Evolutionary Active Vision Toward Three Dimensional Landmark-Navigation

Mototaka Suzuki and Dario Floreano

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Laboratory of Intelligent Systems, CH-1015 Lausanne, Switzerland
Mototaka.Suzuki@epfl.ch, Dario.Floreano@epfl.ch
http://lis.epfl.ch/activevision

Abstract. Active vision may be useful to perform landmark-based navigation where landmark relationship requires active scanning of the environment. In this article we explore this hypothesis by evolving the neural system controlling vision and behavior of a mobile robot equipped with a pan/tilt camera so that it can discriminate visual patterns and arrive at the goal zone. The experimental setup employed in this article requires the robot to actively move its gaze direction and integrate information over time in order to accomplish the task. We show that the evolved robot can detect separate features in a sequential manner and discriminate the spatial relationships. An intriguing hypothesis on landmark-based navigation in insects derives from the present results.

1 Introduction

Active vision emphasizes the role of vision as a sense for robots and other real-time perception-action systems [1,2,3,4]. It picks out the properties of images which are necessary to perform its assigned tasks, and ignores the rest. In this context, there is no clear need for the sort of detailed reconstructions of the visible world that have been an accepted, traditional goal of machine vision [5].

Active vision may be useful to perform landmark-based navigation where landmark relationship requires active scanning of the environment. In this article we explore this hypothesis by evolving the neural system controlling vision and behavior of a mobile robot equipped with a pan/tilt camera so that it can discriminate visual patterns and arrive at the goal zone.

The experimental setup employed in this article has a notable characteristic: the visual landmarks are identical if the elevation of a robot's camera is fixed with the body. In that case, the robot could be unable to discriminate one from the other[1]. It needs to actively move its gaze direction and integrate information over time in order to differentiate these patterns. The sequential detection of spatially separate visual landmarks has been largely neglected in the literature. Instead most machine vision systems process an entire image of their large visual field every time step.

[1] The use of a panoramic camera which provides larger field of view is discussed in section 4.

S. Nolfi et al. (Eds.): SAB 2006, LNAI 4095, pp. 263–273, 2006.

We show that the best evolved robots successfully perform the task by using an effective scanning strategy. The evolved active scanning trajectory covers only a small region of the entire visual field and, more importantly, consists of a sequence of feature-driven, anticipatory, and context-dependent gaze movements. We address the advantages of the present method and neural architecture in terms of algorithmic, computational and memory resources.

The rest of this paper is organized as follows: the next section details the experimental setup, i.e. the environment, the simulated robot and the task for the robot. The neural network embedded in the robot and the genetic algorithm for developing the synaptic weights in the neural network are also described. Results and the analysis of the best evolved individual are described in Section 3. Finally an intriguing hypothesis on landmark-based navigation deriving from the present results and conclusions are discussed in Section 4 and 5 respectively.

2 Methods

The neural control system of a mobile robot equipped with a pan/tilt camera is evolved by means of a genetic algorithm to perform goal-directed navigation in an enclosed space using only visual information (Fig. 1). The evolutionary algorithm evaluates each neural controller with random mutations until an evolutionary stable control strategy is found [6]. In order to collect data from several independent runs and perform rigorous statistical analysis, we used fast, physics based simulations of the robot and its environment (Fig. 1).

Fig. 1. Left: The original six-wheeled robot Koala equipped with a pan/tilt camera. Right: The robot's perspective in a simulated environment. The robot can access the world with 5 by 5 retina at the center of the image.

We simulated the robot and the environment using physics-based Vortex libraries[2]. The robot has six wheels, but only the central wheel on each side is motorized. The robot base is 30cm(W)×32cm(L)×20cm(H). The pan and tilt angles of the camera are controlled by two separate and independent motors.

