Comparison of active SIFT-based 3D object recognition algorithms

Mogomotsi Keaikitse
MIAS (CSIR)
South Africa
mkeaikitse@csir.co.za

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

Jonathan Warrell
MIAS (CSIR)
South Africa
jwarrell@csir.co.za

***Abstract*—Active object recognition aims to manipulate the sensor and its parameters, and interact with the environment and/or the object of interest in order to gather more information to complete the 3D object recognition task as quickly and accurately as possible. It can leverage the mobility of robotic platforms to capture additional viewpoints about an object as single images are not always sufficient especially if objects appear in cluttered human environments. Active vision algorithms should reduce the number of viewpoints required to recognise an object and hence reduce the computational time as well.**

This paper compares two active object recognition systems. Both systems use SIFT features for object recognition, but use contrasting models, update and viewpoint selection strategies. The methods for integrating information across views used by the two systems are investigated. This is essential as this module is used to select the next best viewpoint. The number of viewpoints and the time taken to recognise objects are used to compare the performance of these two methods.

I. Introduction

Human vision is an active process and a strong motivation behind active computer vision. Humans can change their position or focus, among others things, to get a better understanding of a scene. Bajcsy[1] defines active vision as "the modeling and study of control strategies for perception, i.e., modeling of the sensors, the objects, the environment, and the interaction between them for a given purpose, which can be manipulation, mobility, and recognition". Using a digital camera as a sensor example, the strategies may include the ability to zoom in and out of the object of interest when required. It may also include moving either the camera or the object of interest itself to capture more information.

Active computer vision finds applications in object recognition, surveillance, scene understanding and reconstruction. All these research areas can benefit from the ability to control the sensor and interact with the environment or the object of interest. This paper concentrates on the application of active vision to object recognition which leverages the mobility of robotic platforms.

Object recognition is difficult because 2-D images are used to recognize 3-D objects. Most model-based[2][3][4] and geometric-based [5] 3-D object recognition systems consider the problem of recognizing objects based on the information gathered from a single image. This is the passive approach to object recognition. However, a single image may not be sufficient to uniquely recognize an object, either because the query object is partially occluded or a number of objects in the database have similar viewpoints.

The solution to the problem is to use images obtained at different viewpoints and, possibly, using different sensor parameters. Equipping artificial systems with the ability to interact with the objects and/or the environment improves recognition rates[6]. This is referred to as active object recognition. More formally, active object recognition is the ability to manipulate a camera or the object of interest to obtain more useful information to complete the object recognition task as quickly and accurately as possible.

Many systems have been proposed for fusing the extracted data and selecting the next best viewpoint. A large number of options exist for the model, update and viewpoint selection strategy. We compare two methods which make contrasting choices for these options, but are based on the same low-level features [7][8]. We compare the methods both as self-contained algorithms, and when elements of one are substituted for elements of the other. This allows us to isolate the effects of different parts of the contrasting systems.

Both methods use interest points detected using the Scale Invariant Feature Transform(SIFT) detector and descriptors[9]. In [8] all the SIFT features extracted from all the training images are clustered using a vocabulary tree[10]. In [7] a *pseudo-3D* model for each training object is generated by retaining features that are visible from any two adjacent viewpoints. To the authors knowledge, no comparison has been done before on two active object recognition systems. The proposed systems are always compared against the randomly selected viewpoint strategy. Moreover, very few of the available active object recognition methods use local features which are robust to occlusion.

The layout of this paper is as follows. Section II outlines the related work. Section III presents the two algorithms in detail. Section IV introduces the datasets used. Section V describes the experiments conducted and the results. Lastly, the conclusions are given in section VI.

II. Related work

A number of different approaches have been proposed for active object recognition[11], [12], [13], [14], [15]. Apart from the representation schemes used, the major differentiating factors are the next viewpoint selection and fusion tasks. With regard to fusion, the favoured approach is Bayes theory. Moreover, next viewpoint selection is often posed as an optimization problem. A case in point is [12] where the viewpoint that minimizes ambiguity is chosen as the next best viewpoint.

