

Unposed Object Recognition using an Active Approach

Wallace Lawson, J. Gregory Trafton

Naval Center for Applied Research in Artificial Intelligence, Washington, DC
{ed.lawson, greg.trafton}@nrl.navy.mil

Keywords:

Object Recognition, Active Object Recognition, Neural Networks

Abstract:

Object recognition is a practical problem with a wide variety of potential applications. Recognition becomes substantially more difficult when objects have not been presented in some logical, “posed” manner selected by a human observer. We propose to solve this problem using active object recognition, where the same object is viewed from multiple viewpoints when it is necessary to gain confidence in the classification decision. We demonstrate the effect of unposed objects on a state-of-the-art approach to object recognition, then show how an active approach can increase accuracy. The active approach works by attaching confidence to recognition, prompting further inspection when confidence is low. We demonstrate a performance increase on a wide variety of objects from the RGB-D database, showing a significant increase in recognition accuracy.

1 Introduction

State-of-the-art-approaches to visual recognition have focused mostly on situations when objects are “posed” (i.e., the camera angle, lighting, and position has been chosen by an observer). When conditions become more variable, the ability to visually recognize objects quickly decreases. In one prominent example demonstrating this affect, [Pinto et al., 2008] produced very good accuracy classifying objects from the Caltech-101 dataset [Fei-Fei et al., 2004], but their state-of-the-art approach was reduced to performing at chance when variation was introduced. Specifically, this meant viewing objects at any arbitrary pan, tilt, scale, and rotation (both in plane and depth).

Unfortunately, such variability is common in the objects that we see scattered throughout our environment. In some cases (see figure 1(a)) it may be difficult for even the most robust visual object recognition approach to recognize an object. What results is a degraded performance from the object recognition system. Figure 1(a) shows two objects from the RGB-D dataset. From left to right the objects are a dry battery, and a hand towel. However, in both cases, the object classes could be mistaken with similar classes. For example, the dry battery could easily be mis-

taken for a flashlight or a pack of chewing gum. The hand towel could easily be confused for a 3 ring binder.

Figure 1(b) shows the accuracy of Leabra (described further Section 3.1) recognizing a dry battery over a range of different pan angles, with a slightly different camera tilt. While performance is generally good, there is a point at which performance drops significantly. A system that had been recognizing objects with an accuracy of about 90% suddenly decreases to an accuracy of 30% when the pan and tilt of the object modified. An image from this region is shown in figure 1(a).

The strategy of improving object recognition through multiple viewpoints is referred to as *active object recognition* [D. Wilkes, 1992]. Several ([Denzler and Brown, 2002, Farshidi et al., 2009, LaPorte and Arbel, 2006]) have proposed probabilistic frameworks for active object recognition. These frameworks serve to both incorporate multiple viewpoints as well as incorporating prior probability. However, most have been evaluated on only a small number of objects, using simple recognition schemes chosen specifically to highlight the benefits of active recognition.

We demonstrate the benefit of active object recognition to improve the results of a state-of-the-art approach, specifically, to improve in ar-

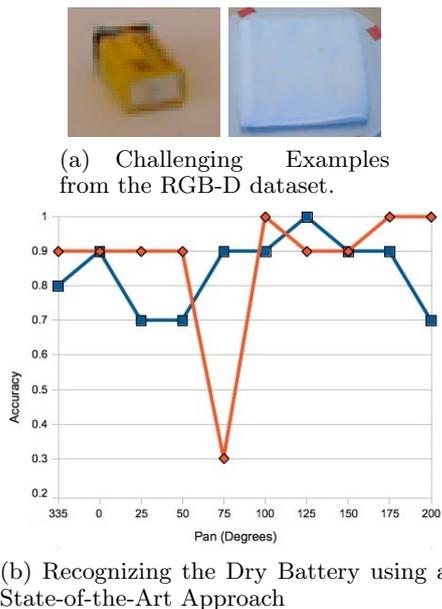


Figure 1: Images from the RGB-D dataset.

eas where performance is affected by the pose of an object. We recognize objects using Leabra¹, which is a cognitive computational neural network simulation of the visual cortex. The neural networks have hidden layers designed to mimic the functionality of the primary visual cortex (V1), the visual area (V4), and the inferior temporal cortex (IT). We extend Leabra by adding a confidence measure to resulting classification, then use active investigation when necessary to improve recognition results.

