Evolving a Neural Network Active Vision System for Shape Discrimination

Derek James
Evolutionary NeuroSystems
5839 Goliad Ave.
Dallas, TX
214-887-0203
djames@gmail.com

Philip Tucker
Evolutionary NeuroSystems
5839 Goliad Ave.
Dallas, TX
214-887-0203
ptucker@gmail.com

ABSTRACT
Previous research has demonstrated the potential for neural network controlled active vision systems to solve shape discrimination and object recognition tasks. 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 can correctly classify shapes presented to it, despite variance in location, scale, and rotation. And, contrary to our initial assumptions, effective discrimination is actually improved when the ability to rotate is disabled.

Keywords
Neuroevolution, Active Vision, Object Recognition

1. INTRODUCTION
Traditional approaches to pattern recognition tasks usually involve highly domain-specific algorithms involving statistical analysis [1], 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 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 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.

Floreano et al [2] implemented such a system, evolving the connection weights of a recurrent neural network with fixed topology for a controller that explored a noisy grayscale image containing either an isosceles triangle or a square. The system 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, a transformation found in most pattern recognition tasks.

Stanley et al [3] 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.

In a virtual aquatic environment, Terzopoulos et al [4] equipped artificial fish with active vision systems with similar functionality that were able to exhibit complex behavior such as tracking other objects in the environment.

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 shape discrimination task.

2. EXPERIMENTAL DETAILS

2.1 The Active Vision System
The active vision system consists of a framework for feeding a 2-dimensional image into a recurrent artificial neural network and allowing that network to scan the image. The receptive field, or artificial retina, is a square region composed of cells, or receptors, that read pixel values from the surface. All experiments use a 5x5 retina.

Just as in [2], 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 are in grayscale TIFF format. Each pixel contains a value between 0 and 255. These are scaled to values between 0 and 1 and input to the neural network (so, our 5x5 retina receives 25 pixel inputs). Any portion of the retina that wanders past the image boundary receives pixel inputs of –1.

Also input to the neural network is the retina's current orientation, an "hourglass", and a bias. The orientation consists of the retina's x and y position, angle of rotation, and zoom factor. The active vision system is allotted a certain number of steps for each image evaluation, and the "hourglass" input is the ratio of steps remaining to total steps allocated. The bias input is a constant value of 1.

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 (ΔΘ), 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 below.

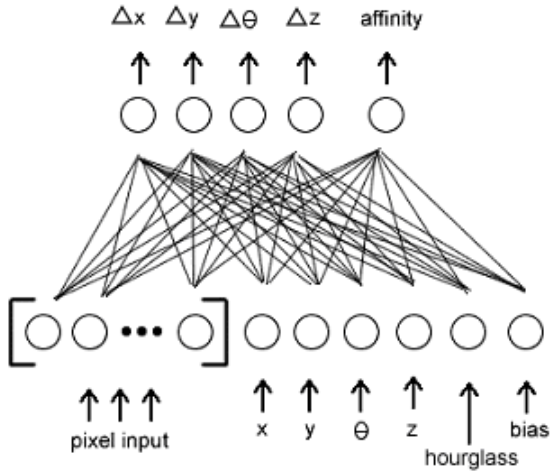


Figure 1: Active vision neural network initial architecture.

The activation function for the output neurons is a modified hyperbolic tangent (tanh). Traditional tanh plateaus at –1 and 1, which seems to make movement control difficult for the system; i.e., it is difficult for it to stay still. The modified tanh (dubbed "tanh-cubic" since it raises the input to the power of 3 as part of the function) adds a plateau at 0. The graph below compares tanh to tanh-cubic.

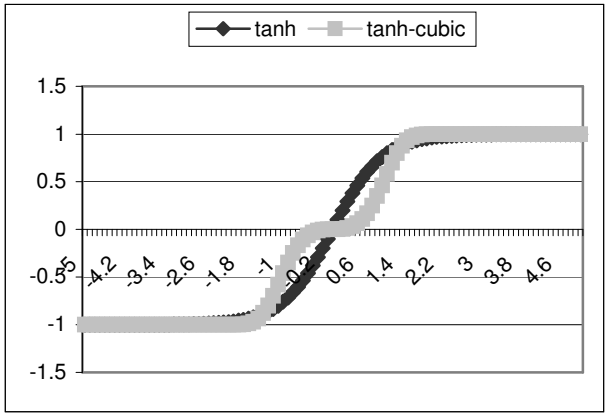


Figure 2: Activation functions.

