Leveraging Cognitive Context for Object Recognition

Wallace Lawson, Laura Hiatt, J. Gregory Trafton
Naval Research Laboratory
Washington, DC USA

Abstract

Contextual information can greatly improve both the speed and accuracy of object recognition. Context is most often viewed as a static concept, learned from large image databases. We build upon this concept by exploring cognitive context*, demonstrating how rich dynamic context provided by computational cognitive models can improve object recognition. We demonstrate the use cognitive context to improve recognition using a small database of objects.*

1. Introduction

One of the reasons why humans are so good at recognizing objects in images is that they understand the world using a combination of recognition of visual features and *context*; *e.g.*, surrounding objects, knowledge of their location, etc. [13]. Exploiting context for computer vision approaches, similarly, has been shown to improve both the speed and accuracy of object recognition. Knowledge of a camera's location, for example, can help object recognition distinguish between an alligator and a crocodile, since both typically inhabit different geographic regions of the world. There are many other examples of ways in which context can help [3, 8, 10].

Context in these approaches, however, is largely static, and based purely on features in the external world. While this does produce effective results, it falls short of the rich, dynamic context that people rely upon [13]. In this paper, we begin to close that gap, by using a computational cognitive model as the source of context. We model human context using ACT-R/E, the embodied version of the cognitive architecture ACT-R [18]. It has been extensively verified against many aspects of human cognition, including memory [2, 16] and similarity [9]; this high fidelity to the human mind makes it the ideal candidate for providing the rich, dynamic context that people rely upon as they perceive the world.

Context in ACT-R/E takes the form of associations between related concepts that are learned incrementally over time. Concepts become associated when they are thought about at roughly the same time; the more they are thought about in proximity to each other, the stronger their association becomes. This means that recognition can be assisted by both external context (*e.g.*, I am currently looking at a desk so I may see my computer next), as well as internal (*e.g.*, I was told to look for a banana so I may be likely to see one soon).

We recognize objects using LVis, a biologically plausible model of the visual cortex [14]. LVis is robust in handling the wide variety of poses that can be expected when viewing objects in real-world environments. It's also been demonstrated to correctly identify objects in challenging conditions such as changes in lighting, ambiguity in pose, and partial occlusion. A biologically plausible model also has a well understood structure, and the features of which can be used to infer a confidence in the classification of objects [12].

By combining contextual information from ACT-R/E and perceptual information from LVis, we can dramatically improve recognition in cases where recognition using visual features alone is difficult. Despite the robustness that LVis shows towards pose, lighting, and occlusion, ambiguities between several similar objects are inevitable. Identification in these cases should be stated with a likelihood. By combining these likelihoods with contextual information, we can arrive at the correct answer. For example, when looking in the kitchen, context may suggest related concepts such as apples or lemons. Any ambiguities that might arise from other similar objects (such as a red ball) can be quickly resolved by incorporating contextual information, resulting in the correct identification.

The remainder of this paper is organized as follows. We present related work on using context to recognize objects in Section 2. We present the LVis and ACT-R/E architectures in Section 3. We present the experimental scenario and results in Section 4. Finally, we discuss the experimental results and conclude in Sec-

tion 5.

2. Related Work

Context is used heavily in the human visual system to bias both where we look and what objects we expect to see [13]. Much of the previous work on context has focused on how to learn this context from large databases of images. Shotton et al. [17] used the location of pixels within an image to improve recognition of objects and image segmentation. Rabinovich et al. [15] used *semantic object context* to improve recognition of related objects within a single image (*e.g.*, tennis racket, tennis ball, tennis court). Divvala et al. [5] used object location and size to improve detection of objects in the PASCAL VOC database.

One problem inherent in these domains, yet rarely addressed, is how contextual information can be learned in a new environment. It is possible to encounter a similar type of object in a new location, for example. If the object representation is learned at the same time as context, this would require re-learning about objects each time they were seen in a different type of environment. Our work focuses on the problem of how to learn context online, in a manner that is separated from object representation. This provides a greater level of flexibility, effectively managing the rich dynamics of real-world environments.

3. Methodology

3.1. LVis

The Leabra vision model [14], or LVis, is a biologically plausible neural network with three hidden layers that are organized in a manner similar to the visual cortex (see Figure 1).