We demonstrate the performance on our system using the RGB-D [Lai et al., 2011] database. The RGB-D contains a full 360° range of yaw, and three levels of pitch. We perform active object recognition on 115 instances of 28 object classes from the RGB-D dataset.

The remainder of the paper is organized as follows. We present related work in the field of active object recognition in Section 2. We discuss our approach in Section 3, then present experimental results in Section 4 with concluding remarks in Section 5.

2 Related Work

Wilkes and Tsotsos’ [D. Wilkes, 1992] seminal work on active object recognition examined 8 origami objects using a robotic arm.

¹<http://grey.colorado.edu/emergent/>

The next best viewpoint was selected using a tree-based matching scheme. This simple heuristic was formalized by Denzler and Brown [Denzler and Brown, 2002] who proposed an information theoretic measure to select the next best viewpoint. They use average gray level value to recognize objects, selecting the next pose in an optimal manner to provide the most information to the current set of probabilities for each object. They fused results using the product of the probabilities, demonstrating their approach on 8 objects.

Jia et al [Jia et al., 2010] demonstrated a slightly different approach to information fusion, using a boosting classifier to weight each viewpoint according to the importance for recognition. They used a shape model to recognize objects, using a boosted classifier to select the next best viewpoint. They recognized 9 objects in multiple viewpoints with arbitrary backgrounds.

Browatzki et al. [Browatzki et al., 2012] used an active approach to recognize objects on an iCub humanoid robot. Recognition in this case was performed by segmenting the object from the background, then recognizing the object over time using a particle filter. The authors demonstrated this approach to recognize 6 different cups with different colored bottoms.

3 Methodology

We use Leabra to recognize objects (section 3.1). Once an object has been evaluated by Leabra, we find both the object pose (section 3.2), and attach confidence to the resulting classification (section 3.5). Finally, when the resulting classification has low confidence, we actively investigate (section 3.6).

3.1 Leabra

The architecture of a Leabra neural network is broken into three different layers, each with a unique function. The V1 layer takes the original image as input, then uses wavelets [Gonzalez and Woods, 2007] at multiple scales to extract edges. The V4 layer uses these detected edges to learn a higher level representation of salient features (e.g., corners, curves) and their spatial arrangement. The features extracted at the V1 layer includes multiple scales, therefore features extracted in the V4 layer have a sense of the large and small features that are

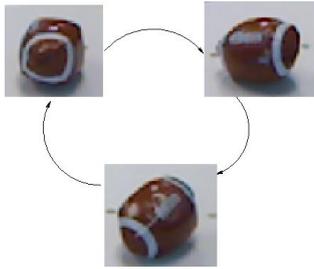


Figure 2: An example of visual aspects from one level of pitch. The images show different visual aspects, and the arrows show how each of these visual aspects are connected.

present in the object. The V4 layer also collapses on location information, providing invariance to the location of the object in the original input image. The V4 layer feeds directly into the IT activation layer, which has neurons tuned to specific viewpoints (or visual aspects) of the object.

3.2 Visual Aspects

Object pose plays an important role in recognition. We consider pose in terms of *visual aspects* [Cyr and Kimia, 2004, Sebastian et al., 2004] (see figure 2). When an object under examination is viewed from a slightly different angle, the appearance generally should not change. When it does not, we refer to this as a “stable viewpoint”, both the original and the modified viewpoint belongs to the same visual aspect V_1 . However, if this small change in viewing angle affected the appearance of the object, we would call this an “unstable viewpoint” representing a transition between two different visual aspects V_1 , and V_2 .

The human brain stores pose in a similar manner. Neurophysiological evidence suggests that the brain has *view-specific encoding* [Kietzmann et al., 2009, Frank et al., 2012]. In this encoding scheme, neurons in the IT cortex activate differently depending on how an object appears. Referring to Figure 2, when we look at the football in the first visual aspect, a certain set of neurons in the IT layer activate. When we look at the football in the second visual aspect, a different set of neurons activate.