Each output ranges from –1 to 1. For movement controls (Δx, Δy, ΔΘ, and Δz) the neuron's output value is multiplied by the maximum delta for that value. The maximum change for location (x and y) each time step is +/-20 pixels. The maximum change for rotation is 3.6 degrees clockwise or counterclockwise, and the maximum change in zoom is +/-1%. The affinity output is scaled to a confidence value between 0 and 1.

The network's final judgment regarding the target shape is a product of this affinity, with greater weight given to affinity responses nearer the end of the evaluation. The weighted sum of affinity values is given in the equation below.

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

Figure 3: Final affinity calculation.
(n ≡ number of steps, affinity_i ≡ affinity at step i).

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.

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 is the distance to the inner edge of the range (0.2 for false and 0.8 for true). The total error of the network is the sum of errors for all shapes presented for evaluation. To calculate fitness, this error is subtracted from the maximum possible total error, and the result is then squared.

2.2 NEAT

The algorithm used to evolve the neural network architectures was NEAT (NeuroEvolution of Augmenting Topologies) [5], a methodology that evolves both the weights and architecture of the neural networks controlling the active vision systems.

NEAT is distinguished by allowing crossover between networks with different topologies. Also, NEAT uses 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, 5, 6].

The version of NEAT used here was an open source version, ANJI [http://anji.sourceforge.net/], written in Java and 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 uniform distribution between -1 and 1. Input node

activation functions were linear, output nodes tanh-cubic, and hidden nodes tanh.

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 survivors. 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 or removed 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 roughly 300 new connections in the population).

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

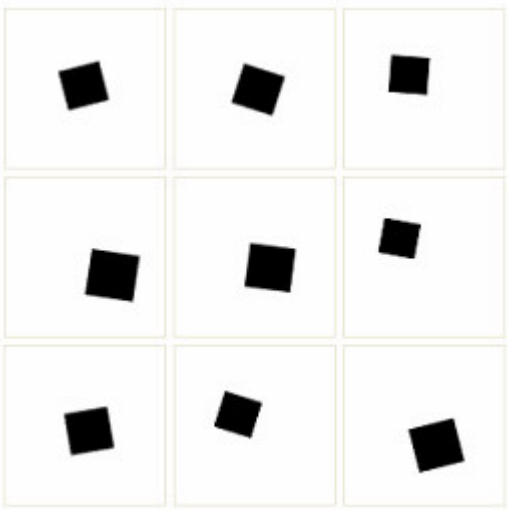
Table 1. Parameters for genetic algorithm.

Parameter	Value
Population size	100
Number of generations	500
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
Remove connection mutation rate	0.005

2.3 The Object Recognition Task

Three distinct shapes were used for these experiments: a square, a circle, and an 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 up to 20 degrees clockwise or counterclockwise. The following figures show an original image and typical randomizations.

a) Original image.



b) 9 typical random transformations.

Figure 4: 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 1:1 ratio of image pixel to retina receptor, and its center point was inhibited from moving off the canvas.

Nearest neighbor interpolation was used for pixel sampling. This means that for a zoomed-out retina the value input for each receptor was the value of the centermost pixel in that receptor's viewing area. This is much more crude than area averaging interpolation, which would compute the average value of all pixels in the receptor area. But, nearest neighbor is much less computationally expensive, and experimentation showed that it did not significantly hurt fitness.

The following figure 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.

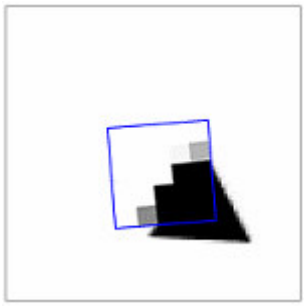


Figure 5: Pixel sampling for artificial retina.

