Active Object Recognition using Vocabulary Trees

N Govender, J. Claassens, P. Torr, J. Warrell

—

Presentation by Aishwarya Padmakumar

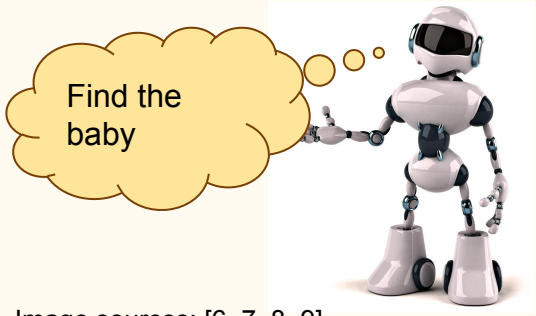
Motivation

Fast and accurate classification of objects is a necessity for robotic manipulation tasks

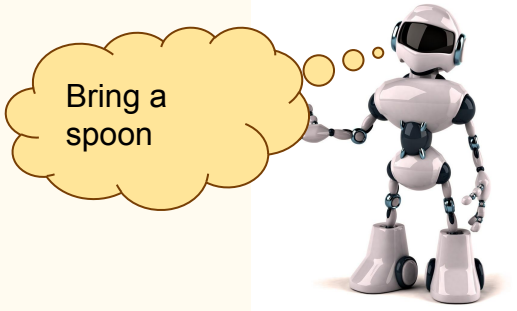
Image sources: [3, 4, 5]

Motivation: Objects may be ...

Occluded



Hidden in clutter

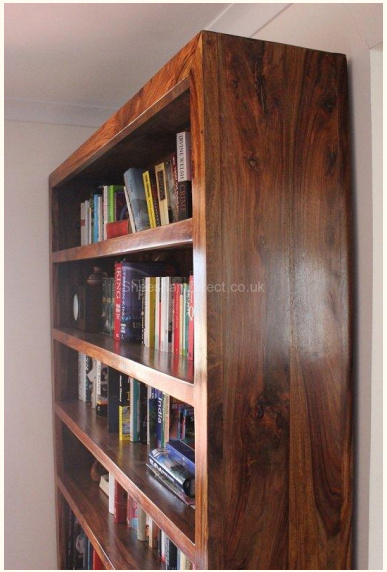


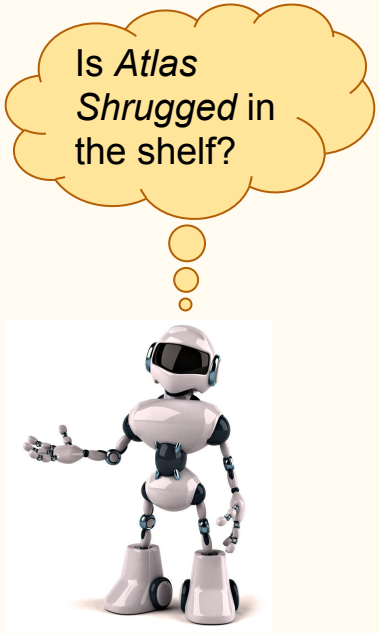
Visually confusing because of similar looking objects

Image sources: [6, 7, 8, 9]

Problems with single viewpoint

Single viewpoint may be of poor quality





Single view may not be enough to identify an object uniquely

Image sources: [9, 10, 11, 12]

Active object recognition

- It is possible to obtain images from different views but there is a cost associated with each additional image to process.
- Cost could be as simple as additional compute time per image - undesirable when fast detection is key
- Goal: Uniquely identify an object using minimum number of images
- Steps -
  - → Selecting next best viewpoint
  - → Integration of relevant information from new image obtained

Differences from prior work

- Number of images and sequence is variable
- Explicitly considers occlusion or clutter
- Select views on based on promised **uniqueness** of features rather than minimizing entropy or some other notion of error

What is a vocabulary tree?

