Network: Computation in Neural Systems
March 2007; 18(1): 11–34

informa
healthcare

Evolution of visually guided behavior in artificial agents

BYRON BOOTS, SURAJIT NUNDY, & DALE PURVES

Department of Neurobiology and Center for Cognitive Neuroscience, Duke University, Durham, NC 27708, USA

(Received 31 July 2006; accepted 27 September 2006)

Abstract
Recent work on brightness, color, and form has suggested that human visual percepts represent the probable sources of retinal images rather than stimulus features as such. Here we investigate the plausibility of this empirical concept of vision by allowing autonomous agents to evolve in virtual environments based solely on the relative success of their behavior. The responses of evolved agents to visual stimuli indicate that fitness improves as the neural network control systems gradually incorporate the statistical relationship between projected images and behavior appropriate to the sources of the inherently ambiguous images. These results: (1) demonstrate the merits of a wholly empirical strategy of animal vision as a means of contending with the inverse optics problem; (2) argue that the information incorporated into biological visual processing circuitry is the relationship between images and their probable sources; and (3) suggest why human percepts do not map neatly onto physical reality.

Keywords: *Vision, neural networks, autonomous agents, genetic algorithms*

Introduction

It has long been recognized that the sources of visual stimuli are not uniquely specified by the light energy that reaches photoreceptors: the same pattern of light focused on a surface can arise from different combinations of illumination, reflectance and transmittance, and from objects of different sizes, at different

Correspondence: Dale Purves, Center for Cognitive Neuroscience, Box 90999, LSRC Building, Duke University, Durham, NC 27708, USA. Tel: (919) 684-6276. Fax: (919) 681-0815. E-mail: purves@neuro.duke.edu

ISSN 0954-898X print/ISSN 1361-6536 online
DOI: 10.1080/09548980601113254

distances and in different orientations. Nevertheless, visual agents must respond to real-world events. The inevitably uncertain sources of visual stimuli thus present a quandary: although the physical properties of a stimulus cannot uniquely specify its provenance, success depends on behavioral responses that are appropriate to the stimulus source. This dilemma is referred to as the inverse optics problem.

For more than a century now, most investigators have surmised that the basis of successful biological vision in the face of the inverse optics problem is the incorporation of information about prior experience in visual processing, presumably derived from both evolution and individual development. This empirical influence on visual perception has variously been considered in terms of ''unconscious inferences'' (Helmholtz 1924), the ''organizational principles'' advocated by Gestalt psychology (Wertheimer 1923), or the framework of ''ecological optics'' (Gibson 1979). More recently, these broad interpretations have been conceptualized in formal terms that reflect the real-world sources to which an animal has always been exposed, either in Bayesian terms (Knill and Richards 1996; Rao et al. 2002), or terms of an empirical ranking of stimulus-source relationships (Howe and Purves 2005a; Howe et al. 2006). Indeed, many of the anomalous percepts that humans see in response to simple visual stimuli can be rationalized in this way (reviewed in Purves and Lotto 2003; Purves et al. 2004; Howe and Purves 2005a).

If visual percepts are indeed determined empirically, then biological visual systems must have instantiated a scheme of neural processing that links inevitably ambiguous retinal images with their behavioral significance. To examine the feasibility of vision on this basis, we have here turned to evolutionary robotics, an emerging field that uses simulated evolution to generate neural network control systems that link ''sensory'' input to motor output in both simulation and physically realized robots (Pfeifer and Scheier 1999; Nolfi and Floreano 2000). In most such work, the autonomous control systems have been evolved to perform simple tasks such as obstacle avoidance (Floreano and Mondada 1994; Salomon 1996), wall following (Dain 1998) or navigational homing (Floreano and Mondada 1996). Although relatively little research has focused on vision as such, some work has included rudimentary visual input to the relevant controllers (Cliff et al. 1997; Smith 1997; Floreano 1998; Nolfi and Marocco 2000; Nolfi and Marocco 2002; Forsyth 2003; see, however, Wyss et al. 2006; Fioreano and Mondada 1998). None of these systems have addressed the inverse optics problem, and most include infrared range finders (Floreano and Mondada 1994; Salomon 1996; Floreano and Mondada 1996; Smith 1997; Nolfi and Marocco 2000; Nolfi and Marocco 2002), thereby removing the input ambiguity that natural visual systems must deal with.

