Enhancing Object Recognition With Dynamic Cognitive Context

Laura M. Hiatt, Wallace E. Lawson, Anthony M. Harrison and J. Gregory Trafton
U.S. Naval Research Laboratory
4555 Overlook Ave, SW
Washington, DC 20375

Abstract

Object recognition continues to be a challenging area of research, especially for objects situated in their real-world environments. Yet, people are able to recognize objects in their daily environments with ease, in part because of their ability to quickly and effectively learn contextual relationships between objects in the world. Here, we show that leveraging the rich, dynamic context of a computational cognitive architecture can facilitate autonomous object recognition in real-world environments with no prior training. We show that, on the real-world NYU Depth V2 dataset, this cognitive context improves object recognition with an 8% gain in mean average precision.

Introduction

Despite recent progress, object recognition in real world environments, such as those encountered by mobile, autonomous systems, is still a challenging problem. Poor resolution, non-iconic viewpoints, specular reflections, and poor focus can all adversely affect recognition performance. Humans, in contrast, can recognize objects in these environments with ease, in part because they quickly and effectively learn contextual relationships between objects in the world, and exploit that knowledge as they encounter future objects (Oliva and Torralba 2007).

In the past, context has been shown to help improve object recognition performance by providing suggestions about what objects are likely to be using information other than their appearance (Divvala et al. 2009; Galleguillos, Rabinovich, and Belongie 2008; Felzenszwalb et al. 2010). In these approaches, context can come from several sources. The most common is pixel level context, which uses information such as the position of an object within an image and the surrounding pixel values to influence recognition. A different context source, semantic context, uses previously identified objects to suggest what may be likely to be seen next. Scene context, in contrast, influences recognition by taking into consideration the setting of an image.

Typically, however, these approaches to context are static in nature, and require training in a comprehensive set of realistic environments and settings. They also are unable to adapt to new settings, environments, or viewpoints, such as those frequently encountered by autonomous systems. In short, these approaches fall short of the rich, dynamic context that people rely upon.

In this paper, we work towards closing that gap by using a computational cognitive architecture to provide dynamic context with no prior training. The architecture, ACT-R/E (Trafton et al. 2013), has been extensively verified against many aspects of human cognition (Anderson et al. 1998; Schneider and Anderson 2011), and has provided cognitively-plausible context in a variety of other settings (Hiatt and Trafton 2013; 2015). Here, context takes the form of associations between related concepts that are learned incrementally over time. The more that concepts are thought about with one another, the stronger their association becomes; then, these associations can be used as a prediction for what is likely to be seen next, based on both what object is currently being looked at (semantic context), as well as knowledge of what type of scene one is in (scene context).

We investigate the efficacy of this approach to dynamic context by studying its improvement over a baseline, state-of-the-art approach to computer vision (Razavian et al. 2014). Importantly, we consider the improvement that dynamic context can produce with varying amounts of training and learning. We show that, on the challenging NYU Depth V2 dataset (Silberman et al. 2012), dynamic, cognitive context boosts object recognition precision by 8%, regardless of whether it has prior training.

Related Work

Context is used heavily in the human visual system to bias both where we look and what objects we expect to see (Oliva and Torralba 2007). Much of the previous work on context in computer vision has focused on how to learn this context from large databases of images. Li et al. (2010) combined scene level cues (e.g., categorization, depth, saliency) in the framework referred to as the *feedback enabled cascaded classification model* (FE-CCM). Divvala et al. (2009) explored a gridded version of scene categorization, which divides the image into smaller regions which are then used for prediction. In addition to scene-level context, semantic context can be learned both on its own (Lawson, Hiatt, and Trafton 2014; Galleguillos, Rabinovich, and Belongie 2008) or simultaneously (Mottaghi et al. 2014).

The problem inherent in these approaches, however, are that they require extensive training and so cannot easily

learn context in a new environment. It is possible, for example, to encounter a known object in a new location; for these approaches to take that into consideration, they would need to continuously be re-learning about objects, a prohibitively expensive approach. Our work, on contrast, focuses on learning context online and incrementally. This provides a greater level of flexibility, effectively managing the rich dynamics of real-world environments.