To find visual aspects, we use the IT layer in the Leabra neural network. We find visual aspects using unsupervised learning, clustering IT activations. We describe this process in section 3.3.

3.3 Finding Aspects

Classifying an object using a Leabra network produces a set of neurons that have been activated in the IT layer. Leabra contains a total of 210 neurons in this layer, with similar activation patterns occurring when an object is viewed in a similar pose. We group activation patterns using unsupervised learning through k -means clustering [Duda et al., 2000].

Some care must be taken to establish the number of clusters, k , since this is synonymous with the number of visual aspects of an objects. This number is variable depending on the complexity of the object. For example a simple, uniformly colored, perfectly symmetric object such as a ball would only have one aspect. That is, a change in viewing angle will never affect the appearance of the object. Contrast this with a more complicated object, such as an automobile. An automobile would likely have a great number of visual aspects because of its complex structure.

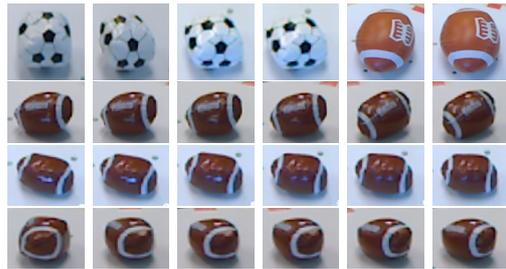


Figure 3: Four different visual aspects found using clustering.

The value of k cannot be estimated a priori, so we set this value using a heuristic based on viewpoint stability. A small change in viewpoint (δ) should generally not result in a new visual aspect. Therefore, when the correct value of k has been found, all of the elements resulting clusters (cl) will mostly all belong to stable viewpoints. We determine the quality of the clustering using the heuristic shown in Eq. 1, where c represents a cluster.

$$m(c) = \frac{\sum_{i \in c} |cl(pose(i)) - cl(pose(i) + \delta)|}{|c|} \quad (1)$$

To determine the correct number of visual aspects, we set k to a large number, then evaluate each resulting cluster. If the majority of the elements of any cluster do not belong to stable viewpoints, k is decreased, then the process is repeated. Some visual aspects from different object classes are shown in figure 3.

3.4 Distinctiveness of Visual Aspects

A basic tenet of active object recognition is that some viewpoints have greater distinctiveness than others. In this section, we establish the distinctiveness of each visual aspect using STRoud [Barbara et al., 2006], a test which evaluates the distinctiveness (or conversely “strangeness”) of the members of each class. The *strangeness* of a member (m) is evaluated using the ratio of the distance to other objects of that class c over the distance to all points of other classes c^{-1} (Eq. 2). In practice, we evaluate this by selecting the k smallest distances.

$$str(m, c) = \frac{\sum_{i=1}^K distance^c(i, m)}{\sum_{i=1}^K distance^{c^{-1}}(i, m)} \quad (2)$$

The sum of the distance to objects in the same class c should be much smaller than the sum of the distances to other classes c^{-1} . Therefore, a distinctive data point would have very low strangeness. When referring to visual aspects (s), the probability that we have correctly identified object (o) in visual aspect s is determined using

$$p(s|o) = \frac{|\forall_{i \in s} str(i, s) \leq str(o, s)|}{|s|} \quad (3)$$

3.5 Recognition from a Single Viewpoint

To recognize object o , we use the IT activation pattern from Leabra (a), then compare this against known classes. We compute the strangeness that the object belongs to each class using Eq. 2. The probability of recognition is conditionally dependent upon the distinctiveness of the visual aspect s , as well as the confidence that the object belongs to visual aspect s . The probability that we have recognized an object of class o is $p(o_{ix}|a_x s_x)$, for object i using image x .

$$p(o_{ix}|a_x, s_x) = p(o_{ix}|a_x)p(o_{ix}|s_x) \quad (4)$$

$$p(o_{ix}|a_x) = \alpha p(a_x|o_{ix})p(o_{ix}) \quad (5)$$

$$p(o_{ix}|s_x) = \alpha p(s_x|o_{ix})p(o_{ix}) \quad (6)$$

Eq. 5 can be interpreted as the probability that we have observed the object given a particular activation pattern. If the activation pattern

observed is quite similar to known activation patterns for object o_i , we expect the probability to be high. Similarly, eq. 6 can be interpreted as the general confidence of recognizing object o in estimated visual aspect s . Combining the two (eq. 4) produces a uncertainty measure that accounts for both similarity of activation patterns as well as the confidence in the visual aspect.