- A technique for organizing any kind of data represented in the form of vectors.
- Obtained using hierarchical k-means
- k is the branching factor of the tree.
- The root is the centroid of the entire dataset
- First, k-means is performed on the entire dataset and the centroids become children of the root

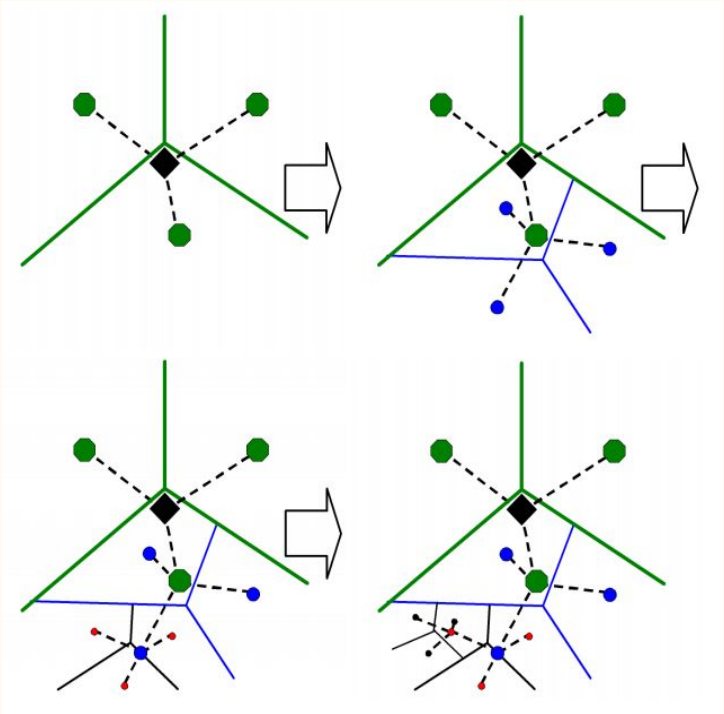


Image source: [2]

What is a vocabulary tree?

- The dataset is partitioned into the k clusters, each of which is associated with the node of its centroid.
- Each node is further split by performing k-means on the data points associated with it.
- Continued till there are sufficiently few data points associated with each node.

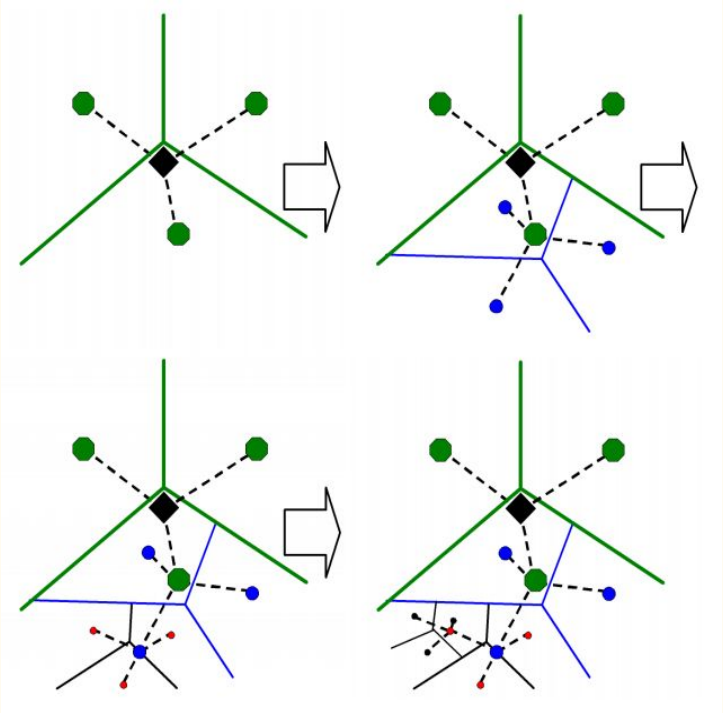


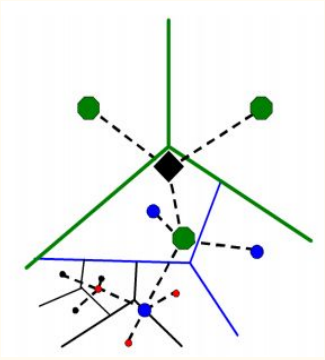
Image source: [2]