Approach

In this section, we begin by describing our baseline system for classifying objects using an existing, state-of-the-art approach to object recognition. We then describe how we learn and incorporate dynamic cognitive context into our object recognition approach.

Baseline Object Recognition

As our baseline object recognition system, we follow in others' footsteps and use the state-of-the-art approach of SVM / off-the-shelf features from convolutional neural networks (CNNs) (Razavian et al. 2014). This approach follows the standard protocol of dividing data up into training and testing sets; training on the training set; and then evaluating its efficacy on the test set.

To do this, the approach begins with a pre-trained CNN; here, AlexNet (Krizhevsky, Sutskever, and Hinton 2012), a trained version of the publicly available caffe (Jia et al. 2014). Then, AlexNet is given each of the training images, and the features for each image are extracted from the fc7 layer. These extracted, "off-the-shelf," features are next used to train a linear SVM to classify object images, using the publicly available libsvm package (Chang and Lin 2011).

Then, for any given test image, the probability distribution of its possible classes is found by minimizing Equation 1 over all classes:

$$\min_{p} \frac{1}{2} \sum_{i=1}^{k} \sum_{j:j \neq i} (r_{ji} p_i - r_{ij} p_j)^2 \tag{1}$$

Here, $p_i \geq 0$, $\sum_{i=1}^{k} p_i = 1$ and r_{ij} is the probability, provided by the SVM, that the sample comes from either class i or class j, estimated using 5-fold cross validation.

Dynamic Cognitive Context

We model dynamic context using the cognitive architecture ACT-R/E (Trafton et al. 2013). ACT-R/E is an integrated, hybrid symbolic/subsymbolic system, and models in ACT-R/E continuously learn as they interact with the world. At ACT-R/E's core is a set of limited-capacity buffers that loosely correspond to working memory; items in the buffers represent what an ACT-R/E model is thinking, such as what it is looking at, or what setting it is in. We chose ACT-R/E because it has successfully provided dynamic cognitive context in a variety of other settings (Anderson and Reder 1999; Hiatt and Trafton 2013; 2015).

In addition to symbolic information (e.g., I am in a kitchen and see an apple), items have a subsymbolic, *spreading activation* value that represents the item's relevance to the current situation. Spreading activation is temporary and sources from the current contents of working memory, allowing items that are the focus of attention to activate related, or *associated*, items for short periods of time (Hiatt and Trafton 2013). At any given time, we consider the set of items that has spreading activation to comprise the current context.

Associations between items can be created for several reasons. Pertinent to our discussion here, they are created between items that are in working memory at roughly the same time. This means that associations are created between objects that are seen in close temporal proximity to one another, as well as between each object and the scene in which it appears.

Once created, associations have a corresponding strength value that affects how much activation is spread along them. Intuitively, an associative strength is a Bayesian-inspired value that reflects how strongly an item in working memory (whether an object the model is looking at or a setting the model knows its in) predicts that an item it is associated with will be seen next. For object recognition, this means that both what a model looked at previously, as well as what scene it is in, can affect its expectations for what it will see next. Thus, objects appearing in common situations (both with other objects they typically appear with, and in a scene they typically are part of) receive the highest amounts of spreading activation, facilitating their recognition. Objects appearing out of place typically receive much lower spreading activation, hindering their recognition. This process occurs incrementally and online, allowing models to have a rich context even with very little experience in the world.

Mathematically, the strength of an association from item j to item i (S_{ji}) is (Hiatt and Trafton 2013):

$$S_{ji} = mas \cdot e^{\frac{-1}{al \cdot R_{ji}}} \tag{2}$$

$$R_{ji} = \frac{f(N_i C_j)}{f(C_j) - f(N_i C_j) + 1} \tag{3}$$

These equations have two parameters: mas, the maximum associative strength; and al, the associative learning rate. The function f tallies the number of times that item j has been in working memory, either independently (C_j) or at similar times to when i has been in working memory ($N_i C_j$). The strengths are thus a function of how often the two items are in working memory at roughly the same time, versus how often each is in working memory without the other.