In the present work, we have asked whether agents using simple neural network control systems could evolve successful behavior in response to inherently ambiguous visual input based solely on interactions with their environment. The results show that evolved agents contend with the inverse optics problem by associating projected images with behavior appropriate to the probable underlying sources experienced in their native environment. When confronted with unlikely relationships between their sensory images and the sources of the images in either novel or distorted environments, evolved agents behave in an anomalous way that

mimics the percepts and behavior of humans presented with improbable geometries. These results support the idea that biological vision resolves the inverse problem according to the probability distributions of the possible stimulus sources, providing a new way of exploring why we see the world in the peculiar way we do, and the neural basis for these anomalies.

Methods

The environments

A series of six virtual environments were created with OpenGL. Each comprised a simple arena with a central obstacle, geometrically similar to arenas used in evolutionary robotics experiments (e.g., Floreano and Mondada 1996; Salomon 1996; Nolfi and Floreano 2000) (Figure 1). The illumination of each environment was anisotropic, much as terrestrial environments are illuminated on an overcast day. The intensity of the light represented by each of the RGB channels in OpenGL was scaled such that a surface that reflected 100% of the incident light was represented by an RGB value of 255 in the simulated environment. The walls of the arena were assigned a reflectance value of 30%, and the floor 60%; the sky was assigned an RGB value of 230. As a result of these fixed assignments, the uncertainty in images projected from the environment (i.e., the inverse optics problem in the simulated environments) was restricted to a conflation in the image plane of the size, distance and orientation of object surfaces (Figure 2A).

Based on their geometrical properties, the environments we used are described in what follows as ''standard'' or ''distorted''. The two standard environments – the square (Figure 1A) and the diamond (Figure 1B) – had a level floor and walls 4 units high. The four distorted environments – the symmetrically distorted square (Figure 1C), the symmetrically distorted diamond (Figure 1D), the large asymmetrically distorted square (Figure 1E), and the small asymmetrically distorted square (Figure 1F) – had oblique floors and varying wall heights. The latter distorted environments closely resemble the ''Ames rooms'' used in human perceptual demonstrations (reviewed in Ittleson 1952). The best known of these are oddly angled rooms that, when seen from a particular viewpoint, are nonetheless perceived as rectangular (Figure 2B and C). Agents operating in such environments are confronted with images similar to those encountered in the standard environments, but these typically will have been generated by geometries different from those experienced in the standard environments.

The agents

Each agent was modeled as a sphere whose position and orientation in the environment were specified by a neural network control system (Figure 3A). The radius of the agent was 2 units, which is roughly half the height of the walls in the standard environments. A simulated 16×16 sensor matrix modeled as a flat