LVis's input is an image that has been isolated from the surrounding environment in some manner. In the first hidden layer (V1), the input image is convolved with wavelets tuned to different orientations and scales. V1 is fully connected to the the second hidden layer (V4). V4 learns higher level representations of salient features in the image (e.g., corners, curves). The multiple scales of wavelets in V1 provides a sense of both scale and orientation of salient features. V4 is fully connected to the third hidden layer (IT). Neurophysiological evidence suggests that the brain has *view-specific encoding* [11, 7]. In this scheme, neurons in the IT cortex activate differently depending on how an object appears. In LVis, when an object is viewed in similar poses, the same neurons will activate in the IT layer. When the viewpoint of the object causes a significantly different appearance, a different set of neurons will activate.

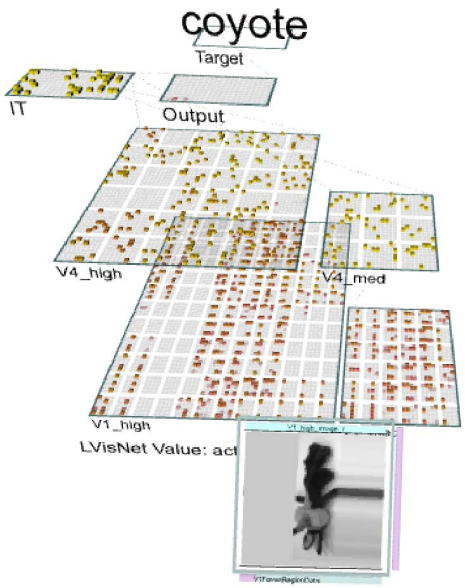

Figure 1. Visual representation of the LVis architecture.

One result of view-specific encoding is that when two objects are difficult to distinguish from each other, a very similar pattern will appear in the IT layer. The IT activations provide both the object class, viewpoint, as well as distinctiveness from other known objects. Intuitively, a small round object can be compared to the class apple. While it may be similar in some respects to an apple (i.e., small and round), it might also look similar to other known small round objects (e.g., lemon). Therefore, while our classification decision might state that we are looking at an apple, there is also a chance that we could be looking at something else.

Unsupervised learning is used to find distinctive viewpoints. The IT activations of the training objects are clustered using k-means clustering [6]. It's important to note that in this scheme, some care must be taken to establish the number of clusters, k, since this is synonymous with the number of distinctive viewpoints. It's difficult to know this value *a priori*, and it's likely that this will change depending on the object in question. A simple object such as a uniformly colored ball would only have one distinctive viewpoint, since a change in viewing angle will never affect appearance. A more complicated object, such as an automobile would likely have a great number of distinctive viewpoints because a change in viewing angle will often times affect the appearance. In practice the appropriate value of k is determined using viewpoint stability, where a small change in viewing angle will generally not result in a new distinctive viewpoint. This number is found by initially setting k to a large number, then decreasing

the value until adequate viewpoint stability has been found.

Confidence comes from the distinctiveness of the viewpoint when compared to other known viewpoints. This is established using a statistical test known as stroud [4]. In stroud, the *strangeness* of a viewpoint (m) measures how "strange" it would be for it to belong to one class, versus all other classes. Formally, strangeness is the ratio of the distance to one viewpoint (c) over the distance to all other known objects and viewpoints c^{-1} (Eq. 1).

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

A distinctive viewpoint has a very low strangeness, since it is very different from anything else that has been observed. Viewpoints that are not distinctive would have a strangeness closer to 1, since it should appear similar to multiple classes.

The probability that viewpoint (s) is distinctive for object (o) can be estimated using leave-out-one cross-validation

$$p(s|o) = \frac{|\forall_{i \in s} str(i,s) \leq str(o,s)|}{|s|} \quad (2)$$

3.2. Recognition

To recognize object o, we use the IT activations from LVis (a), then compare this against known viewpoints. We compute the strangeness to each viewpoint Eq. 1. The probability of recognition is conditionally dependent upon both the distinctiveness of the viewpoint 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 dependent upon the probability that the IT activation patterns belong to the viewpoint (left part of Eq. 3), as well as the distinctiveness of that viewpoint (right part of eq. 3).

$$p(o_{ix}|a_x, s_x) = p(o_{ix}|a_x)p(o_{ix}|s_x) \quad (3)$$

Where the left hand side of the equation, $p(o_{ix}|a_x)$, is the probability that we have observed the object given particular IT activations. If the observed activations are similar to known IT activations for object o_i, we expect the probability to be high. The right hand side of the equation, $p(o_{ix}|s_x)$, can be interpreted as the general confidence of recognizing object o in estimated viewpoint s.

