Biol. Cybern. 90, 218–228 (2004)
DOI 10.1007/s00422-004-0467-5
© Springer-Verlag 2004

Biological
Cybernetics

Coevolution of active vision and feature selection

Dario Floreano[1], **Toshifumi Kato**[2], **Davide Marocco**[3], **Eric Sauser**[1]

[1] Autonomous Systems Lab, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
[2] Hewlett-Packard Japan, Ltd., Tokyo, Japan
[3] Institute of Cognitive Science and Technology, National Research Council (CNR), Rome, Italy

Received: 6 January 2003 / Accepted: 19 January 2004 / Published online: 19 March 2004

Abstract. We show that complex visual tasks, such as position- and size-invariant shape recognition and navigation in the environment, can be tackled with simple architectures generated by a coevolutionary process of active vision and feature selection. Behavioral machines equipped with primitive vision systems and direct pathways between visual and motor neurons are evolved while they freely interact with their environments. We describe the application of this methodology in three sets of experiments, namely, shape discrimination, car driving, and robot navigation. We show that these systems develop sensitivity to a number of oriented, retinotopic, visual-feature-oriented edges, corners, height, and a behavioral repertoire to locate, bring, and keep these features in sensitive regions of the vision system, resembling strategies observed in simple insects.

1 Active vision and feature selection

In this paper we show that the computational complexity of visual processing can be greatly simplified by the codevelopment of active vision and feature selection. *Active vision* is the sequential and interactive process of selecting and analyzing parts of a visual scene (Bajcsy 1995, 1988; Ballard 1991). This process can simplify the computation involved in vision processing by reducing the information load on the system and by selecting only characteristics of the visual scene that are relevant for the task to be solved (Aloimonos 1993). Active vision is largely inspired by ways in which both mammals (Yarbus 1967) and insects (Srinivasan and Venkatesh 1997) gather information from their environments. For example, a doctor examining an X-ray plate for the presence of fractures sequentially directs its gaze to several points in the image (Krupinski and Nishikawa 1997), as shown in Fig. 1. And Drosophila flies, trained to discriminate between two shapes, move their body to bring selected parts of the image within matching receptive fields (Dill et al. 1993).

Feature selection instead consists in filtering the image to enhance features that are relevant for the task to be solved and discard all the rest. In computer vision, filtering can be accomplished by convolving images with a set of operators such as the Difference of Gaussians (Marr 1982) to detect edges and discard overall illumination. In biological vision, filtering is implemented by means of matched receptive fields and lateral connections that respond maximally only to some properties of the image. An example is the pattern of center-surround, antagonist synapses found in the early stages of both verterbrate and invertebrate vision systems. An introduction to computer and biological feature selection can be found in Mallot (2000).

However, the combination of active vision and feature selection has been investigated only to a limited extent. The dominant approach in computer vision consists in defining the set of features that an active vision system exploits to explore a visual scene. For example, Rimey and Brown (1994) make use of Bayes nets and decision theory to optimally position a vision sensor in an image, taking advantage of prior knowledge of environmental relations and geometrical structure. Using a different technique, Terzopoulos and Rabie (1997) describe pursuit behavior of an artificial fish that exploits active vision to find and track red spots in the visual scene. Interestingly, most models do not take into account that the number and type of visual features that an organism is sensitive to depend also on the sensory-motor and behavioral characteristics of the organism in its environment (Gibson 1979). The codevelopment and interaction of feature sensitivity and active vision behaviors are still largely unexplored. From a design perspective, an interesting complication is that behavior is determined by visual information and at the same time affects what type of visual information is gathered. This is probably the reason why either one aspect or the other is predefined and fixed in most engineered systems. A notable exception is the work by

Correspondence to: D. Floreano
(e-mail: Dario.Floreano@epfl.ch, Fax: +41-21-6935859)

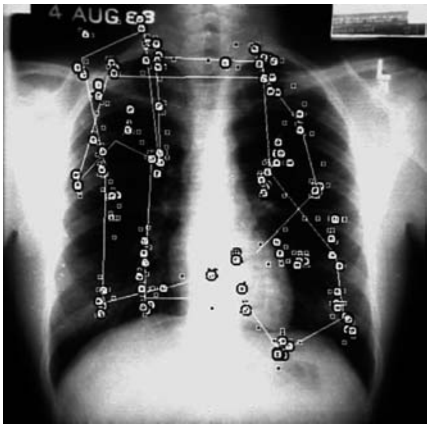

Fig. 1. Patterns of eye movements of a doctor scanning an X-ray image for the presence of bone fracture. *Dots* represent fixation points (Krupinski and Nishikawa 1997)

Franceschini et al. (1992) on biomimetic, vision-based robots that exploit sensory-motor loops to dynamically interact with and move around environments. Their approach capitalizes on a combination of insect study and specially crafted machines that reproduce biological circuitry and are free to interact with the environment.

Artificial evolution of neural architectures for autonomous robots is another methodology to address codevelopment and interaction of active vision and feature selection because it does not separate perception from behavior (Nolfi and Floreano 2000). This methodology, also known as evolutionary robotics (Cliff et al. 1993), consists in encoding the parameters of a neural system (architecture, connection weights, time constants, sensor position, etc.) of a robot into an artificial genome and evolve a population of such genomes according to a fitness function. Each genome is decoded into a correponding neural network that is interfaced to a simulated or physical robot whose fitness is measured while the robot freely interacts with its environment. Genomes with higher fitness are reproduced by making a number of copies with genetic crossover and random mutations while the remaining genomes are discarded (Holland 1975).

Within this context, Nolfi (1998) and Scheier et al. (1998) described evolved robots that exploit active perception to perform tasks that require perceptual discrimination (the robots are equipped with proximity sensors that indicate the distance to objects, not with vision). Resorting to a problem classification theory developed by Clark and Thornton (1997), these authors showed that such evolved robots turn difficult sensory classification problems into simpler ones by means of active behavior. Harvey et al. (1994) described evolution of sensory and neural morphology for a robot asked to reach for a triangular shape while avoiding a rectangular shape painted on a wall. Evolved robots solve the problem exploiting only two visual neurons whose receptive fields are aligned with a lateral edge of the triangle. The sequential activation of these neurons, caused by the sweeping of the image over the retina while the robot rotates, is sufficient to trigger the correct approaching, or avoidance, behavior. Despite the relative simplicity of such evolved systems, it has been argued that they may provide a new perspective on perceptual mechanisms with respect to conventional knowledge-based models (Cliff and Noble 1997).