How they build the vocabulary tree

Image

SIFT features

Hierarchical K-means clustering



Vocabulary tree (Nodes are clusters of SIFT features)

- The complexity depends only on the number of training images - not number of degrees of freedom in viewpoints.
- What about less textured objects?
- CNN features - Instance vs category recognition

Image source: [16]

Scoring features

- Each node i is associated with a uniqueness score -

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

- M - total number of images in the database
- M_i - number of images in the database having some feature in the cluster i
- Uniqueness score of a feature - Sum of w_i's on the path from the root to it
- Uniqueness score of a viewpoint - Sum of scores of features present in it ★

Object verification

Input image

SIFT matching, Hough transform

Closest training image

View Selection

Pose Estimate

Next best view

Object hypothesis

Observer

Object Belief

Image sources: [13, 14, 15, 16]

View selection for object verification

Relative to the current pose estimate, the view selection component selects a view that

- → Has not been previously visited
- → Has the largest uniqueness weighting for that object

View selection for object verification

Relative to the current pose estimate, the view selection component selects a view that

- → Has not been previously visited
- → Has the largest uniqueness weighting for that object

View selection for object verification

Relative to the current pose estimate, the view selection component selects a view that

→ Has not been previously visited
→ **Has the largest uniqueness weighting for that object**

- Requires calculation of uniqueness score for all possible viewpoints
- Requires calculation of SIFT features of all possible viewpoints

Object Recognition

- ➢ Overall pipeline is similar to verification
- ➢ Input: Image (no object hypothesis)
- ➢ Next best view is the one
  - → Which has not been previously visited
  - → With highest combined uniqueness score across all objects in the database
- ➢ Maintain a belief for each possible object

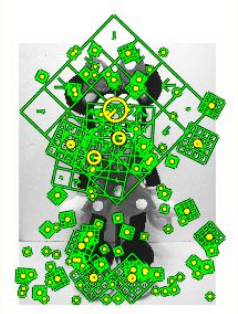
Observer

- Integrates information from a new view to update object belief
- Modifications to vocabulary tree -
  - Leaf nodes store the probability of the feature occurring at least once given each object (discrete density function) - $P(N|O)$
  - Calculation - smoothed normalized counts of features occurrences in training images

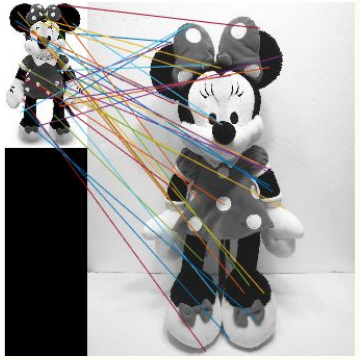
Observer is independent from viewpoint selection - Advantage or Disadvantage

Observer - Processing a new image (viewpoint)

Input image



Extract SIFT features



Find closest training image using Lowe's method and verification using Hough transform ★

Retain features satisfying Hough transform

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

Calculate probability of object given each feature

Object Belief
(assuming independence of features)

Image sources: [14, 16]

So what does their method really save on ...

- ➢ Computation saved by their method - observer component for each image not used by the active system.
  - → Assuming SIFT features of training images are stored, observer component still needs nearest neighbour comparison with each training image.
  - → In case of object recognition, needs comparison with every training image in the DB

But if you had a dataset of the size of ImageNet, you can't do this even for a few views.