Combining the two (eq. 3) produces a uncertainty measure that accounts for both similarity of IT activations as well as the confidence in the viewpoints. Not surprisingly, this uncertainty measure becomes more reliable when given a greater amount of object classes, and instances per class.

3.3. ACT-R/E

We model context using the cognitive architecture ACT-R/E [18]. At a high level, ACT-R/E is an integrated, production-based system. At its core is a set of limited-capacity buffers that loosely correspond to working memory; they indicate what the model is thinking, including what the model is looking at and its current goal representation. At any given time, there is a set of *productions* (if-then rules) that may fire because their preconditions are satisfied by the current contents of the buffers. The fired production can either change its internal state (e.g., by creating new knowledge) or its physical one (e.g., by pressing a key on a keyboard).

Buffer contents take the forms of "chunks", or facts. In addition to symbolic information (*e.g.*, I ate an apple), chunks have a subsymbolic, *spreading activation* value that represents the chunk's relevance to the current situation. Spreading activation is temporary and sources from the current contents of the buffers, allowing chunks that are the focus of attention to activate related, or similar, memories for short periods of time [9].

3.4. Context in ACT-R/E

Context in ACT-R/E relies on ACT-R/E's spreading activation mechanisms. Activation is spread between related concepts via *links* between those concepts [1]. Links are directional, and are created between chunks when they are thought about in close temporal proximity. For example, if chunk j is in a buffer, and chunk i follows it in the buffer, chunk j will prime chunk i. Concepts also prime themselves, since one typically thinks about a concept over a period of time. This link structure is inherently sequential; this allows the model to learn from experience which chunks are related in a truly sequential way, in which case there will be links between them in only one direction, and which chunks typically just co-occur, in which case there will be links between them in both directions. There are other ways of creating links, as well, but we do not utilize them in this model and so forgo their discussion.

Once established, links have an associated strength value which affects how much activation is passed along the link from chunk j to chunk i. Link strengths, intuitively, reflect the probability that chunk i will be needed or will be thought about given that chunk j is currently in working memory; the strength is updated iteratively whenever the model thinks about chunks i

and j (*i.e.*, has chunks i and j in working memory), whether alone or together.

We define the chunks in working memory plus their outgoing links as the current *context*, which spreads activation to chunks with incoming connections to those links. Context, then, can come from both external, environmental sources as well as internal, goal-oriented or introspective sources, depending on what the model is thinking about and looking at. At any given time, this rich, dynamic context causes associated concepts to be primed, indicating that these concepts are likely to be relevant to the current context and, as we will show, can help perception interpret what it sees in the world.

Given a set of concepts that are primed, ACT-R/E can then translate their spreading activation values to probabilities that each object will be needed in working memory next. To calculate this probability, we adapt the cognitively-plausible ACT-R equation that calculates the probability that a chunk i will be retrieved given its total activation values to account for spreading activation only [2]:

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

where the variable S_i is the spreading activation of chunk i, $\sum_k$ iterates over the set of all concepts (including i), and t equals $0.5 \cdot \sqrt{6}/\pi$. For this paper, we limit the set of all concepts considered by $\sum_k$ to be those concepts that can be seen (*e.g.*, an apple, not the model's goal).

3.5. Biases Perception with Context

The probability of each object from LVis (Eq.3) and cognitive context (Eq. 4) are combined using decision level fusion. In this case, context is treated as a prior probability when computing the probability of each object.

4. Experimental Results

In this section, we experimentally verify the benefits of using context to improve the object recognition system.

4.1. Training

We train LVis using a set of 6 objects (apple, banana, raisins, coyote, lemon, wire) learned as they are encountered upon a table. During training, a robot approaches the table from a distance of approximately 1m, typically producing about 20 images per object. While training LVis, each image is synthetically rotated and scaled to show examples of how the object might appear at different angles and distances.

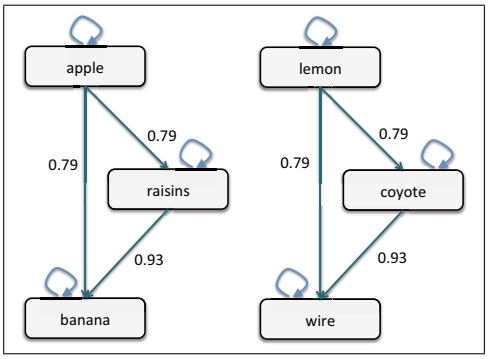

Figure 2. Connections between the recognized objects, with example link strengths.