3.6 Active Recognition

When confidence is low, a single image may not be sufficient to correctly recognize the object. In these cases, we make a small local movement to view the object from a slightly different perspective, then combine the measurements. The probability that the object belongs to class i , as was suggested in [Denzler and Brown, 2002], is estimated using the product of all measurements that have been taken over time (n). This also has the potential for incorporating a prior probability, which we have set to a uniform probability.

$$p(i) = \prod_{x=1}^n p(o_{ix}|a_x, s_x) \quad (7)$$

4 Experimental Results

We experimentally validate our approach using the RGB-D dataset [Lai et al., 2011]. This particular dataset was selected due to its large number of object classes, many instances of each class, and the range of poses where each instance was imaged. A few examples of training images are in Figure 5. Our experiments are conducted using 115 instances of 28 object classes. RGB-D has images of objects when viewed from three different levels of camera pitch, rotating the object a full 360° at each pitch level. We use 39 randomly selected images per object for training (approximately 5% of the images). One third of the remaining images were used for validation, the remaining images are used for testing (52,404 images).

We extract the object using the foreground mask provided in the RGB-D dataset. The foreground mask represents the part of the region that is not on the table, as estimated using the Depth information provided by the Kinect. The size of the object was normalized in the same manner as was previously described in [Pinto et al., 2008]. The purpose of foreground

extraction and size normalization is to remove irrelevant size cues and to provide a measure of scale invariance.

Table 1 shows the recognition rates for Leabra (i.e., single viewpoint or static object recognition), and active object recognition. Active object recognition has been set for very high confidence ($p=0.99999$) and therefore will only recognize an object when it is extremely confidence in the results. Note that this is a confidence on a decision-level basis, and does not necessarily predict the overall performance of the system, as performance is driven by the variability of the testing data.

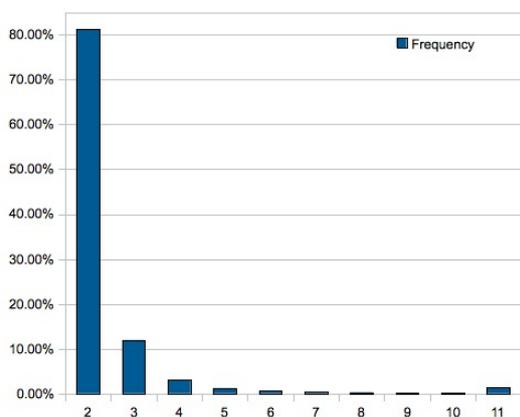


Figure 4: Frequency and number of positions used during the active object recognition process.

During active investigation, on average, objects are examined at 2.4 positions before they are recognized. The frequency of the positions used during examination are shown in 4.

Across all of the objects, the static approach has a precision of 90.55%, and the active approach has a precision of 96.81%. Furthermore, the standard deviation of precision varies greatly with the approaches. The standard deviation for static is 9.27%, the active approach is 4.95%. This indicates that not only is the accuracy of the system improving, but the number of objects with a low level of accuracy is also improving.

5 Discussion

State-of-the-art approaches to object recognition have been demonstrated to perform very well on posed objects. We have shown that unposed objects can be more difficult to recognize, particularly in degenerate viewpoints. Further, an

active strategy can boost the performance of the system even when considering a simple approach to next best viewpoint selection. Using only a random movement strategy, we demonstrated a 6% boost in improvement without significantly impacting the recognition speed of the system (requiring only 2.4 positions on average).

Acknowledgment

This work was partially supported by the Office of Naval Research under job order number N0001407WX20452 and N0001408WX30007 awarded to Greg Trafton. Wallace Lawson was also funded by the Naval Research Laboratory under a Karles Fellowship.