We sought to evolve a "square recognizer" active vision system. Each generation 10 randomized versions of each shape were

generated, for a total of 30 images, and each individual was presented all 30 images in random order. The squares were "match" images, the triangles and circles "mismatch" images. Those individuals responding with high affinity for squares and low affinity for triangles and circles received higher fitness.

3. RESULTS

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

Intially, three runs were performed in which the active vision system had all navigational features enabled. The results of these evaluations are presented in Table 2.

Run	True Pos	False Pos	True Neg	False Neg	Overall Match Rate
1	442	1	999	58	96.07%
2	445	0	1000	55	96.34%
3	495	15	985	5	98.67%

Table 2. Evaluations of champions from runs with all navigational features enabled.

Table 3 shows the results for champions from a set of ablation runs in which the active vision system's ability to rotate was disabled.

Run	True Pos	False Pos	True Neg	False Neg	Overall Match Rate
4	499	5	995	1	99.60%
5	497	2	998	3	99.67%
6	497	1	999	3	99.73%

Table 3. Evaluations of champions from runs with rotation disabled.

Table 4 shows the results for champions from a set of ablation runs in which the active vision system's ability to zoom was disabled. The retina was zoomed completely out, giving it a low-resolution view of the full image canvas.

Run	True Pos	False Pos	True Neg	False Neg	Overall Match Rate
7	299	41	959	201	83.87%
8	304	19	981	196	85.67%
9	243	80	920	257	77.53%

Table 4. Evaluations of champions from runs with zoom disabled, fully zoomed out.

Table 5 shows the results for champions from a set of runs in which the ability to zoom was disabled, but the retina began fully zoomed in to the center of the canvas.

Run	True Pos	False Pos	True Neg	False Neg	Overall Match Rate
1	458	42	70	930	92.53%
2	488	12	4	996	98.93%
3	417	83	40	960	91.80%

Table 5. Evaluations of champions from runs with zoom disabled, fully zoomed in.

The behavior of the evolved systems with the ability to zoom (tables 2 & 3) closely resembled the behavior of the systems evolved in [2]. Successful individuals varied in the precise strategy used to discriminate between shapes, but there were many similarities in their behaviors.

Individuals with the ability to zoom always zoomed in to an a specific edge or corner of the target shape. Most often, the retina would disengage from the shape and drift toward a particular corner of the canvas to signal a negative affinity output. For positive identifications, the retina would focus on a particular corner and continue to scan that corner while outputting a positive affinity response.

Some networks always began a discrimination sequence by outputting a positive affinity, then switching to a negative output once the shape was scanned and an identification was made. Most output a negative affinity to begin with, before switching to positive after encountering a square. Some networks focused on a given upper corner, while others either focused on lower corners or lower edges. There was a variation in the region scanned, but the aspects of scanning a particular corner or edge, disengaging non-matches and remaining engaged in matches, were the predominant commonalities.

Networks with the ability to rotate often began by rotating slightly (e.g., 10 degrees) in one direction before spending the rest of the evaluation rotating in the other direction, up to 90 degrees. These networks did not rotate to a particular orientation when zoomed in on a particular feature such as a corner or edge. The rotation speed and direction was about the same for all images, and did not appear to contribute much to shape recognition.

4. DISCUSSION

We have presented an evolved neural network controlled active vision system that elaborates on previous models by introducing rotation into the range of navigational features, and have tested the usefulness of that feature in discriminating samples of shapes randomized with respect to size, location, and rotation.

The most surprising result was that in ablation runs, individuals without the ability to rotate were able to evolve to perform better than those with all features intact. Rotational variance does not affect the appearance of circles, but it does affect the appearance of 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.

The behavior and performance of the rotation-enabled networks suggests not only that the ability to rotate was not exploited as a helpful navigational feature, but that the added complexity was detrimental to the search.

Before performing the experiments, our assumption was that networks given the ability to rotate would outperform those that could not when attempting to discriminate between shapes that had been randomly rotated. This turned out not to be the case. It remains to be seen if this generally true of active vision systems, or if the ability to rotate becomes useful, or even necessary, for more complex image recognition tasks.