Concurrently, we expose the robot to each of the objects across two different environments; the robot has no prior knowledge of any objects or any potential associations between them. First, the robot looks around a kitchen. There, it sees the apple, raisins, and banana. Although it did not explicitly understand the semantics of where it was, it created links between the newly created *apple*, *raisins*, and *banana* concepts. Based on the order in which the model perceives the objects, the links are created such that apple primes raisins and banana, and raisins primes banana. Next, the robot went to our laboratory, where it sees a plastic *lemon*, a stuffed *coyote*, and a *wire*. These new concepts also became linked because the robot saw them sequentially. Recall that each concept also primes itself. The robot's contextual network after these two training scenarios is shown in Figure 2.

4.2. Testing and Results

To test the system, the robot sees the six objects, divided into the same two sets as before (apple, raisins and banana vs. lemon, coyote and wire). In this case, however, the robot views the objects from a relatively close distance. This was an intentionally difficult scenario for the object recognition images, since the testing images will be at least a little bit different than the training images. Each object is evaluated 5 times, for a total of 30 different evaluations across the 6 different objects.

The robot performs this test both with and without cognitive context. Without context, the robot relied on the results from LVis alone. With context, the robot incorporates context into the object recognition process. Initially, context indicates that all objects have an equal likelihood of being seen. Once LVis determines that the robot is looking at an apple, however,

context suggests that an apple is most likely to be seen next (since it primes itself); raisins and banana are tied for the second-most likely object to be seen next. When raisins are seen, context again suggests that raisins are most likely to be seen next, with banana the second-most likely to be seen next. This pattern repeats itself for the second set of objects.

Numerically, LVis alone correctly recognizes the objects 90% of the time, as shown in Table 1. Two observations can be made from these results. First, LVis recognizes objects quite well, and the recognition results in this case are consistent with the recognition rates that have been observed in previous works [12, 14]. Second, we see a consistent ambiguity between two classes of objects, in this case confusing the coyote and the raisins.

	T1	T2	T3	T4	T5	# correct
apple	A	A	A	A	A	5
raisins	R	R	R	R	A	4
banana	B	B	B	B	B	5
lemon	L	L	L	L	L	5
coyote	R	C	R	C	C	3
wire	W	W	W	W	W	5

Table 1. Results of classifications from LVis without the benefit of cognitive context. T1-T5 are the 5 test evaluations. A = apple, R = raisins, etc.

We used the same set of 30 images to recognize the objects when LVis is given the benefit of cognitive context. Doing so improves the recognition rate to 100% (Table 2). Interestingly, context plays two main roles when correcting recognition errors. We had foreseen that context would primarily be useful in suggesting which objects are likely to be viewed next, as it is when it corrects the first recognition error of coyote: given that the model was looking at a lemon, raisins was very unlikely to be seen next, whereas coyote was likely to be seen next. Additionally, however, since objects prime themselves, context also provided a sort of "visual inertia" to recognition. This leads to an improvement in the errors that occur while still looking at the same object, as happened with raisins and coyote.

5. Discussion

Overall, the experiment showed a large benefit to utilizing cognitive context while performing object recognition. While we were very pleased with the results in this paper, it also emphasized some basic, difficult research questions we will need to answer before our theory of context and perception is complete. In this section we discuss a few of these questions before concluding the paper.

	T1	T2	T3	T4	T5	# correct
apple	A	A	A	A	A	5
raisins	R	R	R	R	R	5
banana	B	B	B	B	B	5
lemon	L	L	L	L	L	5
coyote	C	C	C	C	C	5
wire	W	W	W	W	W	5

Table 2. Results of classifications from LVis with cognitive context from ACT-R/E. T1-T5 are the 5 test evaluations. A = apple, R = raisins, etc.

5.1. Perception

Combining context and perception at the decision level works best when both modalities fail gracefully. In perception, the correct object class should always have a reasonable likelihood of being observed. This was one area of difficulty in our experimental results. At times, LVis can produce very low likelihoods for the correct class, while producing high likelihoods for a different (incorrect) class.

Recall that earlier we mentioned that a wide variety of objects leads to more accurate probabilistic information. With a relatively small set of objects it may be difficult to accurately state the probability of recognition. It's possible that the robot could be initially trained using objects from image databases, which are then extended as the robot encounters new objects (or new instances of existing objects) in the environment. This would also have the added benefit of providing additional training examples, which will yield additional data that could be used to train LVis.