In this paper, we proceed further on this line of investigation and describe a series of experiments on coevolution of active vision and feature selection for behavioral systems equipped with primitive retinal systems and deliberately simple neural architectures. In a first set of experiments, we show that sensitivity to very simple features is coevolved with, and exploited by, active vision to perform complex shape discrimination. We also show that such a discrimination problem can be very difficult for a similar vision system without active behavior. In a second set of experiments, we apply the same coevolutionary method and architecture for driving a simulated car over roads in the Swiss Alps and show that active vision is exploited to locate and fixate simple features while driving the car. In a third set of experiments, we apply once again the same coevolutionary method and architecture to an autonomous robot equipped with a pan/tilt camera that is asked to navigate in an arena located in an office environment. Evolved robots exploit active vision and simple features to direct their gaze at invariant features of the environment and perform collision-free navigation. The experiments on shape discrimination and robot navigation were described in Kato and Floreano (2001) and Marocco and Floreano (2002), respectively. In addition to presenting the new experiments on car driving, in this paper we show that evolved systems display similar strategies for active vision and feature selection across all experiments.

The next section describes the architecture of the neural system and the evolutionary method used in the three experiments. Minor modifications of the architecture due to experimental constraints are described at the beginning of each experimental section. The discussion section compares the solutions discovered across various experimental settings and provides a unified framework for understanding the principles exploited by the coevolutionary active vision and feature selection system.

2 Architecture and evolutionary method

The system consists of a feedforward neural network of artificial neurons with evolvable thresholds and discrete-time, fully recurrent connections at the output layer (Fig. 2). A set of visual neurons (a), arranged on a grid, with nonoverlapping receptive fields receive information about the gray level of the corresponding pixels in the image (b). The size of the receptive fields (zooming factor) can be dynamically changed by one output neuron at each time step. Values of the zooming factor depend on the constraints of the experiment, but the total area covered by the visual neurons is always smaller than the visual scene. We can think of the total area spanned by receptive fields as an artificial retina.

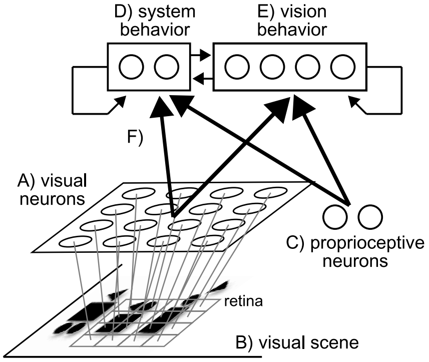

Fig. 2. The architecture is composed of **a** a grid of visual neurons with nonoverlapping receptive fields whose activation is given by **b** the gray level of the corresponding pixels in the image; **c** a set of proprioceptive neurons that provide information about the movement of the vision system; **d** a set of output neurons that determine the behavior of the system (pattern recognition, car driving, robot navigation); **e** a set of output neurons that determine the behavior of the vision system; and **f** a set of evolvable synaptic connections. The number of neurons in each subsystem can vary according to the experimental settings

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 within the receptive field (top left corner). The choice between these two activation methods can be dynamically changed by one output neuron at each time step.

A set of proprioceptive neurons (panel c) provides information about the movement of the vision system in head-centered coordinates. This information can be as simple as a binary value signaling end-of-range when the vision system cannot move any further or more detailed in the form of distance and angle from a straight, forward-looking direction. The choice of proprioceptive information can vary across experiments, depending on constraints of the physical system.

Output neurons have sigmoid activation functions $f(x) = 1/(1 + \exp(-x))$ in the range $[0, 1]$, where x is the weighted sum of the inputs minus the threshold. Thresholds are implemented as a weight from an input neuron with an activation value set to -1. Output neurons are logically organized in two blocks. One block (d) is used to determine the behavior of the system at each time step. For example, in the shape discrimination experiments, their outputs will signal the type of shape recognized by the system; in the car driving experiments, they will encode acceleration, breaking, and steering parameters of the car; in the robot navigation experiment, they will encode the speeds of the wheels of the robot. The other block (e) is used to determine the behavior of the vision system at each time step. It includes two neurons to control the movement of the camera, encoded as angle and distance relative to the current position; one neuron to define the activation method of visual neurons; and one neuron to define the zooming factor (i.e., the size of receptive fields). Small variations in the composition of this block are applied depending on the physical constraints of the experimental settings.

The system is updated at discrete time intervals. At each time interval, the following steps are performed: (i) the activation of the visual and proprioceptive neurons is computed; (ii) the activation of the output units is computed using the current weighted input values and the weighted output values computed at the previous time step; (iii) the pattern recognition system, car, or robot is updated, the vision system is shifted to its new location, and its parameters (zooming and activation method) are reset to the new values defined by the network output.

The strengths of feedforward and recurrent connections (f) are encoded in a binary string along with the threshold values of all output neurons. Connection strengths and thresholds can take values in the range $[-4.0, 4.0]$ and are each encoded on 5 bits. This binary string represents the genotype of the system and is evolved using a genetic algorithm (Holland 1975). A population of n genomes is randomly initialized by the computer. Each genome is decoded into the corresponding neural network and tested for a number of trials during which its fitness is computed. The best 20% individuals (those with highest fitness values) are reproduced, while the remaining 80% are discarded, by making an equal number of copies so as to create a new population of the same size. These new genomes are randomly paired, crossed over with probability 0.1 per pair, and mutated with probability 0.001 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, a copy of the best genome of the previous generation is inserted in the new population at the place of a randomly chosen genome (elitism).

3 Shape discrimination

In this experiment we ask the system to discriminate between triangles and squares that can appear at random locations in the visual scene (320 pixelswide × 240 pixelshigh) and can take a random size between 20 and 100 pixels in height (Fig. 3). Each image includes only one shape. Since triangles are isosceles, the base is always set twice the height so that the total area is equal to that of a square of equal height. We do this to prevent the system from recognizing a shape by its area. Shapes are black (pixel value = 0) against a white background (pixel value = 255). In addition, some noise is added to the entire image by inverting the value of each pixel (black to white or vice versa) with a probability of 0.005 per pixel.