Many systems have been proposed for active 3D object recognition. The most popular method to integrate information extracted from multiple images is Bayes theory. Other methods used for fusion include discriminative approaches [7], [16], Dempster-Shafner theory [17], [18] and particle filters [13].

Systems generally use different types of data extracted from images to update the fusion component. These include parametric eigenspace data [11], [13], entropy maps [15] and SIFT features [7], [8]. One of the reasons we chose to compare Koostra et.al. and Govender et.al systems is that they use SIFT features extracted from the images. SIFT features are robust to changes in scale and illumination, and affine transformations. Global features such as the eigenspace representation tend to be sensitive to occlusion [19]. Another reason for choosing these two methods is that they were originally tested on images with objects appearing in a real-world environment with a degree of clutter. The database used in these experiments have objects appearing in cluttered environments with significant occlusion. The other systems [11][13][15] all use very simplistic datasets which are either synthetic images or images with single objects appearing in no clutter.

While [8] use a Bayesian update strategy, and an a priori next viewpoint selection mechanism, [7] use a discriminative model which they update additively, and use an online selection mechanism. This allows us to compare the merits of quite different approaches.

III. Active object recognition methods

A. Active object recognition using vocabulary trees

The authors of [8] 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 system uses the Scale Invariant Feature Transform (SIFT) [9] detector and descriptor to extract relevant object features. The structure of their system is, however, not SIFT dependent and thus any other descriptor or detector can be used for feature extraction.

The automatic viewpoint selector uses a vocabulary tree structure[10]. 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 is constructed using hierarchical k-means clustering where similar features are clustered together. For each node in the tree a TFIDF-like (Term Frequency Inverse Document Frequency) metric is calculated to capture the node's uniqueness:

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

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.

Every viewpoint for all objects in the database is 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 discretise the feature space to reduce feature dimensionality when considered in the observer component.

For object recognition no object hypothesis is given to the system. The criteria for selecting the next best viewpoint is based on the viewpoint with the highest combined weighting across all objects in the database and has not been previously visited.

Following the approach used in [11], the observer component updates an object belief probability with current view information in a recursive Bayesian manner using a prior determined from previous views. The system only captures and processes the next best viewpoint if the probability is less than a pre-defined threshold. 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 using a different feature extraction method and it would not affect the observer component and vice-versa.

B. Active object recognition: The method by Kootstra et. al.

The experimental setup in [7] entails a mobile platform on which a camera is mounted. In collecting the training images, the platform follows a circular trajectory with the object of interest at the center. It stops at regular angular intervals of 10 degrees to take a picture of the object. The assumption is that the ground is flat.

The ability to change viewpoint is used for model creation and active object recognition. During model creation, it is used to find stable keypoints and segment the object from the background. A stable keypoint is a keypoint that is visible from two images of the same scene taken from different viewpoints. The use of stable keypoints removes all the keypoints that are very sensitive to rotation, translation and other affine transformations. The ability to change viewpoints is used during active recognition to collect additional information for recognition. This is important when an object cannot be uniquely identified after a sequence of viewpoints. In such cases a more informative viewpoint must be chosen. The major contribution of [7] is the algorithm for selecting the next best viewpoint.

1) Model creation: Object segmentation is achieved by noting that, as the robot rotates about the object, the background stable keypoints are displaced a lot more than the object stable keypoints. Moreover, the assumption is that the robot moves on a flat surface and as such there is little change in the vertical components of the positions of the keypoints. Thus, the task of segmenting the object is to find stable keypoints that satisfy the following condition:

$$(|x_i - x_j| \leq x_T) \wedge (|y_i - y_j| \leq y_T)$$

where (x_i, y_i) is the location of the keypoint in the current image. (x_j, y_j) is the position of the same keypoint as seen in either the previous or next image. Two keypoints are a correct match if the distance between their descriptors is less than 0.6

This results in a *pseudo*-3*D* model of the object which is assigned an ID and pose, θ. Models of different objects are kept separate.