[2] `http://www.cm-labs.com`

2.1 Experiment and Task

Figure 2 shows the experimental setup where each of two facing white walls has two squares placed at different heights. The task of the robot is to visually discriminate one wall from the other in order to arrive at the goal zone at the end of each trial. There is no other identification of the goal than the visual patterns. Importantly this experimental setup is designed such that it does not allow the visual field of the robot to cover both of the two black squares at any given moment. Therefore the robot cannot discriminate the two walls by keeping the

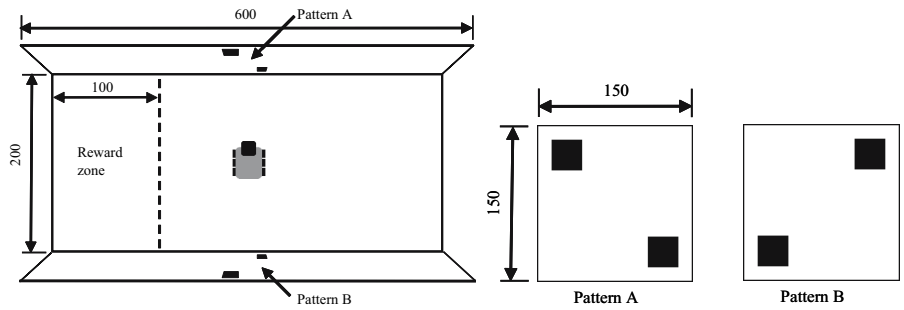

Fig. 2. The arena (200cm×600cm) and two visual patterns used in the simulation. The visual field of the robot can not cover both of the two black squares at any given moment. The difference of the two walls resides in the spatial relationship of the two squares (right). The position and direction of the robot are randomized at the beginning of each test.

vertical angle of the camera constant because both walls have an identical black square in the same height. The difference of the two walls resides in the spatial relationship of the two squares (Fig. 2, right). The robot needs to discriminate one pattern from the other by using active, sequential scanning of the two black squares of each pattern and integrating the information over time.

2.2 Neural Architecture and Genetic Algorithm

The neural network is characterized by a feedforward architecture with evolvable thresholds and discrete-time, fully recurrent connections at the associative layer (Fig. 3). A set of visual neurons, arranged on a grid, with non-overlapping receptive fields receives information about the gray level of the corresponding pixels in the image provided by the camera on the robot. The receptive field of each unit covers a square area of 48 by 48 pixels in the image. We can think of the total area spanned by all receptive fields (240 by 240 pixels) as the surface of an artificial retina. The activation of a visual neuron, scaled between 0 and 1, is given by the average gray level of all pixels spanned by its own receptive field or by the gray level of a single pixel located within the receptive field. The choice between these two activation methods, or filtering strategies, can be dynamically changed by one output neuron at each time step. An object detector unit

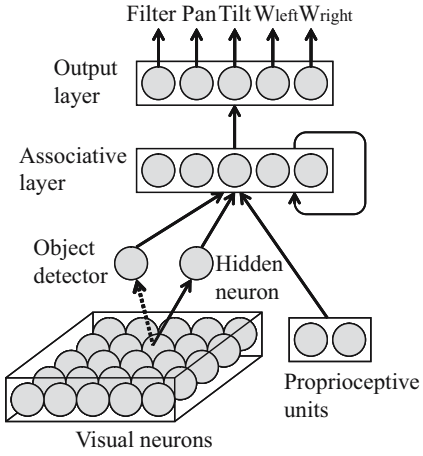

Fig. 3. The neural architecture is composed of a grid of visual neurons with non-overlapping receptive fields whose activation is given by the gray level of the corresponding pixels in the image; an object detector unit that is activated when any visual neuron is strongly activated; a hidden unit with incoming synapses from visual neurons; a set of proprioceptive neurons that provide information about the movement of the camera with respect to the chassis of the robot; a set of output neurons that determine at each sensory motor cycle the filtering used by visual neurons, the new pan and tilt speeds of the camera and the rotational speeds of the two wheels of the robot; a set of associative neurons with recurrent connections. Solid arrows between layers represent fully connected synaptic weights. Dashed arrow represents a predetermined (non-evolvable) `OR` filter (see main text for more detail).