There are nine visual neurons (arranged on a 3 × 3 grid) and one proprioceptive neuron whose activation is switched from 0 to 1 when the system attempts to move beyond a boundary of the visual scene (in that case, the system is left at its current location). Two output units of the behavior block (Fig. 2d) encode the type of shape recognized by the system (triangle and square), the most active unit being used as the network response at each time step. Four output units in the

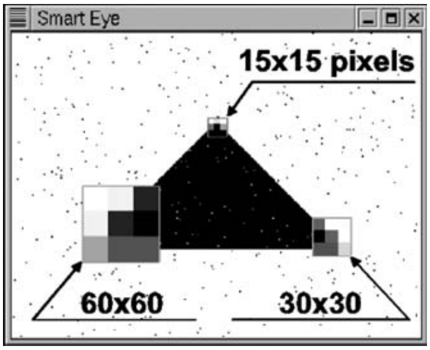

Fig. 3. A triangle with superimposed snapshots of the visual system at three different zooming factors. *Numbers* indicate the total area spanned by the receptive fields of the visual neurons

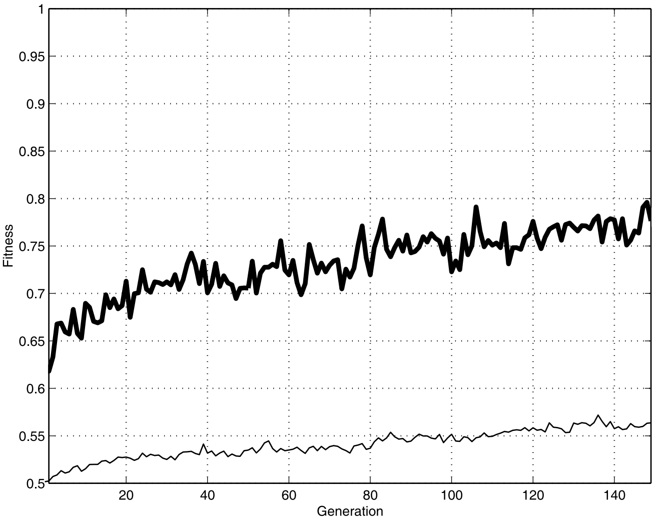

Fig. 4. Fitness data for the shape discrimination experiments. Fitness values can be read as percentages of correct response computed along the entire exploration phase. *Thick line* = best fitness of the population. *Thin line* = average population fitness. Each data point is an average of five evolutionary runs

vision block (Fig. 2e) select the activation method of visual neurons, one of three zooming factors (the side of a neuron receptive field can be 5, 10, or 20 pixels long), and define the displacement of the vision system as distance (in the range $[0, 50]$ pixels) and angle from the current location. The entire network consists of 96 connection weights and 6 threshold values that are encoded in a binary genome and evolved as explained in Sect. 2.

Each individual of the population is presented with 20 images, 10 containing a triangle and 10 containing a square. The location and size of the shapes are randomly computed anew for each image. Whenever a new image is presented, the values of the output units are reset to zero and the retina is positioned at the center of the image, after which the active vision system is free to move and change the zooming factor and sampling strategy 50 times while its response is recorded at each time step. The fitness function F of an individual is proportional to the number of correct responses recorded during the entire exploration of the visual scene for all 20 images:

$$F = \frac{1}{I * S} \sum_{i=1}^{I} \sum_{s=1}^{S} R_s^i \ , \tag{1}$$

where R_s^i is 1 if the system gives a correct response at step s for image i and 0 otherwise, S is the number of steps per image (50 in these experiments), and I is the total number of images (20 in these experiments). Notice that since the system is asked to provide a discrimination response at every time step and the probability of being presented with a triangle (or a square) is 0.5, it is easy to obtain a fitness values of 0.5 by always producing the same response, irrespective of the shape presented in the image.

A population of 100 individuals was evolved for 150 generations. Five evolutionary runs were performed, each starting with a different random initialization (Fig. 4). Notice that a fitness value of 1.0 cannot be reached because the system is asked to provide a response even before having a chance to find the shape in the image. If one counts only the final response after 50 time steps for each image, best evolved individuals are capable of correctly discriminating all shapes at any location and of any size.

Evolved strategies vary slightly across the five evolutionary runs, but all share some basic features. The vision system always starts with a fixed response (square or triangle, depending on the evolutionary run) and then moves toward the shape. Once over the shape, the retina slides back and forth along one of its vertical edges. If the edge is straight, it sets its response to square, otherwise to triangle. Figure 5 shows the trajectories of the retina in the case of two squares, and Fig. 6 shows the trajectories in the case of two triangles. A variation on this basic strategy consists in scanning the corners of the shapes instead of the edges. Once the shape has been recognized, the vision system may move away from the shape but maintains the correct response (this is made possible by recurrent connections among output units).

Most of the time (61% on average over five evolutionary runs), the activation of visual neurons is given by the value of a single pixel in the receptive field, instead of pixel averaging. Given that evolved discrimination strategies are based on the perimeter of shapes (edges and corners), this activation method provides stronger contrast between shape and background. Evolved individuals almost always use the smallest zooming factor, i.e., the largest retinal size. On the one hand, a large retinal size gives the vision system a better chance to locate the shape in the image. On the other hand, shapes are big enough (minimum height is 20 pixels, maximum is 100) to be discrimated correctly with a small zooming factor. To check the latter hypothesis, we performed five new evolutionary runs using shapes that can be smaller (height ranges from 5 to 100 pixels) and thus cannot be resolved at the smallest zooming factor. In these new conditions, best evolved individuals always change the zooming factor while they explore the scene but still display the exploration strategies described above

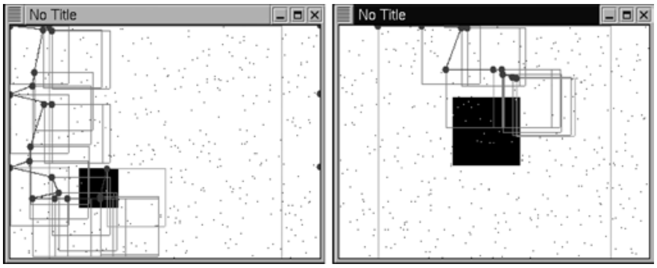

Fig. 5. Examples of trajectories of an evolved individual. The retina moves with respect to its top leftmost corner, here marked by a *dot*. The *dots* drawn after every retina movement are connected by a *line*. For graphical clarity, the values of the cells are not shown, only the retinal perimeter. *Left*: The retina starts with its initial size at the center of the image, signaling "triangle". It then shrinks to the top left corner and moves down toward the square, where it slides along its left-hand and lower edges and starts signaling "square". Finally, it ends up on the right-hand edge maintaining the correct response. *Right*: The same individual begins signaling triangle and then moves toward the square, where it moves to the right-hand edge, changing the response into "square"

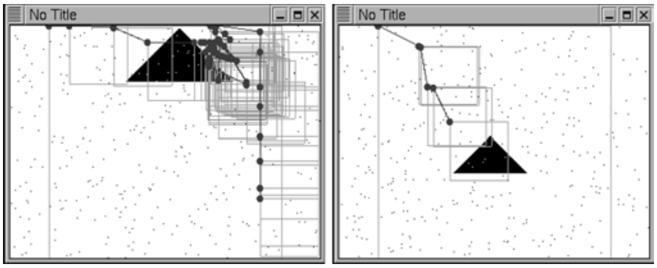

Fig. 6. Examples of trajectories of an evolved individual, as in Fig. 5 above. *Left*: Recognition of a triangle is made by exploring its right corner and then drifting away while maintaining the correct response. *Right*: Recognition is performed by looking at the left edge of the triangle

(searching for edges and corners). In these new evolutionary runs the top performance shown in Fig. 4 is reached much earlier (after about 100 generations, instead of 150). These results suggest that the ability to switch resolution more frequently helps also in the case of larger shapes. Indeed, best evolved individuals change resolution more often also when presented only with large shapes. (For details of those experiments see Kato and Floreano 2001).