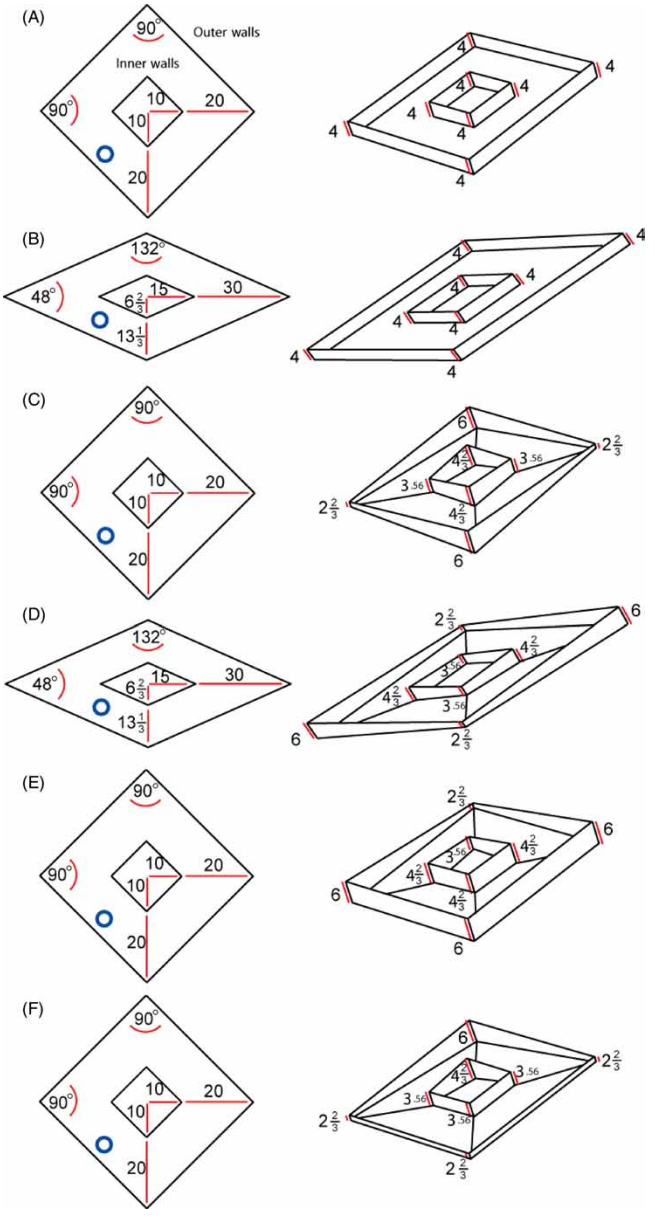

Figure 1. Environmental geometries. Left: bird's eye view showing the dimensions of the inner and outer walls of the arenas (the units are arbitrary). For comparison, an agent (blue circle) is 4 units in diameter. Right: Three-quarter view of the arenas showing the height of the walls at junctures. (A) The standard square environment. (B) The standard diamond environment. (C) The symmetrically distorted square environment, designed to project images similar to images generated by the diamond environment. (D) The symmetrically distorted diamond environment, designed to project images similar to those generated in the square environment. (E) The large asymmetrically distorted square environment; this arena is based on the standard square environment, but asymmetrically distorted so as to present larger wall panels compared to the symmetrically distorted environments. (F) The small asymmetrically distorted square environment is similarly distorted asymmetrically, but in such a way as to present smaller wall panels than the symmetrically distorted environments.