2) Object recognition: Object recognition may only take place once the keypoints database, $\mathbf{\Lambda}$, of the known objects is in place. These, together with the keypoints of the query image, are then used to determine the activation value of each model in the database. The closer the query object image is to one of the viewpoints of a model the higher the activation of that viewpoint. Hence, the higher the activation value of the model.

The first step towards determining the activation of a model is to match the query object keypoints to those that are in the database and belong to that model. This yields M pairs of matching keypoints. An activation level a_i may be calculated for the i^{th} pair as follows:

$$a_i = e^{-|\mathbf{p}_i - \mathbf{k}_n|}$$

where $\mathbf{p}_i$ and $\mathbf{k}_n$ are query object and database keypoints respectively. Note that the notion of a match between two keypoints is as defined in the above section.

The activation level of a model(ID+θ) given the query object keypoints, as observed from viewpoint δ, and the keypoints database is given by:

$$A^{\delta}_{\mathrm{ID},\theta} = \frac{\Sigma^{M}_{i=1} a_i}{\sqrt{|\mathbf{\Lambda}_{\mathrm{ID},\theta}|}} \tag{1}$$

where $\sqrt{|\mathbf{\Lambda}_{\mathrm{ID},\theta}|}$ is the number of keypoints in the database associated with the object ID and pose θ.

As the robot switches from one viewpoint to the next it collects information for recognition which is fused as follows:

$$A_{\mathrm{ID},\theta}(t) = \Sigma_{\delta \in \mathrm{E}} A^{\delta}_{\mathrm{ID},\theta} \tag{2}$$

where $A_{\mathrm{ID},\theta}(t)$ is the accumulated activation for object ID with pose, θ, at time, t, and $\mathrm{E} = \{\phi_0, \phi_1, ..., \phi_t\}$ is the set of viewpoints from where the observations are made.

3) Next viewpoint selection: Active object recognition leverages, among other things, the mobility of the robot to collect additional information needed to resolve ambiguities. The selected viewpoint ϕ_{t+1} is the angle that maximizes the expected activation of one of the models, that is:

$$\phi_{t+1} = \arg\max\nolimits_{\gamma \in (\Theta - \Phi)} \mathrm{E}(A_{\mathrm{ID},\theta}(t+1))$$

where Θ is the set of all possible viewpoints and $\Phi = \{\phi_0, \phi_1, ..., \phi_t\}$ is the set of previous viewpoints. The expected activation of the model(ID+pose) when viewed from viewpoint γ at time $t+1$ is given by:

$$\begin{aligned} \mathrm{E}(A_{\mathrm{ID},\theta}(t+1)) &= A_{\mathrm{ID},\theta}(t) + \mathrm{E}(A^{\gamma}_{\mathrm{ID},\theta}|O_{\mathrm{ID},\theta})P(O_{\mathrm{ID},\theta}) \\ \mathrm{E}(A^{\gamma}_{\mathrm{ID},\theta}|O_{\mathrm{ID},\theta}) &= \sqrt{|\mathbf{\Lambda}_{\mathrm{ID},\theta+\gamma}|} && (3) \\ P(O_{\mathrm{ID},\theta}) &= \frac{A_{\mathrm{ID},\theta}(t)}{\Sigma^{N}_{i=0} A_{o_i,\alpha_i}(t)} && (4) \end{aligned}$$

Equation (3) can be inferred from equation (1) by assuming that all keypoints belonging to object ID and pose, $\theta + \gamma$, have been perfectly matched.

We should note that a stopping condition is not specified in the original paper. However, one was latter communicated by the authors. The condition was to place a threshold, a lower bound, on the ratio of the largest to the second largest activation values. If this ratio exceeds the threshold then the object with the largest activation value corresponds to the query object.