3.1 Stationary discrimination

In another set of experiments, we attempted to train a stationary neural network to perform the same discrimination task by means of a supervised learning algorithm (backpropagation of error between correct response and network response; Rumelhart et al. 1986). The network has only two output units that are used for the discrimination response. Since the network cannot move across the image, it is provided with a larger number of visual neurons in order to cover the entire image.

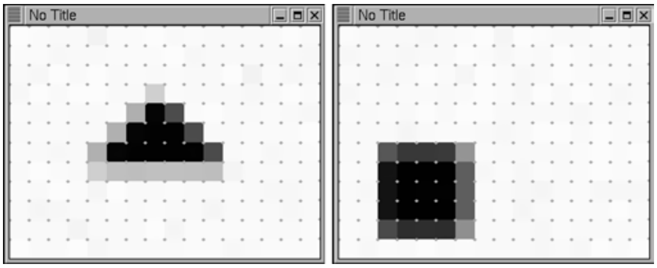

Fig. 7. Examples of image preprocessing before presentation to the stationary neural network shown in Fig. 8. The image is divided into 192 cells of 20 × 20 pixels each, and the average value of the 400 pixels in each cell is taken as the input value of the corresponding visual neuron

The image is divided up into 192 cells (receptive fields of corresponding visual neurons), each measuring 20 × 20 pixels (equal to the size of a receptive field of the active vision system at the smallest zooming factor), as shown in Fig. 7.

The average value of the 400 pixels in a cell represents the input of a corresponding visual neuron (same results were obtained using the single-pixel activation method). The neural network has 192 visual neurons and two sigmoid output units, each standing for one of the two shapes, triangle and square (Fig. 8).

The network is trained on a balanced set of images (half triangles and half squares) by randomly presenting a shape drawn at a random location with a random size (height range is between 20 and 100 pixels, as in the first set of evolutionary experiments described above). The same computer code used for generating the images in the evolutionary experiments is used here too. The connection strengths are initialized to random values in the range $\pm 1/N$, where N is the number of connections in the network (including thresholds). The error between the correct response and the network response is computed and accumulated for each presentation of 10 squares and 10 triangles and is used to update the connection strengths.

Each training session consists of 15,000 batches of 20 images (always created anew), corresponding to the number of individuals evaluated during an evolutionary run described above (150 generations with a population size of 100 individuals). We have trained networks without hidden units, with 5, 10, and 15 hidden units. Each network architecture has been trained 5 times, each time starting with new random weights. We have also tried several combinations of learning rates (0.1, 0.5, and 1.0) and momentum constants (0.1, 0.5, and 1.0).

None of these networks has ever been capable of learning to discriminate between squares and triangles, their performances always oscillating around chance level. Although we cannot exclude that different architectures and input formatting techniques can learn this discrimination task, the results indicate that a straightforward generalization of the simple vision system described in this paper cannot easily perform the task if it is deprived of active vision.

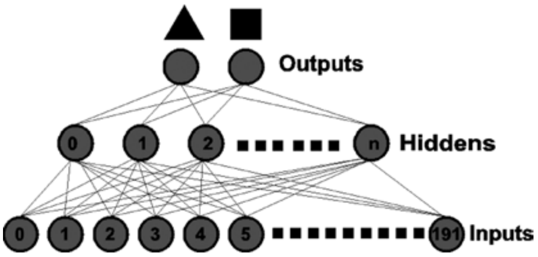

Fig. 8. Architecture of the stationary neural network trained with the backpropagation algorithm. Several architectures with a different number of hidden units (including one without) have been trained

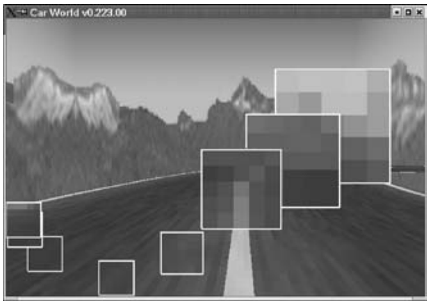

Fig. 9. Car driving. View from the windscreen with a sequence of retinal movements of an evolved individual at the beginning of a race

4 Car driving

In this experiment we ask the system to drive a simulated car over sweeping roads in the Swiss Alps. The visual scene corresponds to the view through the windscreen of the car, which spans an image of 600 × 400 pixels (Fig. 9). The open source simulator CarWorld (http://carworld.sourceforge.net), which is based on Newtonian physics, has been modified and extended to include the evolutionary active vision system, automatic circuit loading, fitness computation, and imaging tools for network analysis. The car has a mass of 1 ton, a width of 2.5 m, and soft suspensions that make it bounce easily. The road has a width of 10 m (corresponding to the diameter of car steering) and is marked by white edges and a yellow dashed line in the center. The road extends over an alpine lansdcape. Color information is mapped to grayscale levels before passing it to the visual neurons. The output units of the neural network control steering and forward/backward acceleration (backward acceleration is equivalent to braking) as well as the position of the retina on the windscreen and its zooming and activation parameters.

The neural network is composed of 25 visual neurons (arranged on a 5 × 5 grid) and two proprioceptive neurons. The side of each receptive field can vary continuously between 10 pixels (highest zooming factor) and 50 pixels (lowest zooming factor). Two proprioceptive neurons encode the vertical and horizontal position of the retina with respect to the center of the windscreen (we assume that the head of the driver does not move from this position). Two output units of the behavioral block (Fig. 2d) determine the steering direction (values above and below 0.5 correspond to right and left steering, respectively) and forward/backward acceleration (values above and below 0.5 correspond to acceleration and breaking, respectively). Two output units of the vision block (Fig. 2e) encode the speeds (max speed = 200 pixels/s) of horizontal and vertical displacements of the retina with respect to the current position. If the retina has reached a border of the windscreen, further movements in that direction have no effect. The third output unit of the vision block encodes the sampling strategy for the visual neurons, and the fourth output unit encodes the zooming factor. The neural network, driving parameters, and vision parameters are updated every 20 ms. The entire network consists of 198 connection weights and 6 threshold values that are encoded in a binary genome and evolved as explained in Sect. 2.

The system is asked to drive the car on three different circuits with variable lengths, bends, and slopes (Fig. 10). An individual is tested twice on each circuit for a maximum of 60 s per trial (real time) starting the car at a random location, orientation, and position on the road. If the car goes off the road, the current trial is terminated. The total testing time for a car that never goes off the road is 60 s × 2 trials × 3 circuits = 360 s. Initially, the retina is positioned at a random location in the windscreen at the lowest zooming factor (see top right snapshot in Fig. 9).