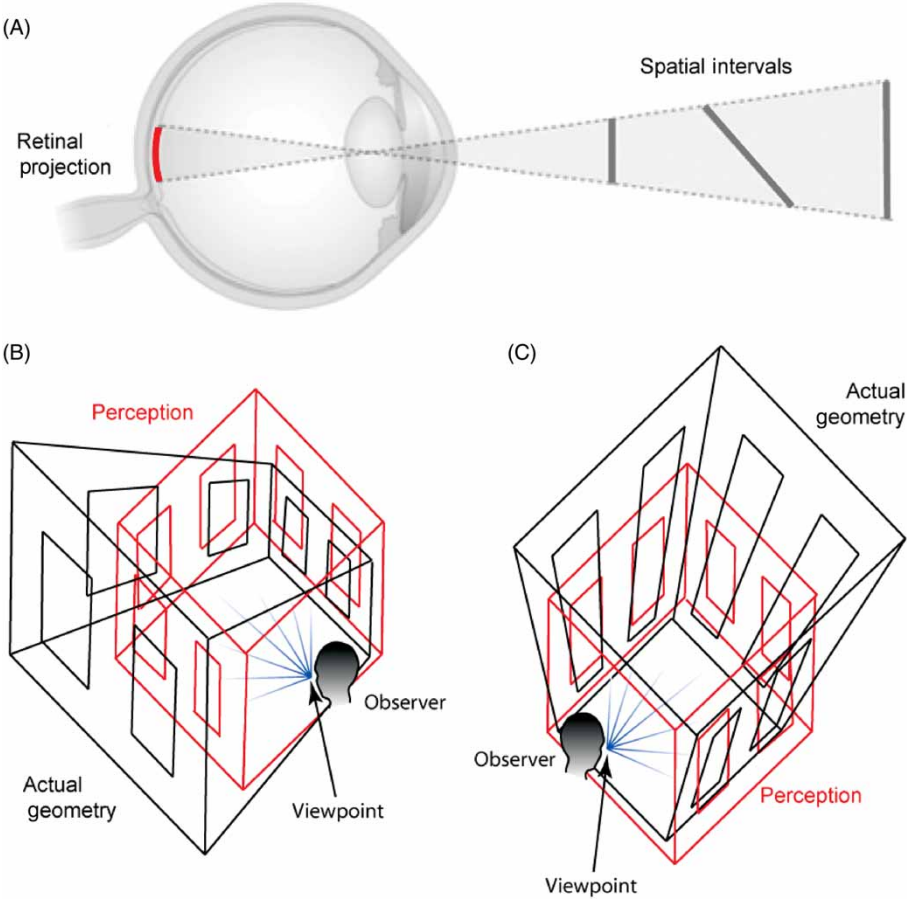

Figure 2. Problems arising in the process of projection. (A) In images projected from any environment the size, distance and orientation of objects are conflated in the image plane, as indicated in this diagram. This conflation illustrates the inverse optics problem. (B, C) Examples of the distorted rooms created by Ames, showing the conceptual basis of the distorted environments used in some of the experiments here. Although the room in (A) consists of an oblique floor, ceiling and rear wall (black outlines), and in (B) a small floor connected by sloping walls to a large ceiling (black outlines), an observer who views room through a monocular port perceives a rectangular room that is normal in appearance (red outlines), leading to anomalies of visually-guided behavior. (A is after Howe and Purves 2005a; B, C are after Ittleson 1952).

surface was located 1 unit from the center of the sphere in the frontal plane, the sphere itself being invisible. The agent's visual field was 45° in both azimuth and elevation, thus providing the agent with an angular resolution of ~2.8° per pixel. Images on the sensor matrix at any moment were simulated by a perspective transformation of the projected (illumination × reflectance) values arising from the agent's position and orientation in the environment at that time (Figure 3B). The resulting 256-element pattern of unprocessed light intensity values was the only input to the agent's neural network, and each behavioral response the agent made produced a new set of projected values.