IV. Datasets

The dataset used was created by the author of [8]. The following is the procedure used for obtaining that dataset. The training and testing datasets were captured using a Prosilica GE1900C camera. Everyday objects such as cereal and spice boxes were used. In compiling the training dataset, each object was placed on a turntable with a plain background as can be seen in Figure 1. The images were then captured at regular angular intervals of 20 degrees. All the images were captured around the y-axis which represents one degree of freedom. For the test set, the objects used for training were placed in cluttered environments with significant occlusion as shown in Figure 2. The testing images were also captured at 20 degree intervals.

V. Experiments

The stopping condition in Govender et. al. is when the probability that an object has been recognised exceeds 0.8. In contrast, the method by Kootstra et. al. places a lower bound on the ratio of the largest to the second largest activation value. If this ratio exceeds the given threshold, set to 1.25 in the experiments, then the model with the largest activation value corresponds to the object for which we are searching. The

Fig. 1. An example of a training image.

Fig. 2. An example of a testing image.

number of viewpoints required to recognise an objects is used to compare the two methods.

Two experiments were conducted. In the first experiment, the two methods were executed in their original format. In the second one, the next best viewpoints selected by Govender et. al.[8] were substituted into the method by Kootstra et. al. to determine if there would be an improvement or decline in performance. This also helps to isolate the contribution of the model and the viewpoint selection strategy on the results.

A. Adapting the method by Kootstra et. al. to our dataset

The method by Kootstra et. al.[7] was changed slightly. This change was implemented because the setup we used to capture the training and testing images differed from that by Koostra et. al. The segmentation process assumes that a robot rotates about a fixed object of interest. This means that keypoints that belong to the object are close to the centre of rotation and thus do not move a lot compare to keypoints that belong in the background. As a result, a stable keypoint belongs to the object of interest if its position between two consecutive images satisfies the condition:

TABLE I
ORIGINAL METHOD BY KOOTSTRA ET. AL.

Object	Ratio	# viewpoints	Recognised (7/10)
Cereal	4.98	1	Yes
Battery	1.77	1	Yes
Curry box	1.433	5	Yes
Elephant	3.26	1	Yes
Handbag	-	-	No
Mr Min	1.62	1	Yes
Salad Bottle	1.88	1	Yes
Spice Bottle	-	-	No
Spray Can	-	-	No
Spray Can2	1.39	1	Yes

$$(|x_i - x_j| \leq x_T = 12) \wedge (|y_i - y_j| \leq y_T = 4),$$

In our case, however, the camera is fixed and the object is placed on a rotating turntable. As a result, a stable keypoint belongs in the background if its location does not change between two consecutive images, i.e.:

$$(|x_i - x_j| \leq x_T = 4) \wedge (|y_i - y_j| \leq y_T = 4)$$

B. Results

Tables I and II show the results obtained using, respectively, the methods by Kootstra et. al. and Govender et. al. Dashes in the tables indicate that the methods could not recognised the objects. The ratio of the number of objects that are correctly recognised to the total number of objects suggests that the method by Govender et. al. outperforms that by Kootstra et. al. These ratios are shown in the heading of the last column of the respective tables. However, when both methods do recognise an object, the method by Kootstra et. al. requires fewer viewpoints. It should be noted that the method by Kootstra et. al. is computationally more intensive. This is because a feature from the query image must be matched against every feature in the database. In contrast, in the method by Govender et. al., a feature is propagated down a tree, and thus matches against nodes of the tree. Recall that a node is represented by a feature and a distance threshold.

Table III shows the results that are obtained by replacing the next viewpoint selection component of the method by Kootstra et. al. with that by Govender et. al. The recognition rate is shown in the heading of the last column of the table. There is a small decrease in the number of viewpoints required to recognise an object for the only object that requires more than one viewpoint. However, this is just too small to conclude that one next best viewpoint selection strategy outperforms the other. However, it may indicate that the data fusion scheme in [7] is primarily responsible for the improved recognition rates.

VI. CONCLUSION

Active vision is important because it reduces the computational costs required to recognise objects. This paper compared two active object recognition methods. The first method uses a vocabulary tree to cluster similar feature and

TABLE II
ORIGINAL METHOD BY GOVENDER ET. AL.