is activated when any visual neuron is strongly activated. Therefore the synaptic weights incoming into this unit can be seen as a predetermined (non-evolvable) `OR` filter. Two proprioceptive units provide input information about the measured horizontal (pan) and vertical (tilt) angles of the camera. These values are in the interval $[-100, 100]$ and $[-25, 25]$ degrees for pan and tilt, respectively. Each value is scaled in the interval $[-1, 1]$ so that activation 0 corresponds to 0 degrees (camera pointing forward parallel to the floor). A set of memory units store the values of the associative neurons at the previous sensory motor cycle step and send them back to the associative units through a set of connections, which effectively act as recurrent connections among associative units [7]. The bias unit has a constant value of -1 and its outgoing connections represent the adaptive thresholds of associative, hidden and output neurons [8].

Associative, hidden and output neurons use the sigmoid activation function $f(x) = 1/(1 + exp(-x))$ in the range $[0, 1]$, where x is the weighted sum of all inputs. Output neurons encode the motor commands of the active vision system and of the robot for each sensory motor cycle. One neuron determines the filtering strategy used to set the activation values of visual neurons for the next sensory motor cycle. Two neurons control the movement of the camera, encoded as speeds relative to the current position. The remaining two neurons

encode the direction and rotational speeds of the left and right motored wheels of the robot. Activation values above and below 0.5 stand for forward and backward rotational speeds respectively.

The present neural architecture has been incrementally developed based on our previous investigations [9,10]. The object detector unit incorporated in the architecture is explicitly designed to simplify the biological visual system capable of monitoring for change in the visual environment[3]. The hidden neuron is incorporated to equalize the contributions of the visual neurons, the object detector unit and the proprioceptive units to the activations of the associative neurons. The roles of the hidden and object detector units are further analyzed in section 3.

The neural network has 106 evolvable connections that are individually encoded on five bits in the genetic string (total length=530 bits). A population of 100 genomes is randomly initialized by the computer. Each individual genome is then decoded into the connection weights of the neural network and tested on the robot while its fitness is computed. The best 20% of the population (those with the highest fitness values) are reproduced, while the remaining 80% are discarded. Equal number of copies of the selected individuals are made to create a new population of the same size. The new genomes are randomly paired, crossed over with probability 0.1 per pair and mutated with probability 0.01 per bit. Crossover consists in swapping genetic material between two strings around a randomly chosen point. Mutation consists in toggling the value of a bit. Finally two copies of the best genomes of the previous generation are inserted in the new population at the places of the randomly chosen genomes (elitism) in order to improve the stability of the evolutionary process.

The fitness function was designed to select robots for their ability to arrive at the goal zone at the end of each life. Each individual is tested for six trials, each trial lasting for 300 sensory motor cycles. A trial can be truncated earlier if the operating system detects an imminent collision into the walls. The fitness criterion F is composed as follows:

$$F = F_{speed}(S_{left}, S_{right}) + F_{goal} \tag{1}$$

where $F_{speed}(S_{left}, S_{right})$ is a function of the measured speeds of the left S_{left} and right S_{right} wheels and F_{goal} is a reward given if the robot reaches the goal at the end of its life[4]. More specifically $F_{speed}(S_{left}, S_{right})$ is defined as follows:

$$F_{speed}(S_{left}, S_{right}) = \frac{1}{ET} \sum_{e=0}^{E} \sum_{t=0}^{T'} f(S_{left}, S_{right}, t) \tag{2}$$

$$f(S_{left}, S_{right}, t) = (S^t_{left} + S^t_{right})(1 - \sqrt{|S^t_{left} - S^t_{right}|/2S_{max}}) \tag{3}$$

[3] In our preliminary studies it seemed difficult to develop the visual system capable of significantly responding to the black squares detected at any location of the retina.