The fitness function F is designed to select individuals that cover the longest distance across all circuits:

$$F = \frac{1}{T * C} \sum_{t=1}^{T} \sum_{c=1}^{C} d_{t,c} \quad , \tag{2}$$

where T is the number of trials per circuit, C is the number of circuits, and $d_{t,c}$ is the normalized distance covered by the car on c circuit during t trial. Notice that if the car goes off the road, the current trial is truncated.

A population of 100 individuals was evolved for 150 generations. Three evolutionary runs were performed, each starting with a different random initialization (Fig. 11).

The performances of best evolved individuals are equal to or better than those of well-trained human drivers tested on the same circuits. Evolved drivers go as fast as possible and steer abruptly when close to the edge of the road (Fig. 12). Often the rear wheels skid sideways, and when the car reaches the bottom of a long and fast descent, it bounces up and down so markedly that the road goes temporarily out of sight. (Video clips of these conditions are available at http://asl.epfl.ch under the research section, active vision project.)

Evolved systems exploit two computationally similar strategies, depending on the evolutionary run and generation number. The first strategy consists of zooming in toward the far edge of the road and keeping it on the same retinal position (Fig. 13). Consequently, when the car approaches the edge of the road, the retina gradually shifts toward the bottom of the visual field. The resulting displacement of the retina from the horizontal position,

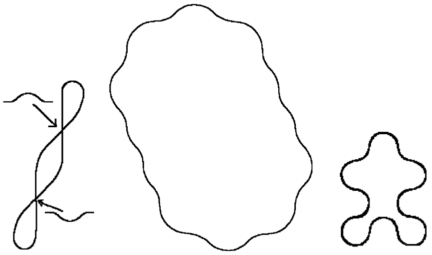

Fig. 10. Three circuits used during evolution. Circuits have different bends, slopes (*arrows*), and length. Traces are obtained by plotting the trajectory of a human driver who follows the dashed line in the middle of the road

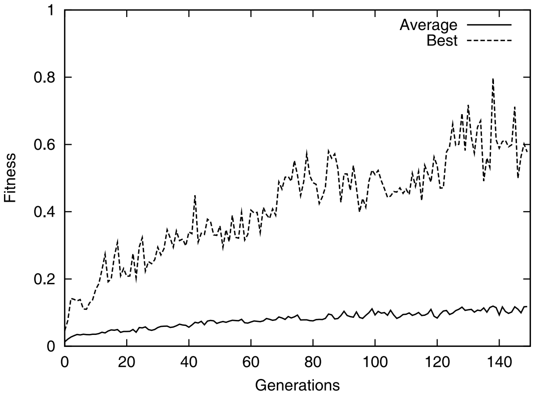

Fig. 11. Fitness data for the car driving experiment. *Continuous line* = average population fitness. *Dashed line* = best fitness of the population. Each data point is the average of three evolutionary runs

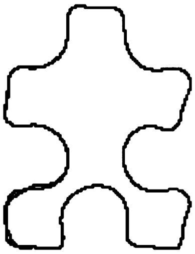

Fig. 12. Trajectory of an evolved individual tested on the rightmost circuit of Fig. 10

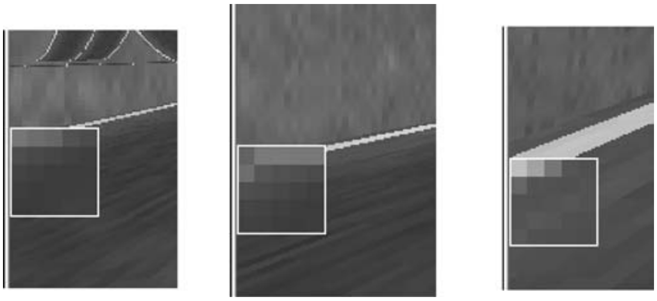

Fig. 13. The retina tracks the far edge of the road. As a consequence, when the car gets closer to the edge of the road, the retina moves toward the lower area of the visual field

which is given as proprioceptive input to the network, is used to steer the vehicle to the right. When the car turns too much to the right or the road bends to the left, the edge shrinks to the top leftmost corner of the retina, causing a left steering. The second strategy consists in zooming out toward one side of the visual field so that the edge of the road shifts over the retina as the car moves. The vertical displacement of the edge over the retina is then correlated to the steering angle and direction of the car.

5 Robot navigation

In these experiments we investigate the ability of the system to steer a real mobile robot equipped with a pan/tilt camera in an office environment. A Koala robot (Fig. 14, left) is positioned in a small square arena and asked to navigate as far as possible for 60 s while its fitness is computed as a function of forward motion over time. Individuals that hit a wall are killed and thus their total fitness is smaller than those that can move around longer. Notice that the pan/tilt camera allows the robot to watch any area of the office (almost up to the ceiling) and visitors are free to come to the office during evolution. The square arena measures 200 cm on each side and is surrounded by white walls 30 cm high (Fig. 14, right).

The robot has six soft rubber wheels but is driven by only two motors, one on each side. The video camera is equipped with two motors that allow both horizontal movement (*pan*) in the range $[-100°, 100°]$ and vertical movement (*tilt*) in the range $[-25°, 25°]$. The zooming option is not used in these experiments to reduce the number of evolvable parameters and thus shorten evolutionary time on the physical robot. The camera returns rectangular video frames to the onboard computer, where they are cropped to a square matrix of 240 × 240 pixels and RGB values are converted to grayscale levels. The onboard computer performs image preprocessing, activation of the neural network, control of the motors

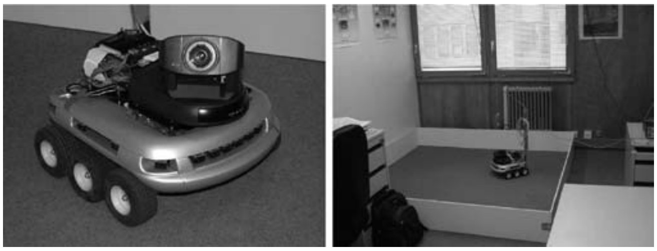

Fig. 14. *Left*: The Koala robot (produced by K-Team S.A.) equipped with a Sony EVI-D31 mobile camera and onboard PC-104 processor. The robot base is 30 cm w, 32 cm l, and 20 cm h; its total weight is 6 kg. *Right*: Evolutionary environment. The robot has visual access to the whole environment, but it can move only within the square arena. Lights were on day and night, and researchers and visitors were free to come to the office during the evolutionary process. The pole on the back of the robot prevents the aerial power supply cable from being trapped in the mobile camera. The other cable visible in this picture is used only after the evolutionary process to download data from the onboard computer to a desktop computer for analysis

of the robot and of the camera, and the evolutionary algorithm, as well as fitness computation and data storage for offline analysis. The robot was connected to a power supply through an aerial serial cable attached to the rear side of the robot and rotating contacts that allow free movement of the robot in the arena.