Object	Probability	# viewpoints	Recognised (8/10)
Cereal	1	1	Yes
Battery	0.9999	1	Yes
Curry box	0.8541	7	Yes
Elephant	0.9183	2	Yes
Handbag	0.9783	3	Yes
Mr Min	0.8586	2	Yes
Salad Bottle	0.9381	15	Yes
Spice Bottle	0.0789	16	No
Spray Can	0.8767	9	Yes
Spray Can 2	0.0542	15	No

TABLE III
GOVENDER ET. AL. NEXT-VIEWPOINT SELECTION + KOOTSTRA ET. AL. FUSION ALGORITHM

Object	Ratio	# viewpoints	Recognised (7/10)
Cereal	4.98	1	Yes
Battery	1.77	1	Yes
Curry box	1.52	4	Yes
Elephant	3.26	1	Yes
Handbag	-	-	No
Mr Min	1.62	1	Yes
Salad Bottle	1.88	1	Yes
Spice Bottle	-	-	No
Spray Can	-	-	No
Spray Can 2	1.39	1	Yes

structure the database[8]. The second one constructs a pseudo-3-D model for each object in the training set. The method in [8] outperforms that in [7] primarily because it can recognize a larger set of the training objects. It is also computationally more efficient.

ACKNOWLEDGMENT

This research was carried out with the financial support of the Council for Scientific and Industrial Research.

REFERENCES

[1] R. Bajcsy. Active perception. *Prec of the IEEE*, 76(8):996–1004, Aug 1988.

[2] A. R. Pope. Model-based object recognition- a survey of recent research. TECHNICAL TR-94-04, University of British Columbia, 1994.

[3] I. Weiss and M. Ray. Model-based recognition of 3d objects from single images. *IEEE Trans. on Pattern analysis and Machine Intelligence*, 23(2):116–128, 2001.

[4] D. G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31(3):355–395, 1987.

[5] J. Mundy. *Towards category-level object recognition*, chapter Object recognition in the geometric era, a retrospective, pages 3–29. Springer-Verlag, 2006.

[6] S. Dutta Roy, S. Chaundhury, and S. Banerjee. Active recognition through next view planning: A survey. *Pattern Recognition*, 37(3):429–446, 2004.

[7] G. Kootstra, J. Ypma, and B. de Boer. Active exploration and keypoint clustering for object recognition. In *Robotics and Automation*, pages 1005–1010. IEEE, 2008.

[8] N. Govender, J. Claassens, and J. Warrell. Active object recognition using vocabulary trees. In *Workshop on Robot Vision*. IEEE, 2013.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. of Comp. Vision*, 60(2):91–110, 2004.

[10] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. volume 2, pages 2161–2168. IEEE, 2006.

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

[12] P-P. Vázquez, M. Feixas, M. Sbert, and W. Heidric. Viewpoint selection using viewpoint entropy. In *Vision modeling and Visualization*, 2001.

[13] F. Deinzer, J. Denzler, and H. Neimann. Viewpoint selection- planning optimal sequences of views for object recognition. In *Comp. Analysis of Images and Patterns*, pages 65–73. Springer, 2003.

[14] F. Deinzer, J. Denzler, and H. Neimann. On fusion of multiple view for object recognition. *Pattern Recognition*, pages 239–245, 2001.

[15] T. Arbel and F. P. Ferrie. Viewpoint selection by navigation through entropy maps. volume 1, pages 248–254. IEEE, 1999.

[16] Z. Jia, Y-J. Chang, and T. Chen. Active view selection for object and pose recognition. In *Computer Vision- 3D Object recognition Workshop*, pages 641–648. IEEE, 2009.

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

[18] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz. A comparison of the probabilistic, possibilistic and evidence theoretic fusion schemes for active object recognition. *Computing*, 62(4):293–319, 1999.

[19] A. Leonardis and H. Bischof. Robust recognition using eigenimages. *Computer Vision and Imgage Understanding*, 78:99–118, 2000.