5.2. Context

For similar reasons, context also needs to be further investigated in order to be more broadly useful to object recognition. An in-depth analysis of our results showed that our results were highly order-dependent. If the error trials for recognizing coyote, for example, had been switched, context would not have successfully helped LVis recognize coyote during the first trial and instead would have recognized it as raisins. Once that error was initially made, then, it would have only been reinforced as the model began to think about the raisins, making it hard for LVis to counteract context's bias and for the model to successfully "see" coyote. One area of future work, then, is to be able to detect, and counteract, this type of self-fulfilling prophecy.

We also found suggestions that the overall results are fairly dependent on our specific goal structure. For example, if we tweaked the model to think about each object it sees for a longer period of time, or to add

different representations of it to working memory (*e.g.* visual representation, conceptual representation, affordances, etc.), the structure and weights (such as what is shown in Figure 2) of spreading activation would change and provide different results. While in this case our approach would still have performed well, it is not hard to think of other examples where these tweaks would have caused problems with recognition. In the near future, we plan on extending this work to other types of models, objects, and context in order to investigate more general ways of interpreting context that will lead to more consistency across domains.

5.3. Conclusions

In this paper, we have shown a system that uses cognitive context to improve machine perception. It learns online about the context surrounding two sets of objects it has recognized in the past to appropriately bias computer vision to correctly identify objects, improving accuracy from 90% to 100% for a limited set of objects.

Blending computer vision and cognitive context in this way opens the door for not only improving object recognition, but also performing more sophisticated reasoning over what an agent sees in the world. Given the current context, for example, different top-down knowledge could be used to interpret what an agent sees in the world. If an agent knows that its teammate has the goal of getting into a locked room, the goal of kicking the door open will be primed by the current context; this may then facilitate the recognition and interpretation of the teammates leg in the air. Reasoning about the world to this level of depth has the potential to drastically improve the functionality we can reasonably expect from our cognitive computer vision systems.

Acknowledgements

This work was supported by the Office of Naval Research (GT) and the Office of the Secretary of Defense / Assistant Secretary of Defense for Research and Engineering (LH). The views and conclusions contained in this paper do not represent the official policies of the U.S. Navy.

References

[1] J. R. Anderson. A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22(3):261–295, 1983. 3

[2] J. R. Anderson, D. Bothell, C. Lebiere, and M. Matessa. An integrated theory of list memory. *Journal of Memory and Language*, 38(4):341–380, 1998. 1, 4

[3] M. E. Auckland, K. R. Cave, and N. Donnelly. Nontarget objects can influence perceptual processes during object recognition. *Psychonomic Bulletin & Review*, 14(2), 2007. 1

[4] D. Barbara, C. Domeniconi, and J. Rodgers. Detecting outliers using transduction and statistical testing. *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006. 3

[5] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert. An empirical study of context in object detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, June 2009. 2

[6] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley Interscience, 2000. 2

[7] M. Frank, Y. Munakata, T. Hazy, and R. O'Reilly. *Computational Cognitive Neuroscience*. 2012. 2

[8] R. D. Gordon. Attentional allocation during the perception of scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4), 2004. 1

[9] L. M. Hiatt and J. G. Trafton. The role of familiarity, priming and perception in similarity judgments. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society (CogSci)*, pages 573–578, 2013. 1, 3

[10] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *International Journal of Computer Vision*, 80(1), 2008. 1

[11] T. C. Kietzmann, S. Lange, and M. Riedmiller. "computational object recognition: A biologically motivated approach". *Biol. Cybern.*, 2009. 2

[12] W. Lawson and J. Trafton. Unposed object recognition using an active approach. *International Conference on Computer Vision Theory and Applications*, 2013. 1, 5

[13] A. Oliva and A. Torralba. The role of context in object recognition. *TRENDS in Cognitive Sciences*, 11(12):520–527, 2007. 1, 2

[14] R. C. O'Reilly, D. Wyatte, S. Herd, B. Mingus, and D. J. Jilk. Recurrent processing during object recognition. *Frontiers in Psychology*, 4(124), 2013. 1, 2, 5

[15] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, Oct 2007. 2

[16] D. W. Schneider and J. R. Anderson. A memory-based model of hick's law. *Cognitive Psychology*, 62(3):193–222, 2011. 1

[17] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*. Springer Berlin Heidelberg, 2006. 2

[18] J. G. Trafton, L. M. Hiatt, A. M. Harrison, F. P. Tamborello, II, S. S. Khemlani, and A. C. Schultz. ACT-R/E: An embodied cognitive architecture for human-robot interaction. *Journal of Human-Robot Interaction*, 2(1):30–55, 2013. 1, 3