[4] One might think that the first term $F_{speed}(S_{left}, S_{right})$ is not necessary, but in our preliminary study the fitness value remained zero without $F_{speed}(S_{left}, S_{right})$, meaning that evolution could not find the solution.

where S_{left} and S_{right} are in the range $[-8, 8]$ cm/sec and $f(S_{left}, S_{right}, t) = 0$ if S_{left} or S_{right} is smaller than 0 (backward rotation); E is the number of trials (six in these experiments), T is the maximum number of sensory motor cycles per trial (300 in these experiments), T' is the observed number of sensory motor cycles; F_{goal} is 10 if the robot reaches the goal area at the end of the test, otherwise it is 0. The reward is given only if at least one lower square and one upper square are detected before reaching the goal. This stronger constraint on F_{goal} is to prevent selecting 'blind' individuals which arrive at the goal by chance without using visual patterns.

At the beginning of each trial the position and orientation of the robot are randomized in the interval $[-50, 50]$ and $[-20, 20]$ for the longitudinal and short axes respectively.

3 Results and Analysis

We performed six replications of the evolutionary run starting with different initial populations. In all cases the fitness reached stable values in less than 30 generations (Fig. 4), and the fitness value of the best evolved individual ranged from 40 to 60.

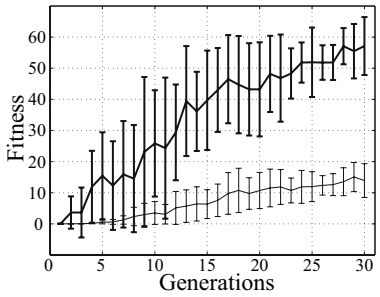

Fig. 4. Evolution of neural controllers for the simple three dimensional landmark navigation. Fitness values of the population average (thin line) and the best individual (thick line) across 30 generations. Vertical bars show the standard deviation. The results are averaged over six evolutionary runs.

We analyzed the behavior of the best evolved individual which arrived at the goal six times out of six trials. Figure 5 shows the scanning strategy, the trajectory of the robot, the camera movement with respect to the chassis of the robot and the activation of neuron 5 in the associative layer when the robot started in the face of pattern A and B. For clarity we show only the activation of neuron 5 because we found that it played the most significant role in the pattern discrimination.

The behavioral strategy of the best evolved robot can be illustrated as follows: 1. The robot searches for a lower square by moving the camera left-downward and turning its chassis counter-clockwise until it finds one; 2. Once it finds one,