The system is composed of 25 visual neurons (arranged on a 5 × 5 grid) and two proprioceptive neurons. The size of each receptive field is 48 × 48. Two proprioceptive neurons encode horizontal (pan) and vertical (tilt) angles of the camera. Each value is scaled in the interval $[0, 1]$ so that an activation of 0.5 corresponds to 0° (camera pointing forward parallel to the floor). Two output units of the behavioral block (Fig. 2d) determine the speeds of the two motors of the robot in the range $[-8, 8]$ cm/s. Activation values above 0.5 stand for forward rotational speed, whereas activation values below 0.5 stand for backward rotational speed. Two output units of the vision block (Fig. 2e) encode the motor speeds of the camera on the horizontal (pan) and vertical (tilt) planes. In this case, the maximum speed in the horizontal plane is 80°/s and in the vertical plane 50°/s. If the camera has reached a maximum allowed position ($-100, 100$ and $-25, 25°$ for pan and tilt, respectively), output speeds in the same direction have no effect. The third output unit of the vision block encodes the activation method of visual neurons. There is no zooming output unit in this experiment. The entire network consists of 160 connection weights and five threshold values that are encoded in a binary genome and evolved as explained in Sect. 2.

The individuals of the population are tested, one at a time, on the same robot for two trials of at maximum 60 s each (200 sensory-motor cycles). A sensory-motor cycle lasts 300 ms (during which the wheels move at constant speed). A trial is truncated if the robot hits a wall. This condition is detected by means of infrared distance sensors located around the body of the robot, but this information is not given to the neural network. At the beginning of a trial, the robot is relocated in the environment at a random position and orientation by means of a motor procedure during which the robot moves forward and turns in a random direction for 20 s. Pan and tilt angles of the camera are set to 0°.

The fitness function is conceived to select individuals capable of moving forward as fast as possible during the time allocated for each trial. It is computed and accumulated after every sensory-motor cycle (300 ms), so that robots whose trials are truncated earlier obtain lower fitness values (Fig. 15).

The fitness $\mathcal{F}(S_{\text{right}}, S_{\text{left}}, t)$ is a function of the measured speeds of the right wheel S_{right} and left wheel S_{left}, and of time t:

$$\mathcal{F}(S_{\text{right}}, S_{\text{left}}, t) = \frac{1}{E * T}\sum_{e=1}^{E}\sum_{t=1}^{T'}((S^t_{\text{right}} + S^t_{\text{left}}) - |S^t_{\text{right}} - S^t_{\text{left}}|) \ , \qquad (3)$$

where $S_{\text{right}}, S_{\text{left}}$ are in the range $[-8, 8]$ cm/s and $\mathcal{F}(S_{\text{right}}, S_{\text{left}}, t) = 0$ if S^t_{right}, or S^t_{left} is less than 0 (backward motion), E is the number of trials (two in these experiments), T is the maximum number of sensory-motor cycles per trial (200 in these experiments), and T' is the number of sensory-motor cycles that the robot can achieve (for example $T' = 34$ for a robot whose trial is truncated after 34 sensory-motor cycles).

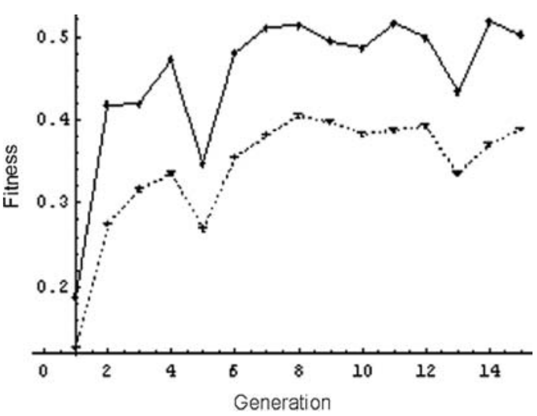

Fig. 15. Fitness data for the robot navigation experiment. *Continuous line* = best fitness of the population. *Dashed line* = average population fitness

A population of 40 individuals was evolved for 15 generations on the real robot, each generation taking approximately 1.5 h. Notice that a fitness value of 1.0 cannot be attained in this environment because the robot must turn to avoid walls, thus increasing the penalizing effect of the component $|S^t_{\text{right}} - S^t_{\text{left}}|$.

After eight generations, we noticed an alternation of two behavioral strategies across generations, one where the camera is continuously moving during navigation and one where the camera is situated and maintained at a stationary position with respect to the body of the robot. The former behavioral strategy disappears after 12 generations, although its fitness performance is equal to the latter behavioral strategy. We will start describing the strategy that exploits continuous movement of the camera because the other strategy represents a particular subset.

Figure 16 shows the trajectory, camera displacement, and visual input of the best individual at generation 12. The robot starts in the position marked by the star. The *horizontal direction* (pan) of the camera is shown by long arrows plotted at each sensory-motor cycle. The gray-scale matrices show the activations of the visual neurons and, for the sake of clarity, are plotted only before, during, and after avoidance of a wall and/or rotation of the camera in the opposite direction. The strategy consists in pointing and maintaining the camera downwards so that the visual system can detect the edge between the dark floor and the white walls of the arena. The activation strategy of visual neurons always uses the value of a single pixel in the receptive field (instead of pixel averaging) so as to enhance the edge between the floor and the wall. This edge is clearly visible in the plots of visual activation shown in Fig. 16. The robot always follows a clockwise trajectory. The camera is moved to the left when the robot is approaching a wall on its left

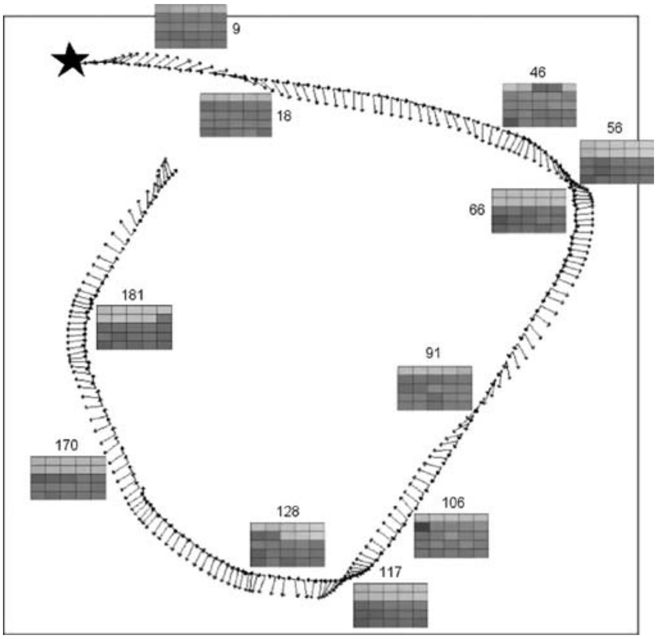

Fig. 16. Robot trajectory (*short arrows*) and horizontal camera displacement (*long arrows*) of the best individual of generation 12 tested for one trial. Grayscale matrices represent the activations of the visual neurons (*black* = 0, *white* = 1) plotted before, during, and after wall avoidance and camera movement. The *numbers* indicate the corresponding sensory-motor cycle

and then moved to the right once it has avoided it. While the camera is pointing to the right, it slowly scans back and forth, and if a wall is detected at a certain distance, it moves back to the left while the avoidance behavior is started.