Dataset

Training -

- 20 everyday objects
- Images captured every 20 degrees against a plain background on a turntable using a Prosilica GE1900C camera
- Objects that share a number of similar views were included

Testing -

- Objects used in the training data captured at every 20 degrees in a cluttered environment with significant occlusion

Dataset

Test setups - finding objects in cluttered settings

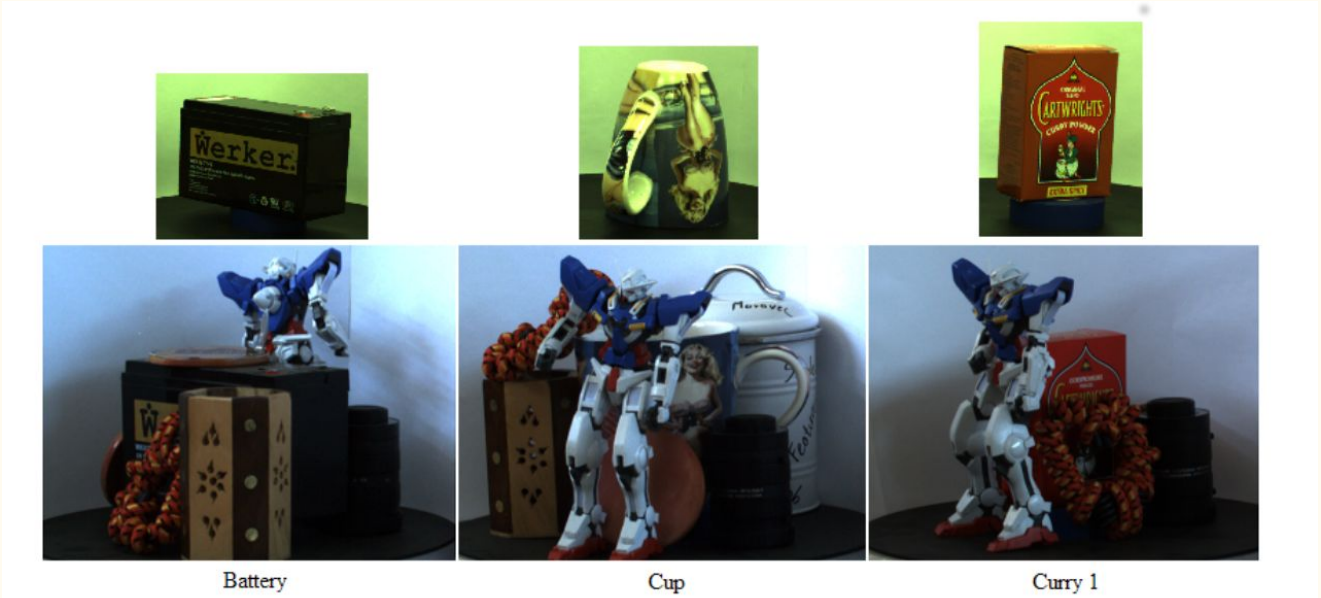


Figure 4. Examples of test images of occluded objects appearing in a cluttered environment which are to be verified

Image source: [1]

Dataset - Discussion points

- Other datasets - NORB dataset
- Is 20 objects really state of the art?
- Using the GERMS dataset - images vs video
- Could context be included? - Theoretically SIFT features can capture some context but in their setup it won't be useful since training images have plain background
- What if the training data had a more cluttered background?

Experiments

Object verification -

- ➢ Retrieves images until belief of hypothesized object reaches 80%
- ➢ Baseline : Random selection of next viewpoint
- ➢ Results

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

Image source: [1]