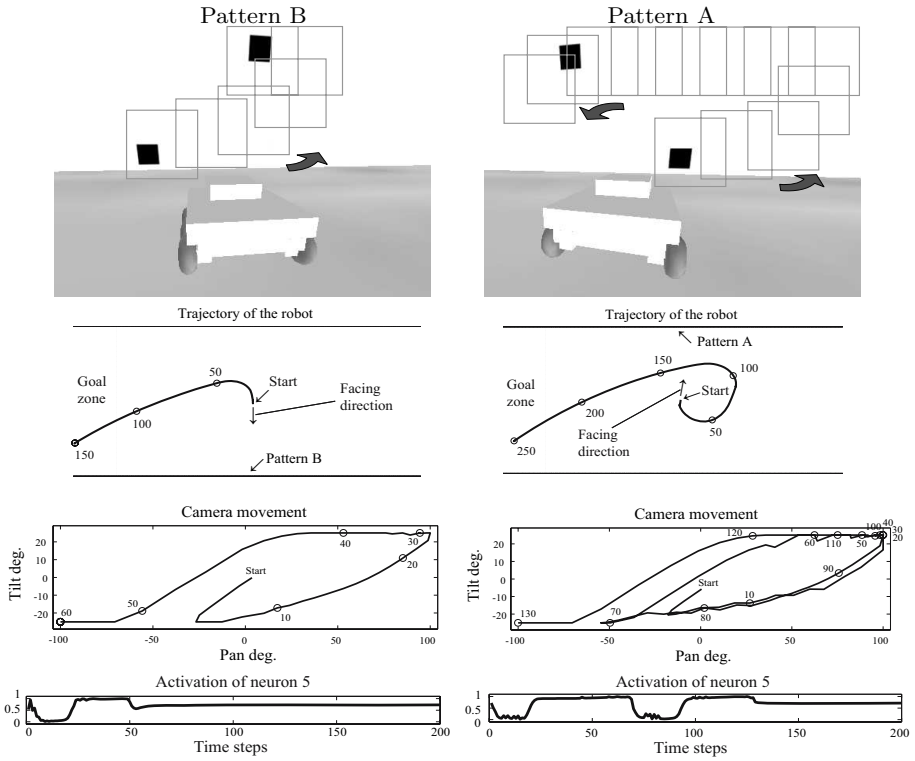

Fig. 5. From top to bottom, the best evolved robot scanning two black squares of each pattern sequentially (gray squares depicting the trajectory of the retinal perimeter), the trajectory of the robot, the camera movement with respect to the chassis of the robot and the activation of neuron 5 in the associative layer during the behavior (shown only for the first 200 sensory motor cycles) when the robot started in the face of pattern B (left column) and A (right column)

it points the camera right-upward to find an upper square; 3. If it finds an upper square after a short delay, it goes toward the goal while moving the camera left-downward. If not, it moves the camera left-downward again while turning its chassis counter-clockwise until it finds another lower square, and then goes back to step 2. Thus the robot always searches for pattern B to go toward the goal.

We studied the role of the hidden and object detector neurons by lesioning one at a time. Their operation was disrupted by clamping the activation value of the neuron to a constant value of 0.5 during behavior. Figure 6 (left) shows that both neurons significantly contribute to the successful performance. The best evolved individual while the object detector neuron was lesioned arrived at the goal zone five times out of 20 trials (10 in the face of pattern A plus 10 in the face of pattern B, at the beginning). However these successes were achieved only when the robot started facing pattern B. If the robot was facing pattern A it went in the opposite direction of the goal. In other words, the robot always goes

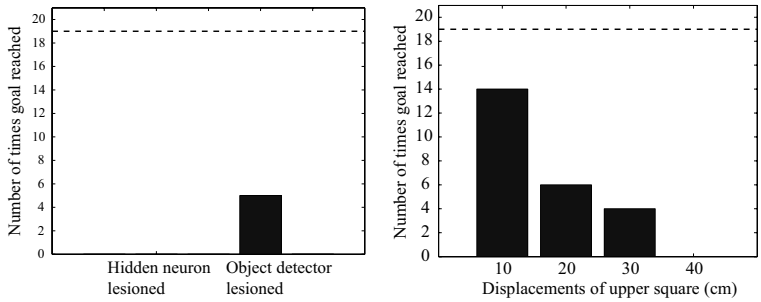

Fig. 6. The number of successful arrivals at the goal is counted in each condition out of 20 trials. Left: Lesion test of the best evolved individual. Horizontal dotted line shows the score of the intact best evolved individual. Right: Test of the best evolved individual when upper squares are displaced. Horizontal dotted line shows the score when upper squares are *not* displaced.

right in the face of both pattern A and B. This result suggests the crucial role of the object detector neuron in the behavior selection or decision making. That neuron significantly contributes to measuring the time interval between looking right-upward and subsequent detection of an upper square. Without the object detector neuron the robot can not measure the time interval and therefore can not discriminate the patterns.

While the hidden neuron was lesioned, the best evolved individual never arrived at the goal. This result suggests that the individual uses not solely the temporal information given by the object detector neuron, but also the visual information given through the hidden neuron.

One might think that the scanning strategy is reactive, i.e. the detection of a lower or an upper square always activates a particular behavior, but it is not. For example, in Fig. 5 (right) the lower square of pattern A was detected for the second time in the left side of the robot around the 130th time step, but this event did not affect the behavior of the robot going toward the goal. Therefore it seems that the behavior had been 'switched on' before the event[5]. The decision might be made when the upper square of pattern B was detected shortly after looking right-upward. If an upper square is detected late after looking right-upward, the robot does not go toward the goal, but resumes searching for a lower square.