The values of output units encoding pan motor and right wheel motor are strongly correlated and are determined by the amount of white on the top half of the visual field. The closer the robot gets to a wall, the larger the white area becomes in the visual field. This information is used to steer the camera toward the wall and slow down the rotation of the right wheel while the speed of the left wheel is maintained constant (which corresponds to a right turn). These data are presented in detail in Marocco and Floreano (2002).

The alternative behavioral strategy (which becomes dominant after generation 12) is similar to that described above, but the camera is not actively used throughout the whole trial. At the beginning of the trial the robot points the camera downwards and to its left, and it keeps it there for the duration of the whole trial. The movement of the body is then sufficient to maintain the edge between the floor and the walls in sight and slow down the right wheel when it gets closer to a wall (signaled by the visual expansion of the white area on the top portion of the retinal image).

6 Discussion

In the three experiments described above, the active vision system develops a sensitivity to position-specific, oriented edges. In the shape discrimination experiment, evolved individuals are sensitive to a vertical edge on the left (or right, depending on the evolutionary run) area of the retinal image while other individuals are sensitive to an inclined (approximately 60°) edge on the right area of the retinal image. In the car driving experiment, evolved individuals are sensitive to an inclined (approximately 60°) edge on the right, or to an inclined (approximately 60°) edge on the left, area of the retinal image. In the robot navigation experiment, evolved individuals are sensitive to a horizontal edge in the middle of the retinal image. All these features are linearly separable categories of input vectors, which is hardly surprising because visual neurons project directly to output neurons. Therefore, the visual system cannot rely on more complex transformations of the image that could be provided by intermediate layers of neurons. The features of these evolved systems resemble features detected by orientation-sensitive receptive fields of neurons in the mammalian visual cortex that receive direct projection from the lateral geniculate nucleus (Hubel and Wiesel 1968, 1977).

It is worth noting that sensitivity to inclined edges has been discovered also in evolutionary experiments on robot discrimination of triangles and squares by Harvey et al. (1994), whose results are summarized in the introduction of this article. Harvey's experiments differ from our experiments on shape discrimination in a number of aspects. For example, in his settings, the relative positions of the square and of the triangle are always the same and at the same height with respect to the robot surface of motion. Also, in those experiments the authors coevolve the number and shape of receptive fields of the visual neurons as well as the architecture and parameters of larger neural networks with time-dependent activation functions. In addition, considering the size of the arena where the robot was evolved, the variation in size of the retinal projections of the images was much smaller than that used in our experiments. Despite these differences, evolved neural controllers discriminated the two shapes by gauging the activation level of a visual neuron sensitive to an inclined edge during a rotational movement of the robot. The rotational movement caused the sequential sweeping of the two images across the area of that receptive field, resulting in higher activation for the inclined lateral edge of the triangle.

The sensitivity to corners found in some of the individuals evolved in our experiments for shape discrimination is a linear combination of two inclined edges. This feature is found in only half of the best evolved individuals, meaning that this strategy is equivalent to one that checks for a single inclined edge. In another set of experiments where the system was evolved to discriminate between convex and concave rectangular shapes, all best evolved individuals developed receptive fields tuned to corners. (Video clips of these experiments are available at http://asl.epfl.ch.)

The behavior of evolved active vision subsystems (Fig. 2e) provides a similar functionality across the three experiments. It moves the retina across the image to locate, bring, and maintain selected features over the receptive fields of matching visual neurons. The combination of feature selection and active vision allows evolved individuals to solve position- and size-invariant tasks using position- and size-variant mechanisms.

Once the relevant features have been brought over the receptive fields of matching neurons, the response of the system behavior neurons (Fig. 2d) varies across the three experiments. In the shape discrimination experiment, the response is a function of the type of feature detected. In the car driving experiment, steering and acceleration are a function of the relative position of the retina or of the position of the inclined edge in the retinal image, depending on the species of evolved individuals. Similarly, in the robot navigation experiment, steering is a function of camera position with respect to forward-looking position or of the amount of white wall in the upper area of the retinal image, depending on the species of evolved individuals.

In all experiments, evolved individuals display a preference for single pixel intensity vs. average pixel intensity for the activation of visual neurons. This choice provides higher contrast for perceived edges and therefore higher dynamic range for output neurons that control the behavior of the system and of the retina. An alternative solution could be to develop synaptic weights from visual neurons with more marked differential values. However, this solution is harder to find in our simple model because it requires precise settings of a higher number of parameters and because the weights are constrained to values in the range $[-4, 4]$ (corresponding to the asymptotes of the sigmoid activation function used for output neurons).

The preferred choice of zooming factor, used only in the first two experiments, corresponds to the lowest resolution sufficient for the visual neurons to detect relevant features. In most cases, this factor is the lowest available resolution (smallest zooming factor). When the size of the smallest shapes in the pattern discrimination experiment is reduced, evolved individuals dynamically change the zooming factor to allow resolution of the edges by adjacent visual neurons. In the car driving experiments, two strategies are possible. One strategy consists in locating one edge of the road and zooming in to precisely maintain it at the same location on the retina by means of active vision. In this case, the relative vertical position of the retina with respect to the straight-ahead direction is the only information used for steering and acceleration. The other strategy consists in positioning the retina on one corner of the visual scene and using the lowest available resolution to detect vertical displacement of the road edge on the retina, which is then used for steering and acceleration.

The system investigated in these experiments has been kept deliberately simple to investigate selection and exploitation of simple features of the visual scene by coevolving active vision behaviors. Therefore, we have not included lateral connections and time-dependent dynamics at the level of visual neurons. The recurrent connections at the output level can provide time-dependent dynamics and lateral interactions only at the behavioral level. Also, we have not included intermediate layers of neurons between visual and output neurons to exclude the possibility that the system may develop sensitivity to more complex features, which may reduce the role of active vision behaviors. Consequently, the feature sensitivity and behavioral strategies coevolved in these experiments are very simple, but the interesting point is that they have been selected out of a huge range of potential visual cues in accordance with the limited computational and architectural abilities of the evolutionary system.