The ablation tests with regard to zoom indicate that the ability to resolve the image at higher resolutions and focus on particular features is important in making correct identifications. Those individuals that began fully zoomed in were able to evolve behavior that allowed for more accurate discrimination than those that began fully zoomed out but could not zoom in, suggesting that the ability to resolve details of local features is in general more useful than course-grained global input.

The results also suggest that, as with certain types of biological organisms, the discrimination strategy does not involve template matching (i.e., memorizing and storing a template of the image and comparing that stored image with the presented image from a particular viewpoint), but rather what some researchers refer to as parameter extraction.

Campan et al [7] studied the ability of two species of bee, *Apis mellifera* and *Megachile rotundata*, to discriminate between black convex shapes on white backgrounds. The target shapes were mounted on tubes, only one of which led to the hive. Bees had to learn correctly to identify a target shape in order to return to their hive. The researchers demonstrated by using both patterned shapes and patterned backgrounds in further tests, that the bees were not using stored templates for comparison, but rather were identifying features on the perimeter of the shapes, such as angles and edges.

Their conclusions were drawn not only from the patterned tests, but from the flight patterns and scanning strategies of the bees. In the case of *A. mellifera*, the authors describe the bees as scanning the regions of the shapes that tended to differ. So that in the case of a diamond and a down-pointing triangle, the bees tended to spend time scanning along the upper part of both images, where they differed. This behavior sounds remarkably similar to the scanning strategies used by the evolved neural networks.

Moller [8] reached the same conclusion with regard to desert ants. They use parameters extracted from images rather than photographic, retinotopical templates. There is still ongoing debate about which approach is used in both invertebrates and vertebrates. Some studies demonstrate apparent template matching in vertebrates such as fish [9] and chickens [10].

It would seem that in all successfully evolved individuals, a strategy much more akin to parameter extraction arose, and that like biological systems, it is robust with regard to variations in scale, location, and rotational orientation.

5. CONCLUSION

The experiments in this paper have demonstrated the efficacy of an active vision system controlled via a recurrent neural network in performing basic shape discrimination tasks with a high degree of reliability. Our model builds upon previous approaches by adding in the ability for the system to rotate, and in the test cases explored in these experiments, that navigational ability actually hampers the evolving system's ability to learn the particular discrimination.

Future work involves expanding the model further, perhaps with higher resolution or multi-resolution retinas, and applying them to more difficult classification tasks, such as automated fingerprint classification (i.e., right loop, left loop, arch, or whorl).

Figure 6. Four different classes of fingerprints (From left to right: Right Loop, Left Loop, Arch, and Whorl). Source: FVC Database.

6. REFERENCES

[1] Wells, W. Statistical approaches to feature-based object recognition. International Journal of Computer Vision, (1997) 21(1/2):63--98.

[2] Floreano, D., Kato, T., Marocco, D. and Sauser, E. Coevolution of active vision and feature selection. Biological Cybernetics, (2004) 90(3), 218-228.

[3] Stanley, K. 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.

[4] Terzopoulos, D., Rabie, T., and Grzeszczuk, R. Perception and learning in artificial animals. Artificial Life V: Proc. Fifth International Conference on the Synthesis and Simulation of Living Systems, Nara, Japan, (May 1996), 313-320.

[5] Stanley, K. and Miikkulainen, R. Evolving Neural Networks Through Augementing Topologies. Evolutionary Computation (2002) 10(2):99-127.

[6] Stanley, K. and Miikkulainen, R. Competitive coevolution through evolutionary complexification. Journal of Artificial Intelligence Research (2004) 21: 63-100.

[7] Campan, R. and Lehrer, M. Discrimination of closed shapes by two species of bee, *Apis mellifera* and *Megachile rotundata*. Journal of Experimental Biology (2002) 205(4):559-572.

[8] Moller, R. Do Insects Use Templates or Parameters for Landmark Navigation? Journal of Theoretical Biology, Vol. 210, No. 1 (May 2001) 33-45.

[9] Schuster, S. and Amtsfeld, S. Template-matching describes visual pattern-recognition tasks in the weakly electric fish *Gnathonemus petersii*. The Journal of Experimental Biology (2002) 205, 549–557.

[10] Dawkins, M. and Woodington, A. Pattern recognition and active vision in chickens. Nature (February 2000) 403:10, 652-654.