Finally, we can translate items' spreading activation values to probabilities that each of the items will appear in working memory next. To calculate this probability for an item i, we rely upon the 'softmax' equation provided as part of the cognitive architecture (Anderson et al. 1998):

$$P(i) = \frac{e^{S_i/t}}{\sum_k e^{S_k/t}} \tag{4}$$

where the variable S_i is the spreading activation of item i, $\sum_k$ iterates over the set of all objects (including i), and t equals $0.5 \cdot \sqrt{6}/\pi$. Intuitively, this equation represents each item's proportional share of spreading activation, translated to a probability value.

(a) (b)

Figure 1: Example images from the NYU Depth V2 database that show some examples of the scene contexts and extreme range of pose present in this dataset. (a) An example 'kitchen' scene. (b) From left to right, example objects of class 'bag', 'microwave', 'monitor', 'television', and 'faucet' taken from scene images.

Dynamic Context and Computer Vision

For each object we are trying to recognize, then, we have two probability distributions of what its label could be – the one from baseline object recognition, and the one from dynamic cognitive context. The final step is to multiply each label's probabilities for each label to come up with a final distribution of the object's classification. For example, if object recognition believes that an object's label is

$$\{\text{computer} = 0.5, \text{microwave} = 0.25, \text{tv} = 0.25\},$$

and dynamic context believes that an object's label is

$$\{\text{computer}=0.15, \text{microwave} = 0.7, \text{tv} = 0.15\},$$

then the combined result for the object's label would be

$$\{\text{computer}=0.075, \text{microwave} = 0.175, \text{tv} = 0.0375\}.$$

The overall classification would thus be microwave.

Experiment and Results

To test our approach, we consider the NYU Depth v2 dataset, which contains 1449 scenes that show objects appearing in situ. This dataset is particularly appropriate since it recreates the types of scenes that both autonomous systems and people encounter as they go about in the world; for the same reason, it is also a particularly challenging dataset. Figure 1 shows examples from the dataset. Both depth and RGB images are included in the data; however, we use only the RGB images. Ground-truth labels for each scene (kitchen, office, etc.) and each object are provided by the dataset; each scene contains between 2 and 24 detected objects. After excluding object classes that had too few instances to train and test on, the dataset included 74 object classes. We split this into training and testing sets, resulting in 5819 training objects and 6012 testing objects, appearing in 724 and 725 scenes, respectively. Our baseline object recognition approach was pre-trained on this training data.

Using this dataset, we ran two different studies. The first study was to test how much using dynamic context could help with object recognition. We first collected data on the precision of object recognition when used alone. Then, we measured performance when biased during evaluation by the dynamic cognitive context we have described. Here, the cognitive model is untrained; then, as each object is viewed for evaluation, the cognitive model is first given the current scene label, similar to an autonomous system localizing itself on a map. Based on spreading activation coming from this scene label, as well as what it has recently seen, the cognitive model suggests what it is likely to be currently looking at. Given this, our approach's classification is determined. At that point, the model is provided with the ground truth of the object it is seeing for the purposes of updating its associations; this is analogous to an autonomous system performing online object recognition and being corrected by a human partner when necessary.

Because dynamic context is sensitive to the order in which objects are seen (it affects what objects are associated with one another), during evaluation we varied this order for each scene. The first "seen" object in an image was selected randomly; subsequent objects were sorted greedily by their proximity to the previously seen object. Because of the stochasticity, we report results for dynamic context that are averaged across five different experimental runs.

We also compare our approach against one that uses a more traditional source of context, static local pixel context (as first proposed by Felzenszwalb et al. 2010). Here, the off-the-shelf feature vector is augmented with the center and area of the detection region. This provides useful context about what objects are likely to appear in different areas of scene images (for example, mugs are usually small and midway in images, on tabletops; chairs are usually lower and larger, as they rest on the floor).

We report results in terms of the standard metric of mean average precision (mAP) (Everingham et al. 2007), which measures both our ability to recognize an object as well as minimizing false positives (i.e., average precision). We show the results in Table 1. Object recognition alone re-

	Object recognition (baseline)	Object recognition + local pixel context	Object recognition + cognitive context
mAP	*40.18*	*41.18*	*47.77*

Table 1: Mean average precision results using baseline object recognition and two types of context on the NYU Depth V2 dataset. The highest precision was attained by using dynamic, cognitive context as part of the recognition process.

