Evolving a Neural Network Active Vision System for Object Recognition

Derek James and Philip Tucker

Abstract

Previous research has demonstrated the potential for neural network controlled active vision systems for shape discrimination and object recognition. However, this approach has not been very well explored, and previous implementations of such systems have been somewhat limited in scope. We present an evolved neural network based active vision system that is able to move about a 2D surface in any direction, along with the ability to zoom and rotate. We demonstrate that a system with such features is able to correctly classify shapes presented to it, despite variance in location, scale, and rotation, and that contrary to our initial assumptions, effective identification is possible even without the full range of navigational features.

Keywords

Neuroevolution, Active Vision, Object Recognition

1. Introduction

Traditional approaches to pattern recognition tasks usually involve highly domain-specific algorithms involving statistical analysis, but recently more biologically-inspired approaches, such as active vision, have begun to develop.

Active vision refers to the process of exploring an image or scene for relevant features, just as biological organisms do. The advantages of such a system are obvious, including attentive focus which excludes processing of areas of the image that are irrelevant, and seemingly providing an elegant method of handling variance in location, scale, and rotation.

Control of an active vision system could be implemented in a variety of ways, but artificial neural networks are an appealing choice because they are more biologically-inspired, and have demonstrated success in both noisy control and pattern recognition tasks. Thus, it seems natural to apply neural networks to an integrated system capable of exploring a scene, locating relevant features, and making determinations based on the information it receives as input.

Kato and Floreano [1] implemented such a system, evolving the connection weights of a recurrent neural network with no hidden units for a controller that explored a noisy grayscale image containing either an isosceles triangle or a square, and identified which object the scene contained based on one of two output values. The objects varied in both scale and location, but not in rotation, an important transformation found in most pattern recognition tasks.

Stanley et al [2] used a similar approach to view and play the board game Go. A 5x5 viewing window controlled via an evolved neural network was given a fixed number of time steps to explore a game board and express a move preference via a given output. The system demonstrated the ability, on small boards, to beat GNU Go, an open source Go-playing algorithm of reasonably high skill (compared to other existing algorithms). The same principles apply as in the research mentioned above, in that active vision allows the system to focus on relevant aspects of the presented surface, whether a 2D image or a game board.

We present a system that expands upon previous approaches and explores the basic paradigm further. Our system consists of an artificial retina capable of processing any 2D surface by panning left, right, up, or down, zooming in and out, and rotating. It is controlled via a recurrent artificial neural network, evolved using a modified version of the NEAT (NeuroEvolution of Augmenting Topologies) methodology, and is applied to a basic object recognition task.

2. Experimental Details

2.1. The Active Vision System

The active vision system consists of a framework for feeding pixel values from a 2-dimensional surface into a recurrent neural network. The receptive field, or artificial retina, is a square region composed of cells, or receptive areas, that read the pixel value from the surface. All experiments mentioned here used a 5x5 retina.

Just as in [1], the retina is able to move across the image vertically and horizontally, as well as zooming in and out. Unlike that system, this one includes the ability to rotate.

All images used were grayscale tiffs, so each pixel contained a value between 0 and 255. These values are input into a recurrent neural network, along with the retina's current orientation, consisting of its x and y position, angle of rotation, and zoom factor. The ratio of evaluation time remaining to total time allocated (i.e., an "hourglass"), is also input. A constant value of 1, or bias, is the final input.

Each time step, neural net outputs are used to update the position and orientation of the retina. The specific movement outputs include change in horizontal location (Δx), change in vertical location (Δy), change in rotation ($\Delta \theta$), and change in zoom (Δz). The fifth output, affinity, represents the confidence the network has that the image contains the target shape.

The network architecture is shown in Figure 1.