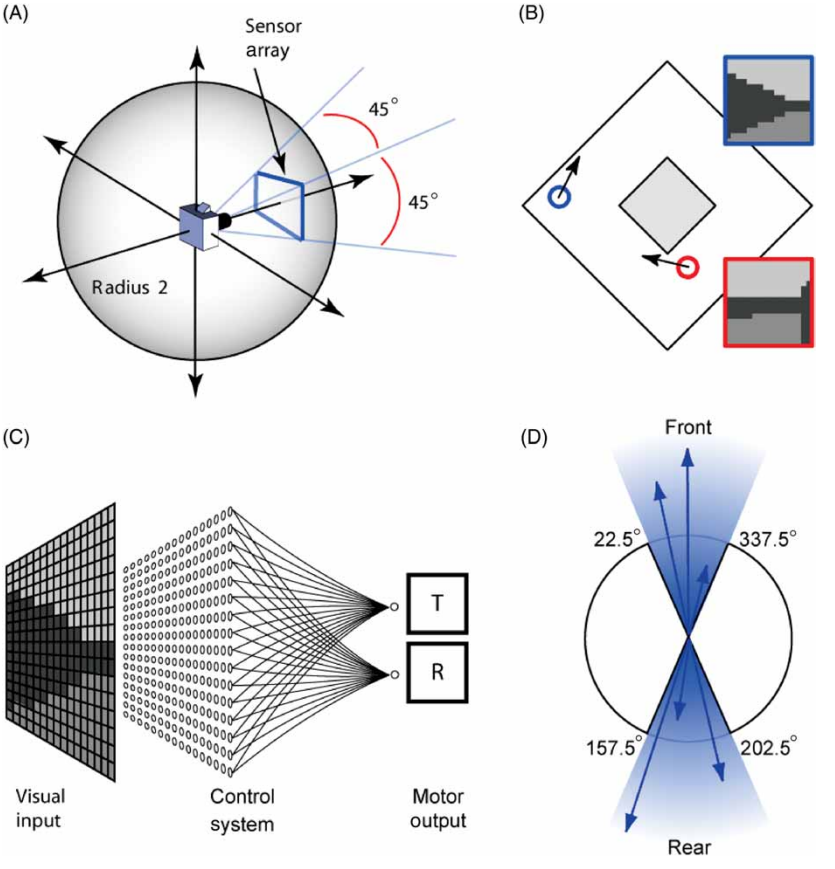

Figure 3. The artificial agents. (A) Each agent was modeled as a sphere with a radius of 2 units. The sensor array was located one unit from the center of the sphere. The images on the array were simulated by a perspective transformation of the objects in the agent's visual field, which was 45° in both azimuth and elevation. (B) Examples of images confronting an agent occupying two different positions in the standard arena (bird's eye view). The image outlined in blue is what the blue agent would have as a visual input when facing in the direction of the arrow; the image outlined in red is what the red agent would see. (C) The neural network control system. Each element of the projected visual input to an agent was fed into a node in a fully connected single layer neural network; each input was in turn connected to each of two output nodes. The two outputs were scaled to translation and rotation motor responses that determined the new position and orientation of the agent. (D) Possible movements of an agent (seen from above). Translational responses were limited to a distance no greater than 4 units (one "body" length), and rotation to no more than ±22.5°. The blue wedges indicate the possible new positions the center of the agent could occupy after a single behavioral response; the arrows show several specific examples of possible forward or backward movements from an initial position defined by the center of the agent.

The neural control system for each agent was a single-layer, feed-forward network with 256 inputs and two outputs (Figure 3C). Formally, the output of the network (y) is

$$y_{1:2} = g\left(w^T \boldsymbol{x} + w_0\right)$$

where $\boldsymbol{x}$ is the matrix of sensory inputs, w is the weight matrix, and w_0 is the bias. The activation function $g(a)$ is the hyperbolic tangent function

$$g(a) \equiv \frac{e^a - e^{-a}}{e^a + e^{-a}}$$

which produces a nonlinear output between ± 1 for any input.

The behavioral response of an agent was determined by scaling the two output values. To make the agents' movements biologically plausible (i.e., covering a variable but limited distance in the arena), the distance that any agent could translate in response to a given visual stimulus was scaled to ± 4 units, or one "body" diameter. Likewise, each agent's rotation was scaled to $\pm 22.5^\circ$, corresponding to half of the agent's field of view of 45° (Figure 3D). The scaled behavioral outputs were executed sequentially, first rotation and then translation, thus moving the agent in two dimensions. In the real world, friction, uneven surfaces, and other factors confound precisely predictable movements. To simulate this uncertainty, we added a small amount of noise to the rotation and translation components of the control system; the noise was sampled from a Gaussian distribution centered at 0, with a standard deviation equal to 5% of the total range of motion.

A collision was said to have occurred if the outer boundary of the agent's spherical body intersected the plane of a wall. To model collision with a wall, the agent was stopped just outside of the point of contact, simulating the translation of an object into a solid surface in the case where neither body has much elasticity.

The evolutionary algorithm

Chromosomes and network initialization. The real-valued weights of each neural network control system were considered the agent's "chromosome" (Holland 1975). Each chromosome was randomly initialized according to a Gaussian distribution centered at 0 with standard deviation of 1/16 (the reciprocal of the square root of the number of inputs to each output node; see Bishop 1995, pp. 261–262). In this way, we created populations of agents in which each individual was somewhat differently suited to the challenge of behaving successfully in the environment.