REFERENCES

- [Barbara et al., 2006] Barbara, D., Domeniconi, C., and Rodgers, J. (2006). Detecting outliers using transduction and statistical testing. *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- [Browatzki et al., 2012] Browatzki, B., Tikhanoff, V., Metta, G., Bulthoff, H., and Wallraven, C. (2012). Active object recognition on a humanoid robot. In *IEEE International Conference on Robotics and Automation*.
- [Cyr and Kimia, 2004] Cyr, C. and Kimia, B. (2004). A similarity-based aspect-graph approach to 3d object recognition. *International Journal of Computer Vision*, 57(1).
- [D. Wilkes, 1992] D. Wilkes, J. T. (1992). Active object recognition. In *Computer Vision and Pattern Recognition (CVPR)*.
- [Denzler and Brown, 2002] Denzler, J. and Brown, C. (2002). Information theoretic sensor data selection and active object recognition and state estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(2).
- [Duda et al., 2000] Duda, R., Hart, P., and Stork, D. (2000). *Pattern Classification*. Wiley Interscience.
- [Farshidi et al., 2009] Farshidi, F., Sirouspour, S., and Kirubarajan, T. (2009). Robust sequential view planning for object recognition using multiple cameras. *Image and Vision Computing*.
- [Fei-Fei et al., 2004] Fei-Fei, L., Fergus, R., and Perona, P. (2004). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE Conferece on Computer Vision and Pattern Recognition*.

Table 1: Object Recognition Results

object	ball	binder	bowl	calculator	camera	cellphone	cerealbox	coffeemug
Leabra	95.04%	88.92%	97.80%	84.46%	69.95%	93.47%	77.64%	94.68%
Active	97.30%	96.54 %	99.92%	93.80%	87.98%	91.84%	96.15%	97.68%
	comb	drybattery	flashlight	foodbag	foodbox	foodcan	foodcup	foodjar
Leabra	86.77%	96.50%	91.65%	82.93%	93.51%	96.38%	94.00%	63.07%
Active	94.06%	100.00%	98.68%	98.29%	99.69%	99.79%	99.19%	79.02%
	gluestick	handtowel	keyboard	kleenex	lightbulb	marker	notebook	pitcher
Leabra	98.97%	80.55%	76.49%	83.63%	91.80%	96.06%	90.93%	80.62%
Active	99.87%	98.92%	85.30%	96.76%	99.01%	98.92%	97.86%	98.12%
	plate	pliers	scissors	sodacan				
Leabra	99.90%	96.62%	93.61%	97.85%				
Active	99.90%	98.63%	98.99%	94.22%				



Figure 5: Examples of training images from the RGB-D dataset.

- [Frank et al., 2012] Frank, M., Munakata, Y., Hazy, T., and O'Reilly, R. (2012). *Computational Cognitive Neuroscience*.
- [Gonzalez and Woods, 2007] Gonzalez, R. and Woods, R. (2007). *Digital Image Processing*. Prentice Hall.
- [Jia et al., 2010] Jia, Z., Chang, Y., and Chen, T. (2010). A general boosting based framework for active object recognition. *British Machine Vision Conference (BMVC)*.
- [Kietzmann et al., 2009] Kietzmann, T. C., Lange, S., and Riedmiller, M. (2009). "computational object recognition: A biologically motivated approach". *Biol. Cybern.*
- [Lai et al., 2011] Lai, K., Bo, L., Ren, X., and Fox, D. (2011). A large-scale hierarchical multi-view rgb-d object dataset. In *International Conference on Robotics and Automation*.
- [LaPorte and Arbel, 2006] LaPorte, C. and Arbel, T. (2006). Efficient discriminant viewpoint selection for active bayesian recognition. *International Journal of Computer Vision*, 68(3):267–287.
- [Pinto et al., 2008] Pinto, N., Cox, D., and DiCarlo, J. (2008). Why is real-world visual object recognition hard? *PLoS Computational Biology*, 4(1).
- [Sebastian et al., 2004] Sebastian, T., Klein, P., and Kimia, B. (2004). Recognition of shapes by editing their shock graphs. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(5):550 – 571.