Although this system is not modeled on biological neuronal architectures, the results described above may help to understand the relatively complex visual performance of insects equipped with simple nervous systems. For example, it has been experimentally shown that some insects rely on simple receptive fields tuned to oriented edges in order to discriminate between oriented textures during goal-oriented flight (Srinivasan et al. 1994). There is also experimental evidence that free-flying Drosophila insects can discriminate relatively complex shapes (triangles and T-shapes) by moving in such a way as to bring the shape over the receptive fields of neurons sensitive to the retinotopic height of a horizontal edge (Dill et al. 1993).

One should notice that active alignment of parts of the shape with appropriate receptive fields does not necessarily require movement of the animal. On a more speculative note, this matching mechanism may also take place within more complex brain by means of attentional mechanisms that sweep across the retinal projection. For example, Crick (1984) suggested that such a mechanism could be implemented by the interplay of thalamic-cortical forward and feedback connection coupled with local lateral inhibition within the reticular structure of the thalamus. This mechanism could be used by mammalian visual systems under some circumstances of retinotopic perceptual learning (Karni and Sagi 1991).

7 Conclusion

The experiments described in this paper indicate that coevolution of visual features and of behavior can address a variety of visual tasks that range from complex shape discrimination to navigation in complex environments by means of very simple architectures and computational abilities. Evolved individuals can solve position- and size-invariant tasks exploiting position- and size-variant receptive fields by actively searching and maintaining simple features of the visual scene over sensitive areas of the retina.

Active behavior affects, interacts with, and supports vision processing by selecting sensory experiences that can be dealt with by the system in a coherent manner. Gibson suggested that what we perceive is what the environment affords us to do (Gibson 1979). These experiments indicate that behavior is not only a variable to be considered in an ecological study of visual perception but is also intimately related to the way in which vision mechanisms develop and are exploited by the system.

Although in the experiments described in this paper we have used a genetic algorithm to shape the synaptic connections of the system, we do not intend to stress the parallelism with evolution of vision in nature. We used an evolutionary algorithm only because it allows individuals to autonomously interact with the environment instead of being guided by principles imposed by a human designer. Any other adaptive algorithm that satisfies that criterion would be suitable for the purpose of these experiments. What really matters is the relationship between developmental time scales of mechanisms responsible for action and of mechanisms responsible for visual processing. In the experiments reported here, both systems develop on the same time scale, but it would be interesting to investigate how differential time scales affect the strategies exploited by the system.

Acknowledgements. The authors thank Claudio Mattiussi, Jesper Blynel, and Shloke Hajela for help with technical preparation of the experiments and discussions. This work was carried out while the authors were at the Swiss Federal Institute of Technology Lausanne and was partially supported by the Swiss National Science Foundation, grant no. 620-58049. Video clips of the experiments and supplementary material can be found on the lab research pages (http://asl.epfl.ch).

References

Aloimonos Y (ed) (1993) Active perception. Erlbaum, Hillsdale, NJ

Bajcsy R (1985) Active perception versus passive perception. In: Proceedings of the 3rd IEEE workshop on computer vision, Bellaire, MI, April 1985. IEEE Press, Los Alamitos, CA

Bajcsy R (1988) Active perception. Proc IEEE 76:996–1005

Ballard DH (1991) Animate vision. Artif Intell 48:57–86

Clark A, Thornton C (1997) Trading spaces: computation, representation, and the limits of uniformed learning. Behav Brain Sci 20:57–90

Cliff D, Harvey I, Husbands P (1993) Explorations in evolutionary robotics. Adaptive Behav 2:73–110

Cliff DT, Noble J (1997) Knowledge-based vision and simple vision machines. Philos Trans R Soc Lond B 352:1165–1175

Crick F (1984) Function of the thalamic reticular complex: the searchlight hypothesis. Proc Natl Acad Sci 81:4586–4590

Dill M, Wolf R, Heisenberg M (1993) Visual pattern recognition in drosophila involves retinotopic matching. Nature 355:751–753

Franceschini N, Pichon J-M, Blanes C (1992) From insect vision to robot vision. Philos Trans R Soc Lond B 337:283–294

Gibson JJ (1979) The ecological approach to visual perception. Houghton Mifflin, Boston

Harvey I, Husbands P, Cliff D (1994) Seeing the light: artificial evolution, real vision. In: Cliff D, Husbands P, Meyer J, Wilson SW (eds) From animals to animats III: proceedings of the third international conference on simulation of adaptive behavior, Brighton, UK, September 1994. MIT Press-Bradford Books, Cambridge, MA, pp 392–401

Holland JH (1975) Adaptation in natural and artificial systems. The University of Michigan Press, Ann Arbor

Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. J Physiol 195:215–243

Hubel DH, Wiesel TN (1977) Functional architecture of macaque visual cortex. Proc R Soc Lond B 198:1–59

Karni A, Sagi D (1991) Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity. Proc Natl Acad Sci USA 88:4966–4970

Kato T, Floreano D (2001) An evolutionary active-vision system. In: Proceedings of the congress on evolutionary computation (CEC01), Seoul, Korea, October 2001. IEEE Press, Piscataway, NJ

Krupinski EA, Nishikawa RM (1997) Comparison of eye position versus computer identified microcalcification clusters on mammograms. Med Phys 24:17–23

Mallot HA (2000) Computational vision. MIT Press, Cambridge, MA

Marocco D, Floreano D (2002) Active vision and feature selection in evolutionary behavioral systems. In: Hallam J, Floreano D, Hayes G, Meyer J (eds) From animals to animats 7: proceedings of the seventh international conference on simulation of adaptive behavior. MIT Press-Bradford Books, Cambridge, MA, pp 247–255

Marr D (1982) Vision. Freeman, New York

Nolfi S (1998) Evolutionary robotics: exploiting the full power of self-organization. Connect Sci 10:167–183

Nolfi S, Floreano D (2000) Evolutionary robotics: biology, intelligence, and technology of self-organizing machines. MIT Press, Cambridge, MA

Rimey RD, Brown CM (1994) Control of selective perception using bayes nets and decision theory. Int J Comput Vis 12(2/3):173–207

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagation of errors. Nature 323:533–536

Scheier C, Pfeifer R, Kunyoshi Y (1998) Embedded neural networks: exploiting constraints. Neural Netw 11:1551–1596

Srinivasan MV, Venkatesh S (eds) (1997) From living eyes to seeing machines. Oxford University Press, Oxford

Srinivasan MV, Zhang SW, Witney K (1994) Visual discrimination of pattern orientation by honeybees: performance and implications for "cortical" processing. Philos Trans R Soc Lond B 343:199–210

Terzopoulos D, Rabie TF (1997) Animat vision: active vision in artificial animals. Videre J Comput Vis Res 1:2–19

Yarbus AL (1967) Eye movements and vision. Plenum, New York