Populations and lifetimes. Populations comprised 100 agents, each with a unique chromosome. The agents occupied the environment one at a time (Nolfi and Floreano 2000), and each agent's lifetime was defined by 300 behavioral responses. To make sure that success did not depend on a particular starting position, the responses over an agent's lifetime were divided into 10 epochs, each of which began with the agent in a random position and orientation. The completed lifetime of all the individuals in the population defined a single generation.

Fitness criteria. The relative success of an agent's behavior (its fitness) was determined by two criteria: (1) the ability to avoid obstacles, measured by the ratio of the number of collisions (see above) to the number of behavioral responses in each agent's lifetime, a lower value indicating fewer collisions and greater fitness; and (2) more complete exploration of the environment, measured by the dispersion

of the positions occupied by each agent during its lifetime. For this determination we used the standard deviation of movements in the arena as an index, higher values indicating more dispersion, and thus greater fitness. The fitness of an individual agent for a single epoch was determined equally by both these measures according to the expression

$$\text{fitness} = \left(1 - \frac{c}{m}\right)\left(\frac{\text{sd}(x_{1:30})}{2} + \frac{\text{sd}(z_{1:30})}{2}\right)$$

where c is the number of collisions, m the number of movements, and $x_{1:30}$ and $z_{1:30}$ the coordinates of the 30 positions occupied during each of the 10 epochs in a lifetime. This fitness for each epoch was calculated and averaged to produce an overall lifetime fitness value for the agent in question.

Selection, reproduction and evolution. Once each member of the population had been assigned a fitness value, a selection mechanism identified the agents best suited to produce the next generation. To identify the parents that could reproduce, the population was evaluated by tournament selection (see Appendix A). The best agent of each pair was then added to a parent pool and the process repeated until 100 parents had been chosen. This standard stochastic method identified a pool of parents with many duplicates, but a higher average fitness than the population from which they had been drawn, thus mimicking the greater reproductive success of more fit biological parents (Eiben and Smith 2003). The chromosomes of pairs of parents selected at random from the pool were then recombined using single-point crossover (Holland 1975) and an uncorrelated mutation of a single step size (Bäck 1996; Eiben and Smith 2003; for details see Appendix B). The result of each such mating was a pair of offspring that was added to the new population of 100 agents that made up the next generation.

This sequence of evaluating each agent's behavioral actions in the environment, assigning a fitness value, selecting individuals for reproduction and applying genetic operators to produce a new population was repeated until the average fitness of subsequent generations reached a plateau.

Results

Evolution

Eight sets of evolutionary experiments, each consisting of five trials, were divided into two categories: visually guided and "blind." The visually guided experiments consisted of evolutions in each of the six environments illustrated in Figure 2. Control experiments included evolutions of "blind" agents in the two standard environments (Figure 2A and B).

In each set of visually guided experiments, five populations of 100 agents were placed in the relevant environment and allowed to evolve for 200 generations. In each of the five trials, the fitness of the agents generated by the relative success of the visually guided behavior improved as a function of the number of generations during evolution (Figure 4A). The increased fitness of all the populations was rapid and

