantly outperform randomly selecting the next viewpoint.

5. Observer Component

In our framework, the observer component updates the system's object belief. The vocabulary tree used in the view selection component is altered to store the statistics necessary for the observer component. This is explained in the next subsection.

5.1. Calculation of feature statistics

The vocabulary tree is built by the view selection component using features from all dataset images. At each leaf node a discrete density function which represents the likelihood of the feature appearing at least once given a certain object, is added. These densities are represented as $P(N|O)$ where O is an integer. A representation of the modified tree in shown in Figure 2.

The discrete density elements are determined as follows. If any feature from an object's training set, when passed through the vocabulary tree, reaches a leaf node N_5, say, then the corresponding element of $P(N_5|O)$ is assigned p_o. Elements that are not reached by this object's training set are assigned p_{no}. Constants p_o and p_{no} are assigned in a 'soft' manner, i.e. no elements are assigned zero. This avoids over committed densities. In our experiments, $p_o = 2$ and $p_{no} = 1$ appeared to work well. Once the leaf node densities are populated they are normalized so that all elements sum to one.

5.2. Pipeline

With the tree constructed, the observer component will proceed in the following manner to update its belief:

1. **Initialization** - A uniform prior is assumed over all object hypothesis:

$$P(O) = 1/N \tag{2}$$

   where N is the number of objects (this initialization is used for both verification and recognition).

2. **Image processing** - When given a new viewpoint which is provided by the viewpoint selection component the observer component proceeds to extract SIFT features from the image. These features are then matched using Lowe's method [11] to the training image of the hypothesized object which best matches the given image. A Hough transform and voting scheme is used, as described in [8], to select only those features that agree with the training image feature geometry. This additional filter removes spurious matches.

3. **Fusion** - Each feature provided by the previous step is cascaded through the vocabulary tree by selecting children with the closest centroids. The leaf node associated with each feature contains a density as described above. Every feature's density is fused recursively with the prior using

$$P(O|N) = \frac{P(N|O)P(O)}{P(N)} \tag{3}$$

   where $P(N|O)$ is the density at the leaf node. $P(N)$ is merely a normalizing coefficient. All nodes are considered independently.

4. **Stopping criteria** - If the posterior belief has a probability of greater than some threshold, ϵ, for the hypothesized object the process terminates. We may also stop if a maximum number of viewpoints have been reached, otherwise we take the resulting posterior belief, request a new view from the selector component and return to step 2.

The pipeline steps described above are used for both object verification and for object recognition.

6. Experiments

6.1. Verification

The main purpose of an *active* object recognition or verification system is to improve the processing time and accuracy required to determine an object's identity. In addition to this, our system also provides a measurement for how certain the system is of an object's identity. Test images were captured with the relevant objects in occluded locations in cluttered environments as shown in Figure 4.

An initial test image is presented to the system at an arbitrary pose. The belief probability is updated at each subsequent view that is processed. The system retrieves the next best viewpoint until a confidence or belief probability of 80% is reached. In accordance with previous state-of-the-art active object recognition systems [18] [12] [7] [22] [20] [17], we compare our results to randomly selecting the next viewpoint. When randomly selecting the next best viewpoint, the experiment was conducted ten times and the average number of views for each object was taken. Both methods correctly verify all objects. We are, however, more interested in the number of views required to correctly verify an object as this greatly influences the processing time of the system. Table 1 displays the number of views required by each method.

Table I describes the number of viewpoints required for each object in the database to reach a confidence level of 80% for verification. For each of the twenty objects, our method requires fewer viewpoints, in some cases significantly so, to reach a confidence of 80%. This indicates that our method is selecting more infomative viewpoints which can significantly decrease the processing time of the system.

Figure 4. Examples of test images of occluded objects appearing in a cluttered environment which are to be verified

Table 1. Number of Views: Object Verification

	Cereal Box	Battery	Can1	Can2	Curry1	Curry2	Elephant	Handbag1	Jewelry 1	Jewelry 2
Our method	1	1	3	4	2	3	1	2	16	15
Random	1	1	6.8	7.5	4.4	7.5	1	2.3	18	16

	Bottle	MrMin	Salad Bottle	Sauce1	Sauce2	Spice1	Spice2	Can1	Can2	Can3	Average
Our method	9	1	15	3	3	6	16	5	5	11	6.1
Random	14.4	1.5	18	5.8	7.1	6.2	18	7.8	7.6	16.3	8.41

The difference in information provided by the varying choice of viewpoints can be shown. Figure 5 displays the increase in belief after each view for the 'Curry 1' object. We can see that even after the second view our method has a much higher belief than randomly selecting a viewpoint for both verification and recognition. After four views in the case of verification and five views for recognition, our method reaches a confidence level of 1.