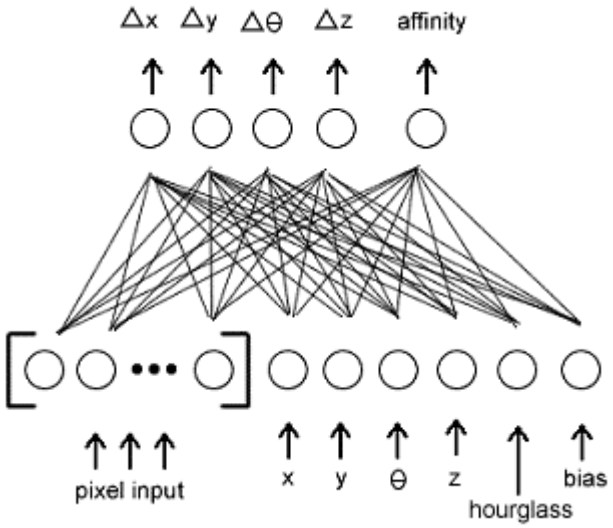


Figure 1. Active vision neural network initial architecture.

The maximum change for location each time step is 20 pixels in any given direction. The maximum change for rotation is 3.6 degrees clockwise or counterclockwise, and the maximum change in zoom is 1%.

To provide evolutionary pressure for efficient neural networks (i.e. to mitigate bloat) the number of time steps provided each network is a product of its complexity. Thus, smaller networks are allocated more time steps than larger networks. The normalization is such that each network should use approximately the same number of CPU cycles to process an image fully.

To determine the network's final judgment regarding the target shape, a weighted sum of affinity values for all time steps is calculated, according to the following equation:

$$\frac{\sum_{i=1}^{n} (\text{affinity}_i \cdot i^2)}{\sum_{i=1}^{n} i^2}$$

Figure 2. Final affinity calculation. (n = number of steps, affinity_i = affinity at step i).

Affinities at later time steps are weighted more than those from earlier time steps, and the net result ranges between 0 and 1.

Each target shape had an associated target range, 0.0-0.2 for a mismatch and 0.8-1.0 for a match. A weighted affinity value within the target range had an error of 0.0. Otherwise, its error was the distance to the inner edge of the range (0.2 for false and 0.8 for true). The total error of the network was the sum of errors for all shapes presented for evaluation. To calculate fitness, this error was subtracted from the maximum possible total error, and the result was then squared.

2.2. NEAT

The algorithm used to evolve the neural network architectures was NEAT (NeuroEvolution of Augmenting Topologies) [3], a methodology that evolves both the weights and architecture of the neural networks controlling the active vision system.

NEAT is distinguished by a disciplined method for crossover, and the use of speciation to divide the population into morphologically similar subgroups. The algorithm has demonstrated the ability to outperform other neuroevolutionary approaches, and perform well at a variety of tasks [3, 4].

The version of NEAT used here was an open source version, ANJI [http://anji.sourceforge.net/], a modified version written in Java, actively maintained by the authors.

Per the NEAT paradigm, the initial neural network architecture consisted of only input and output nodes, fully connected with only feed-forward connections. Initial weight values were taken from a normal distribution between 0 and 1. All input nodes were linear, and all other nodes in the network used the tanh activation function.

Each generation, upon receiving a fitness score as mentioned above, the best performing 20% of the population was selected for survival and reproduction. For all experiments, a population

size of 100 was used, so after selection there were always 20 surviving individuals. The population was then replenished back to 100 individuals: the 20 survivors, plus 20 mutated versions of those survivors, plus 60 "offspring" the result of both crossover and mutation.

The three mutations in standard NEAT are 1) mutate connection weight, 2) add new connection, and 3) add new node. ANJI adds a fourth, 4) remove connection, to combine both simplification and complexification dynamics to the search. Mutations in ANJI are handled differently than in standard NEAT. In standard NEAT, a mutation rate indicates the probability that a particular individual will be mutated (e.g., an add connection mutation rate of 0.03 with a population of 100 would mean that 3 individuals per generation would receive a new connection). In our implementation, a topological mutation rate indicates the probability that a new topological feature will be added among all locations where such a mutation would be possible (e.g., if in the entire population there are 10,000 possible locations where an add connection mutation could occur, a 0.03 mutation rate would result in 300 new connections in the population).