Experiments

Results - increase in belief after each view

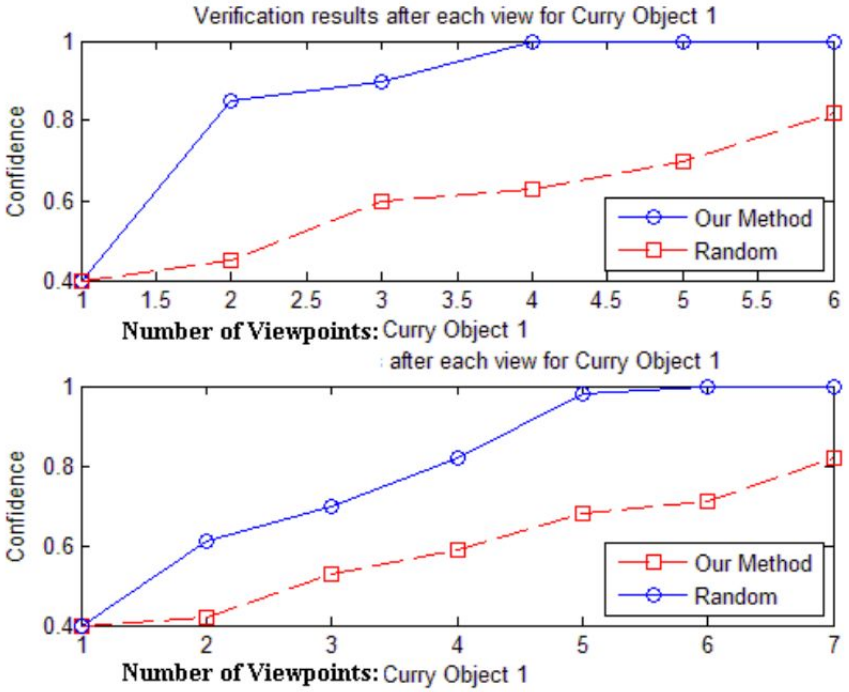


Figure 5. Confidence values after each view for verification and recognition

Image source: [1]

Experiments

Concerns:
- Small dataset
- Why 80% confidence?
- Other baselines/ comparisons

Object recognition -

➢ System retrieves next best viewpoint till belief for some object reaches 80%
➢ Results

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

Image source: [1]

Thank You!

References

[1] Active Object Recognition using Vocabulary Trees.
N Govender, J. Claassens, P. Torr, J. Warrell.
Workshop on Robot Vision, 2013.

[2] Scalable Recognition with a Vocabulary Tree
David Nister, Henrik Stewenius
Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) 2006

Image Sources

[3] http://a.abcnews.go.com/images/GMA/140121_gma_mathison_822_wg.jpg

[4] https://mercedesbenzblogphotodb.files.wordpress.com/2011/03/japan-after-2011-earthquake.jpg

[5] http://www.dotemu.com/sites/default/files/product/screenshots/screen_space_colony_7.png.jpg

[6] https://lightspinner.files.wordpress.com/2011/06/115-scared-kid.jpg

[7] http://img.8-ball.xyz/2015/09/24/dirty-messy-kitchen-l-dd76c382377da96b.jpg

[8] https://s-media-cache-ak0.pinimg.com/736x/b7/69/fb/b769fbf3c9d2d06b41aaba3665914e29.jpg

[9] http://wall.wallrage.com/wp-content/uploads/Cute-Robot-Wallpaper-for-Desktop.jpg

[10] http://www.sheeshamdirect.co.uk/wp-content/gallery/bespoke-bookcase/sheesham-bookcase-side-view.jpg

[11] http://www.feelmorebetter.com/shop/media/catalog/product/cache/1/image/9df78eab33525d08d6e5fb8d27136e95/m/u/mug3_zoom.jpg

[12] http://www.feelmorebetter.com/shop/media/catalog/product/cache/1/image/9df78eab33525d08d6e5fb8d27136e95/m/u/mug7_zoom.jpg

Image Sources

[13] http://ecx.images-amazon.com/images/I/317PGe5s9cL.jpg

[14] https://s-media-cache-ak0.pinimg.com/736x/c2/49/bc/c249bc88826d06e3a5fd4988cec8d79b.jpg

[15] https://upload.wikimedia.org/wikipedia/en/6/67/Minnie_Mouse.png

[16] http://i.ebayimg.com/00/s/OTgwWDU1OA==/z/WfoAAOSwDwtUnd~7/$_1.JPG?set_id=880000500F