This hypothesis was supported by another set of analyses where the upper square in each pattern was horizontally shifted toward the center (Fig. 6, right). The robot can not discriminate the two patterns any more if the upper square is shifted more than 20 cm.

The importance of the proprioceptive inputs is validated by another set of evolutionary runs without proprioceptive inputs (Fig. 7, left). Despite the shorter length of the genetic string (total length=480 bits), the best evolved individuals in all six evolutionary runs could reach the goal only three times out of six trials at max-

[5] The stable activation of the neuron 5 around 0.7 (see Fig 5, bottom) seems to reflect such a fixed behavior after the decision making.

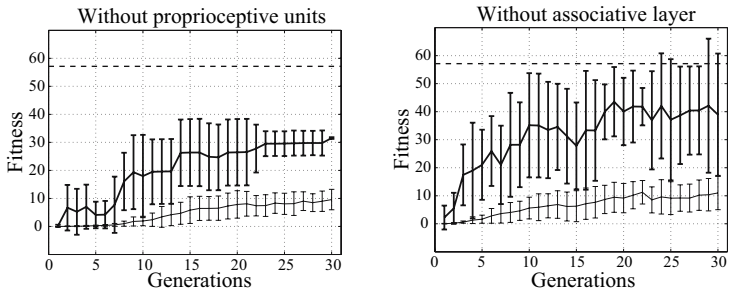

Fig. 7. Left: Evolution *without* proprioceptive inputs encoding pan and tilt movements. Right: Evolution *without* the associative layer. Fitness values of the population average (thin line) and the best individual (thick line) across 30 generations. Vertical bars show the standard deviation. Averaged over six evolutionary runs. Horizontal dotted line shows the averaged fitness value of the best evolved individual with the original neural architecture (see Fig. 4).

imum. Their behavioral analysis shows that these individuals always go left (or right depending on the evolutionary run) in the face of both pattern A and B. In other words they do not differentiate one pattern from the other.

One more set of evolutionary runs with another neural architecture which has fully recurrent connections at the output layer and does not have the associative layer shows worse fitness values than those with the original neural architecture (Fig. 7, right).

4 Discussion

We have shown that the evolved robot can detect two separate features in a sequential manner and discriminate the spatial relationships. If the system can perform active vision and sequentially store the events of visual feature detection, we do not need expensive computational power nor large memory storage capacity which would be required to resort to image memorization and matching. Although it has been shown that insects may indeed adopt such an image memorization and matching strategy [11], it is tempting to speculate that their tiny brain with restricted memory capacity may favor a more economical strategy as shown in this paper.

The evolved robot was able to effectively scan small regions of the broad visual field in an anticipatory manner in order to sequentially detect separate features. Such a characteristic of the evolved scanning strategy is in agreement with the evidence shown in [12,13] that people direct their gaze to points of the scene where information is to be extracted. Land et al. recorded human eye movements while playing cricket and table tennis. The eyes are very active and their activity takes roughly the same path as the ball. Contrary to popular belief, they do not follow the ball, but work in an anticipatory way. For example, eyes anticipate the position of the ball before it bounces, making saccades to positions where there is, as yet, no visible stimulus. In this article we have shown a computational model capable of an anticipatory "eye" movement in a freely moving behavioral system.

In order to detect the spatially separate features, the evolved robot executes a particular scanning sequence in front of the visual patterns. That is, after detecting a lower square, the robot routinely directs its gaze right-upward. Such a scanning sequence might be reminiscent of the human 'scanpath' during facial recognition [14,15]: Noton and Stark claimed that when a particular visual pattern is viewed, a particular sequence of eye movements is executed and furthermore that this sequence is important in accessing the visual memory for the pattern. The evolved scanning strategy presented in this article is similar to the 'scanpath' in that the moving sequence is crucial to identify a particular pattern. However, notice that the evolved scanning strategy is not for accessing the visual memory, but rather is tightly coordinated with the behavior of the robot.