The parameters for the NEAT algorithm used in all experiments are listed in Table 1.

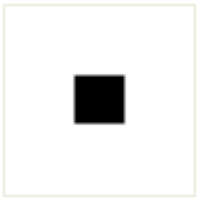
Parameters	Value
Population size	100
Number of generations	300
Weight mutation rate	0.75
Survival rate	0.2
Excess gene compatibility coefficient	1.0
Disjoint gene compatibility coefficient	1.0
Common weight compatibility coefficient	0.4
Speciation threshold	0.2
Add connection mutation rate	0.002
Add neuron mutation rate	0.001
Delete connection mutation rate	0.005

Table 1. Parameters for NEAT, the genetic algorithm used in all experiments.

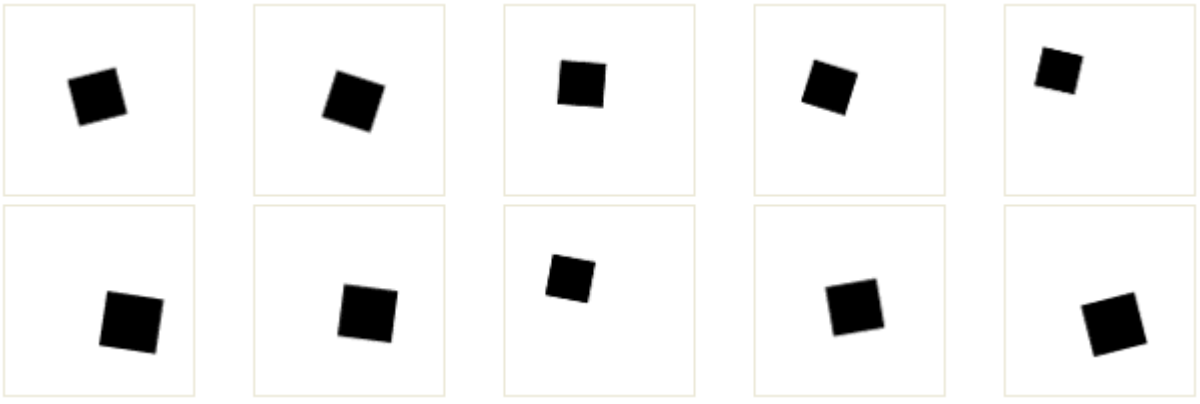
2.3. The Object Recognition Task

Three distinct shapes were used for these experiments: a square, circle, and equilateral triangle. All images were grayscale tiffs, with pixel values ranging between 0 and 255. All images were 100 pixels square, and each of the shapes were 30 pixels across at their widest points. The shapes were black and the backgrounds white. For both evolution and evaluation, the shapes were randomly varied according to the following parameters: their center points were translated randomly up to 20 pixels along the x and y axes; they were scaled randomly up to 20% larger or down to 20% smaller; and they were randomly rotated between 20 degrees clockwise and counterclockwise.

Figure 3a shows an original image, and 3b shows 10 typical randomizations.



a) Original image.



b) 10 typical random transformations.

Figure 3. Random transformation of shapes for evaluation.

The active vision system began all evaluations fully zoomed out and snapped to the edges of the canvas, and was not allowed to zoom out further. It was allowed to zoom as small as a 1-to-1 pixel ratio, and its center point was inhibited from moving off the canvas. Grayscale pixel values were scaled to values between 0 and 1, with 0 being white and 1 black. Off-canvas input was read as –1.

Nearest neighbor interpolation was used for pixel sampling (??? Definition reference ???). Figure 4 shows an image being viewed by the active vision system, and how the region covered by the active vision system is interpreted into pixel values for input into the neural network.