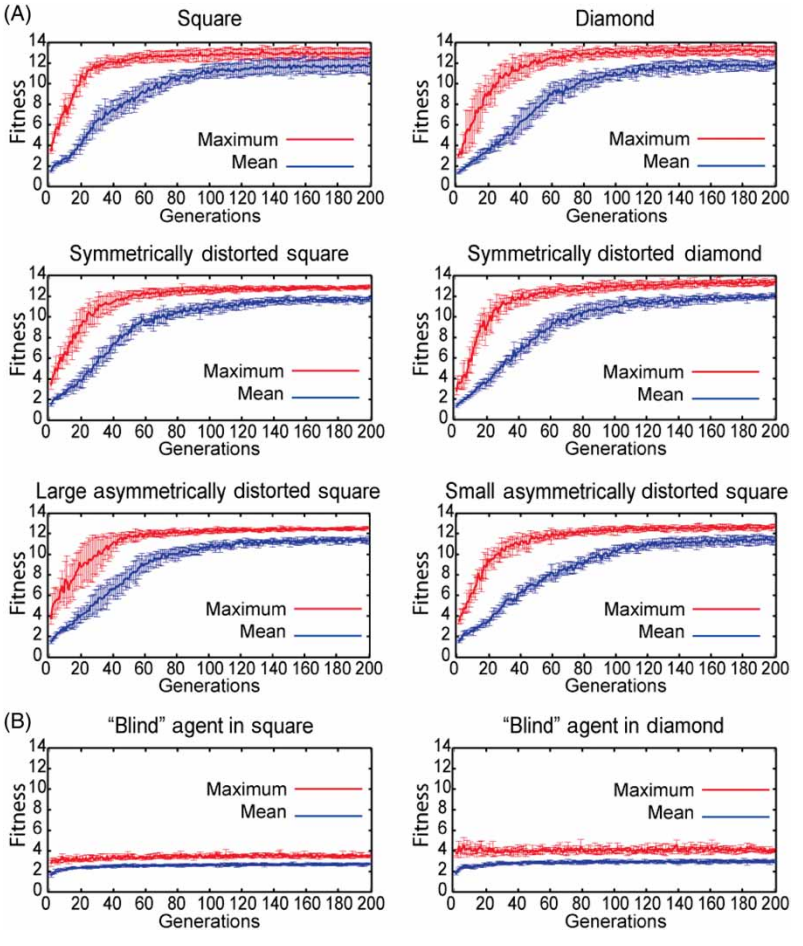

Figure 4. Fitness of agents over evolutionary time. Red points indicate the mean fitness of the most successful agent in each of the populations generated in the five trials of evolution carried out in each arena, plotted against generation number; blue points indicate the mean fitness of the evolved population. Bars show the SE. (A) Trials with visual agents. (B) Trials with ''blind'' agents placed in the standard square or diamond environment.

robust, with initial improvement that slowed and reached a plateau after ~50 generations; only slight improvement occurred over the 150 (or more) additional generations. The enhanced fitness was apparent by both measures of fitness; i.e., the frequency of collisions decreased, and agents explored the environment more extensively as evolution progressed.

To insure that non-visual factors played little or no role in the evolution of successful behavior, in the ''blind'' control experiments populations were evolved in the standard environments with luminance values sampled from a random uniform distribution between 0 and 1 as their only input. Little or no improvement in fitness was observed under these conditions (Figure 4B), indicating that structured visual input is essential to the evolution of increasingly fit behavior.

Dependence of behavior on the full input pattern

We next asked how the information accumulated during evolution is captured in agent's neural network control systems by examining the weight matrices of agents evolved in all six environments. There were no apparent patterns in the weights at the two output nodes of agents evolved in any of the environments, whether analyzed as individuals or in terms of the mean weights of all the evolved agents in five trials (i.e., 500 agents) (Figure 5A). Weight matrices were also evaluated in