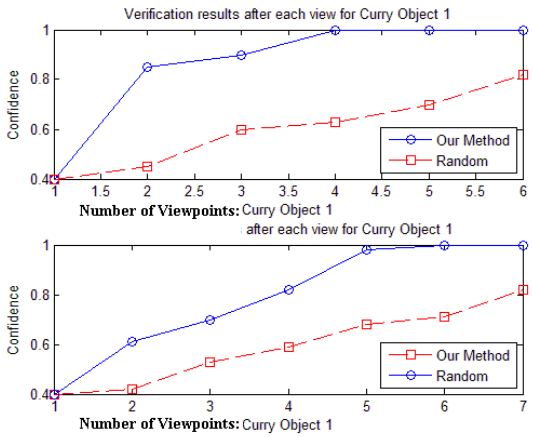

Figure 5. Confidence values after each view for verification and recognition

6.2. Recognition

The system was then tasked to *recognise* occluded objects in cluttered scenes. This differs from verification in that, the object's identity is not known to the system. It has to determine the identity based on which object has accumulated the greatest belief probability given the current database. The system retrieves the next best viewpoint until a confidence or belief probability of 80% is reached for any of the objects in the database. The next best viewpoint is selected based on which viewpoint has the highest combined weighting over all objects. Both methods for selecting the next best viewpoint (our method and random selection) correctly recognize all objects. As mentioned before,

Table 2. Number of Views: Object Recognition

	Cereal Box	Battery	Can1	Can2	Curry1	Curry2	Elephant	Handbag1	Jewelry 1	Jewelry 2
Our method	1	1	5	10	4	7	2	3	16	15
Random	1	2	18	18	5.8	8.8	8.3	3.1	18	18

	Bottle	MrMin	Salad Bottle	Sauce1	Sauce2	Spice1	Spice2	Can1	Can2	Can3	Average
Our method	14	2	15	4	4	10	16	9	11	15	8.2
Random	16	2.1	18	10	7.1	11	18	17.5	13	18	11.58

the measure of interest is the number of viewpoints required to correctly identify an object. Table 2 displays the number of views required by each method.

Table II describes the number of viewpoints required for each object in the database to reach a confidence level of 80% for object recognition. Our method clearly out performs randomly selecting the next viewpoint. It requires fewer views for all objects to attain a confidence of greater than or equal to 80%. This leads to a significant decrease in processing time for recognising objects which are occluded in cluttered environments.

A number of methods have explored active object recognition previously [18] [13] [5] [12] [17], but using experimental set-ups not directly comparable to ours. We adapted [18] to run on our data, and found our performance to be comparable and better in a number of cases, but do not quote results since we did not try to optimize their performance on our problem. We will make our data and code available on request to facilitate future comparisons.

7. Conclusions

Our experiments show the successful use of active object exploration for 3D object verification and recognition for significantly occluded objects in extremely cluttered environments. The active vision system performs considerably better than randomly selecting the next viewpoint. The system also provides a measure of certainty for the object's identity.

We introduce a new framework for active object verification and recognition consisting of an selector and an observer component. The selector determines the next best viewpoint and the observer component updates the belief hypothesis and provides feedback. The observer component works independently from the selector and thus any exploration or manipulation of an object can occur without interfering with the observer component. This framework, which has proven to work efficiently, can be applied to any active vision task.

To select the next best viewpoint, features appearing in every viewpoint were weighted based on their uniqueness in the given dataset using a vocabulary tree. For verification, the viewpoint with the highest weighting for the object to be verified was then selected as the next view. In the case of object recognition, the viewpoint with the highest weighting over all objects was selected as the next viewpoint. Both these methods proved to be significantly better than randomly selecting the next viewpoint. Bayesian methods are used to update the belief hypothesis and provide feedback. The path of each matched feature in the test image was traced through the vocabulary tree and the statistics contained in the leaf node were used to update the belief hypothesis.

New images were only captured when the belief was below a pre-defined threshold. This reduces the computational time because only the minimal number of images will be processed to perform the task.

Our system uses test images where the object to be verified or recognized is significantly occluded and appears in a cluttered environment. Even with these difficulties, our system correctly verifies and recognizes all objects requiring fewer viewpoints than randomly selecting the next viewpoint, in some cases significantly so.