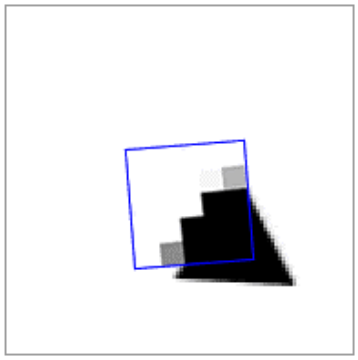


Figure 4. Pixel sampling of area viewed by active vision system.

The system viewed 10 of each shape to be matched, and 10 of each non-matching shape (e.g., when selecting for squares, it viewed 10 squares, 10 triangles, and 10 circles), for a total of 30 evaluations. Images were viewed in random order.

3. Results

The evaluation set for each individual each generation was 30 (10 matches and 20 mismatches). At the end of each run, the best performer from the last generation was evaluated with a larger test set, randomized as mentioned in section 2.3, for a total of 1500 images (500

matches and 1000 mismatches). A weighted affinity value of >= 0.5 indicated a positive match, and < 0.5 indicated a mismatch.

The results of these evaluations are presented in Table 2.

	S1	S2	S3	C1	C2	C3	T1	T2	T3
True positives	377	397	126	16	480	446	474	403	403
False positives	89	70	346	74	100	380	16	136	121
True negatives	911	930	654	926	900	620	984	864	879
False negatives	123	103	374	484	20	54	26	97	97
Overall match rate	**85.87%**	**88.47%**	**52.00%**	**62.80%**	**92.00%**	**71.07%**	**97.20%**	**84.47%**	**85.47%**

Table 2. Test results for normal runs. (S1 = Square 1, C1 = Circle 1, and T1 = Triangle 1).

A match rate of 66.66% could be achieved by a network simply outputing a negative response for all images, since the match/mismatch ratio is 1:2. Though match rates for successful runs did not reach 100%, they were well above the threshold for chance.

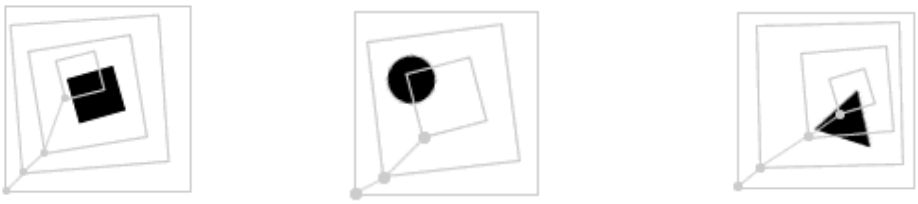
Table 3 shows the results for the two sets of ablation runs, each identifying squares. In the first three, the rotation was disabled for the active vision system. For the other three, zoom was disabled, with the retina centered and completely zoomed in.

	S1-NR	S2-NR	S3-NR	S1-NZ	S2-NZ	S3-NZ
True positives	443	321	453	10	347	387
False positives	86	54	167	4	50	52
True negatives	914	946	833	996	950	948
False negatives	57	179	47	490	153	113
Overall match rate	**90.47%**	**84.47%**	**85.73%**	**67.07%**	**86.47%**	**89.00%**

Table 3. Test results for ablation runs. (S1-NR = Square 1, No Rotation, S1-NZ = Square 1, No Zoom).

When compared with the square runs with all navigational features intact, the ablation tests perform comparably, and in some cases better.

The viewing strategies of the successful active vision systems varied significantly. Figure 5 shows the viewing strategy for the champion of the second square run.



a) Viewing a square b) Viewing a circle c) Viewing a triangle

Figure 5. Behavior of the best evolved active vision system for identifying squares.

As mentioned previously, the artificial retina begins fully zoomed out and centered on the canvas. For this particular individual, the retina always zoomed and moved into a position to sample the edge of the object, rotating counterclockwise. When identifying a square, the retina generally zoomed most of the way in, then panned along a single edge. For a circle, it behaved similarly, zooming in, then sliding along the circle's edge, usually sampling about a 90 degree arc. And with the triangle, the retina usually zoomed in, then concentrated mostly on a corner.