From an engineering point of view one may argue that a panoramic camera could allow the robot to cover the entire visual field and discriminate the two patterns. However this approach would be computationally expensive if the entire image is to be uniformly processed in high resolution to extract tiny features out of a vast visual field as we have shown in this article. Active vision applied to an omnidirectional image is studied in a separate article [16].

Although the present neural architecture shown in Fig. 3 was investigated in the lesion test and additional evolutionary runs with modified neural architectures, further investigations must be done. We intend to identify the minimum components necessary for the neural controller of the robot to detect spatially separate features in the three dimensional visual environment.

5 Conclusions

In this paper we have shown that active vision may help not only to locate important features of the environment, but also to capture spatial relationships between those features that could provide behaviorally relevant information.

From these results it can be hypothesized that landmark-based navigation in insects and robots could be mediated by similar mechanisms instead of resorting to image memorization and matching [11]. We are currently exploring this hypothesis with simulated and physical robots.

Acknowledgments

Thanks to Danesh Tarapore and Claudio Mattiussi for enhancing the readability of this article. Two anonymous reviewers also provided helpful comments on the draft of this paper.

References

1. R. Bajcsy. Active Perception. *Proceedings of the IEEE*, 76:996–1005, 1988.
2. J. Aloimonos, I. Weiss, and A. Bandopadhay. Active Vision. *International Journal of Computer Vision*, 1(4):333–356, 1987.
3. J. Aloimonos. Purposive and Qualitative Active Vision. In *Proceedings of International Conference on Pattern Recognition*, volume 1, pages 346–360, 1990.

4. D. H. Ballard. Animate Vision. *Artificial Intelligence*, 48(1):57–86, 1991.
5. B. Horn. *Robot Vision*. McGraw-Hill, New York, 1986.
6. S. Nolfi and D. Floreano. *Evolutionary Robotics: Biology, Intelligence, and Technology of Self-Organizing Machines*. MIT Press, Cambridge, MA, 2000.
7. J. L. Elman. Finding Structure in Time. *Cognitive Science*, 14:179–211, 1990.
8. G. E. Hinton and T. J. Sejnowski, editors. *Unsupervised Learning: Foundations of Neural Computation*. MIT Press, Cambridge, MA, 1999.
9. D. Floreano, T. Kato, D. Marocco, and E. Sauser. Coevolution of Active Vision and Feature Selection. *Biological Cybernetics*, 90(3):218–228, 2004.
10. D. Floreano, M. Suzuki, and C. Mattiussi. Active Vision and Receptive Field Development in Evolutionary Robots. *Evolutionary Computation*, 13(4):527–544, 2005.
11. S. P. D. Judd and T. S. Collett. Multiple Stored Views and Landmark Guidance in Ants. *Nature*, 392:710–714, 1998.
12. M. F. Land and S. Furneaux. The Knowledge Base of The Oculomotor System. *Philosophical Transactions of the Royal Society of London, Series B*, 352:1231–1239, 1997.
13. M. F. Land and P. McLeod. From Eye Movements to Actions: How Batsmen Hit The Ball. *Nature Neuroscience*, 3(12):1340–1345, 2000.
14. D. Noton and L. Stark. Scanpaths in Saccadic Eye Movements while Viewing and Recognizing Patterns. *Vision Research*, 11:929–942, 1971.
15. D. Noton and L. Stark. Scanpaths in Eye Movements during Pattern Perception. *Science*, 171:308–311, 1971.
16. M. Suzuki, J. van der Blij, and D. Floreano. Omnidirectional Active Vision for Evolutionary Car Driving. In T. Arai, R. Pfeifer, T. Balch, and H. Yokoi, editors, *Proceedings of the 9th International Conference on Intelligent Autonomous Systems*, pages 153–161, March 7 - 9, 2006, Tokyo, Japan, 2006. IOS Press.