	Learning during testing	Learning during training	Learning during training and testing
mAP	*47.77*	*48.06*	*47.64*

Table 2: Mean average precision results using CV and cognitive context, and various learning modes on the NYU Depth V2 dataset. Notably, an untrained cognitive model achieves near identical performance as trained cognitive models.

sults in a mAP of 40.18%, and adding static pixel context increases mAP to 41.18%. By far, the best results are produced when object recognition is augmented with dynamic cognitive context, with a mAP of 47.79%, an improvement of 8% over the baseline system with no context.

We next ran a second study to test our claim that cognitive context performs well even without training. To that end, we attained results from several different learning modes of cognitive context: learning associations only during testing (as in the first study); learning associations during training, but not during testing; and learning associations both during training and testing. The results are shown in Table 2. Importantly, all three conditions achieved roughly the same precision, indicating that dynamic cognitive context can effectively provide context and improve object recognition without any prior training.

Conclusions

In this work we have explored the benefits of using dynamic cognitive context to assist with object recognition without any prior training. It learns incrementally and online to improve the mean average precision of object recognition by 8% on the challenging NYU Depth v2 dataset.

This approach has two main advantages in addition to improving object recognition. First, since it requires no training, it avoids the challenge of gathering enough training data, and capturing enough objects in their natural settings, to be useful. Second, it also opens the door for including the human in the loop by providing extra context (e.g., "I saw that cup last week"), or feedback (e.g., "that is not a cup") that can then be incorporated into the cognitive model. Reasoning about the world to this level of depth has the potential to drastically improve the functionality we can expect from our cognitive computer vision systems.

References

Anderson, J. R., and Reder, L. M. 1999. The fan effect: New results and new theories. *Journal of Experimental Psychology: General* 128:186–197.

Anderson, J. R.; Bothell, D.; Lebiere, C.; and Matessa, M. 1998. An integrated theory of list memory. *Journal of Memory and Language* 38(4):341–380.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27.

Divvala, S.; Hoiem, D.; Hays, J.; Efros, A.; and Hebert, M. 2009. An empirical study of context in object detection. In *CVPR*.

Everingham, M.; Gool, L. V.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2007. The PASCAL visual object classes challenge 2007 (VOC2007) results. http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html.

Felzenszwalb, P.; Girshick, R.; McAllester, D.; and Ramanan, D. 2010. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Galleguillos, C.; Rabinovich, A.; and Belongie, S. 2008. Object categorization using co-occurrence, location, and appearance. In *Computer Vision and Pattern Recognition (CVPR)*.

Hiatt, L. M., and Trafton, J. G. 2013. The role of familiarity, priming and perception in similarity judgments. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Hiatt, L. M., and Trafton, J. G. 2015. An activation-based model of routine sequence errors. In *ICCM*.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 1097–1105.

Lawson, W.; Hiatt, L.; and Trafton, J. G. 2014. Leveraging cognitive context for object recognition. In *Proceedings of the Vision Meets Cognition Workshop (held at CVPR)*, 381–386.

Li, C.; Kowdle, A.; Saxena, A.; and Chen, T. 2010. Towards holistic scene understanding: Feedback enabled cascaded classification models. In *NIPS*.

Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.; Lee, S.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. The role of context for object detection and semantic segmentation in the wild. In *CVPR*.

Oliva, A., and Torralba, A. 2007. The role of context in object recognition. *TRENDS in Cognitive Sciences* 11(12):520–527.

Razavian, A. S.; Azizpour, H.; Sullivan, J.; and Carlsson, S. 2014. CNN features off-the-shelf: an astounding baseline for recognition. *Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Schneider, D. W., and Anderson, J. R. 2011. A memory-based model of hick's law. *Cognitive Psychology* 62(3):193–222.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from RGBD images. In *ECCV*.

Trafton, J. G.; Hiatt, L. M.; Harrison, A. M.; Tamborello, II, F. P.; Khemlani, S. S.; and Schultz, A. C. 2013. ACT-R/E: An embodied cognitive architecture for human-robot interaction. *Journal of Human-Robot Interaction* 2(1):30–55.