The behavior of the various systems is somewhat difficult to generalize, but nearly all successful individuals with the full range of navigational features tended to zoom most or all of the way in and move along an edge of the shape.

Successful individuals in the no-rotation runs exhibit strategies similar to those of individuals with the full range of navigational features. They typically zoomed nearly all the way in while moving to the edge of the shape, before moving along its edge.

Successful individuals in the no-zoom runs began fully zoomed in. If the shape was within the retina's field of view, it would move outward until finding the edge of the shape, then slide along it similar to other strategies. In some cases, the shape was not initially in the retina's field of view. In these cases, the active vision system searched for the shape by moving towards a particular corner. One champion with zoom disabled always drifted toward the upper left corner when searching; another always drifted toward the lower right corner. Upon reaching the far corners, these individuals simply got stuck. They did not move out of the corner to continue the

search. And yet, such individuals were able to perform comparably to systems with the full range of navigational features, even though there was a small percentage of test cases in which the shape was never actually viewed.

4. Discussion

The results suggest that a neural network-based active vision approach to object recognition could be fruitful. Although not all of the runs achieved successful match percentages, and those that did were not completely successful, the results were promising. There remain a large number of avenues to explore including variance in evolutionary parameters (e.g., mutation rates and survival rate), neural network parameters (e.g., initial topology and activation functions) and imaging parameters (e.g., retina size and resolution, alternative methods of navigational control for the active vision system, and different pixel sampling methods).

Whether or not this approach can lead to very strong computer vision systems for real-world tasks remains to be seen. In addition, these representations may serve as useful models for understanding how biological visual and neural systems function.

Our primary intent was to implement a neural network-based active vision system similar to those in previous research, but with additional features that built upon earlier models. Thus, we implemented zoom capabilities that were much smoother, allowing the active vision system to zoom along a continuum, rather than simply toggling between discrete zoom factors. And we implemented rotation, predicting that this would facilitate better performance when trying to recognize shapes that had been rotated.

The most surprising result was that in both sets of ablation runs, individuals without the full range of features were able to evolve to perform as well or better than those with all features intact. Rotational variance does not affect the appearance of circles, but it does affect both triangles and squares. Successful individuals were adequately able to sample enough

information from the edges and corners of both scaled and rotated shapes to make accurate identifications in the vast majority of cases.

This seems to indicate that such active vision systems are highly robust and may not be very tightly coupled to a specific design or representation. Moving forward, it seems that the most important aspects of such a system are the ability to filter and constrain relevant input into the neural network, thereby increasing the efficiency of the search and identification. It is also important to evolve the neural network controller in the context of an efficient, robust genetic algorithm.

5. Conclusion

The experiments in this paper have demonstrated the efficacy of an active vision system controlled via a recurrent neural network in performing simple object recognition tasks to a fair degree of reliability. They have also demonstrated that, contrary to initial assumptions, individuals evolved with limited navigational control are still able to perform comparably to individuals with a wider range of navigational features.

These preliminary results also suggest that this biologically-inspired approach to actively viewing scenes should be further explored.

References

1. Floreano, D., Kato, T., Marocco, D. and Sauser, E. (2004) "Coevolution of active vision and feature selection", *Biological Cybernetics*, 90(3), 218-228, Springer-Verlag.
2. Stanley K.O., and Miikkulainen, R., "Evolving A Roving Eye For Go", *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2004)*. New York, NY: Springer-Verlag, 2004.
3. Stanley K.O., and Miikkulainen, R., "Evolving Neural Networks Through Augmenting Topologies," Evolutionary Computation, 10 (2), 99-127.
4. Stanley K.O., and Miikkulainen, R., "Competitive Coevolution Through Evolutionary Complexification," *Journal of Artificial Intelligence Research, 21*, 63-100.

List of Figures and Tables

Figures

Tables