To summarize, we have developed an active 3D object recognition and verification framework which can be applied to any active vision task. The next viewpoint selection algorithm significantly outperforms randomly selecting the next viewpoint. Our system only captures a new image when necessary and successfully deals with occluded objects in cluttered environments which may be visually similar to other objects contained in the database. It also provides a measure of certainly of the object's identity.

8. Future Work

The robotic arm manipulator for which this system was designed, does not have a complete 360 degree range. To combat this problem we would like to create 3D models of the objects, so in the event that the next view cannot be executed by the arm, the object itself can be manipulated to achieve the desired viewpoint.

9. Acknowledgments

We would like to thank Deon Sabatta for the use of his vocabulary tree implementation.

References

[1] A. Zisserman, D. Forsyth, J. Mundy, C. Rothwell, J. Liu, and N. Pillow, "3d object recognition using invariance," in *Artificial Intelligence 78*, pp. 239–288, 1995.

[2] D. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision*, pp. 1150–1157, 1999.

[3] A. Collet, D. Berenson, S. Srinivasa, and D. Ferguson, "Object recognition and full pose registration from a single image for robotic manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2009.

[4] A. Selinger and R. Nelson, "Appearance-based object recognition using multiple views," in *Conference on Computer Vision and Pattern Recognition*, vol. 1, p. 905, 2001.

[5] Z. Jia, "Active view selection for object and pose recognition," in *International Conference on Computer Vision (ICCV) 3D Object Recognition Workshop*, 2009.

[6] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 5, 2006.

[7] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz, "Appearance-based active object recognition," in *Image and Vision Computing 18*, pp. 715–727, 2000.

[8] D. Lowe, "Local feature view clustering for 3d object recognition," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 682–688, 2001.

[9] V. Ferrari, T. Tuytelaars, and L. Gool, "Simultaneous object recognition and segmentation from single or multiple views," in *International Journal of Computer Vision*, vol. 67, pp. 159–188, 2006.

[10] P. Moreels and P. Perona, "Evaluation of feature detectors and descriptors based on 3d objects," in *International Journal of Computer Vision*, vol. 73, pp. 263–284, 2007.

[11] D. Lowe, "Distinctive image features from scale invariant keypoints," in *International Journal of Computer Vision*, vol. 60 of *91-110*, 2004.

[12] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz, "Active object recognition in parametric eigenspace," in *British Machine Vision Conference (BMVC)*, pp. 629–638, 1998.

[13] S. A. Hutchinson and A. C. Kak, "Planning sensing strategies in a robot work cell with multi-sensor capabilities," in *IEEE Transactions on Robotics and Automation*, vol. 6, pp. 765–783, 1989.

[14] S. J. Dickinson, H. I. Christensen, J. Tsotsos, and G. Olofsson, "Active object recognition integrating attention and view point control," in *Computer Vision and Image Understanding*, vol. 3, pp. 239–260, 1997.

[15] S. Roy, S. Chaudhury, and S. Banerjee, "Isolated 3-d object recognition through next view planning," in *IEEE Transactions on Systems, Man and cybernetics-Part A: Systems and Humans*, vol. 1, pp. 67–76, 2000.

[16] K. D. Gremban and K. Ikeuchi, "Planning multiple observations for object recognition," in *International Journal of Computer Vision (IJCV)*, pp. 137–172, 1994.

[17] F. G. Callari and F. P. Ferrie, "Active recognition: Using uncertainty to reduce ambiguity," in *International Conference on Pattern Recognition*, pp. 925–929, 1996.

[18] G. Koostra, J. Ypma, and B. de Boer, "Active exploration and keypoint clustering for object recognition," in *IEEE International Conference on Robotics and Automation(ICRA)*, 2008.

[19] D. Sabatta, D. Scaramuzza, and R. Siegwart, "Improved appearance-based matching in similar and dynamic environments using a vocabulary tree," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1008 – 1013, 2010.

[20] F. Callari and F. Ferrie, "Active object recognition: Looking for differences," in *International Journal of Computer Vision*, pp. 189–204, 2001.

[21] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Ninth IEEE International Conference on Computer Vision*, vol. 2, p. 1470, 2000.

[22] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz, "A comparision of probabilistic, possibilistic and evidence theoretic fusion schemes for active object recognition," in *Computing*, pp. 293–319, 1999.