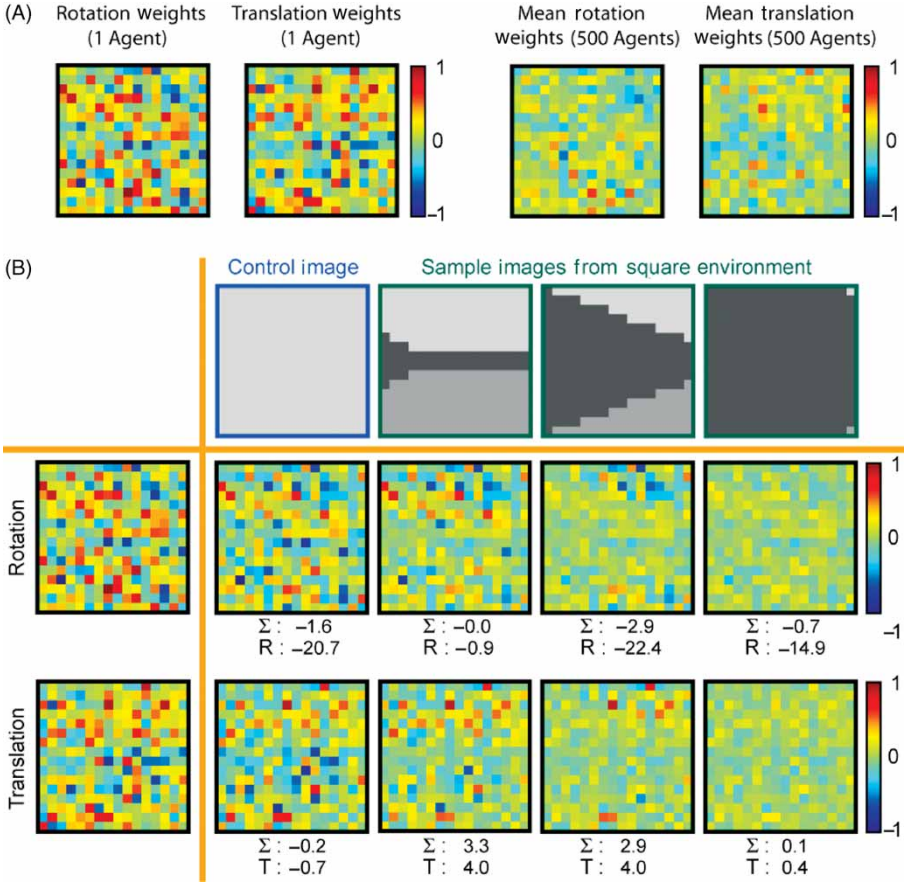

Figure 5. Analysis of neural network weights in evolved agents. (A) Neural network weights from a single evolved agent and the mean weights for all five populations of agents evolved in the standard square environment. The weights at the two output nodes (i.e., rotation and translation) are represented in a 16×16 grid corresponding to the agents' visual sensor arrays. No pattern is apparent in either the individual or mean distribution of network weights. (B) Analysis of network weight changes in response to the presentation of individual images. The test images used (top row) were multiplied by the rotation and translation weight matrices (left column) from an evolved agent; the resulting effect on the weight matrix is shown underneath each test image. The sum (Σ) of the weighted inputs and the response (rotation [R], or translation [T]) indicated underneath each matrix show that the input changes the neural network weights in a manner that leads to well adapted behavior, despite the lack of apparent overall patterns in the matrices.

response to specific images generated by the virtual environments (Figure 5B). It is clear that the luminance values projected onto the sensor arrays modulated the values network weights, thus changing the summed activity at the output nodes and the agents' behavioral responses. However, the lack of patterns in Figure 5 indicates that no simple heuristic can explain the improved performance of the agents over the course of evolution. Agents appear to use the full range of information on their sensor arrays to link structured images to well-adapted behavioral responses.

Dependence of improved fitness on behaviorally relevant sources

It is reasonable to suppose that the dependence of improved fitness on the full range of information in images is because the agent associates the image presented at any given moment with a response that is appropriate to the environmental geometry underlying that image. To evaluate this supposition, identical copies of 10 evolved agents were placed at two different random locations in the various environments at the beginning of each of the 10 epochs that defined a lifetime (Figure 6A). One agent received a stream of visual input appropriate to its positions and orientations in the arena, as before. The second agent, starting from a different random location, received visual input from its twin that was thus not appropriate to its situation in the environment. The fitness of the two copies of the agent was then compared after a single lifetime (300 behavioral responses).

The fittest 10% of agents in each evolved population was evaluated in this way, i.e., 50 agents from each of the eight evolutionary experiments (see Figure 4; the ''blind'' experiments were included as controls) (Figure 6B). The results indicate that agents with appropriate visual input were far more successful than the yoked twins whose visual input was randomly related to their local circumstances. Thus, successful agent behavior in response to sensory stimuli depends not on structured stimuli *per se*, but on its relation to the geometry of the underlying source.

Understanding the strategy being used

These results indicate that the evolution of improved visually guided behavior is a consequence of associations between images on an agent's sensor array and relatively successful behavior. This conclusion, however, does not indicate how the agents make this link, i.e., the strategy that instantiates this association.

The nature of the problem for the agents. To understand the basis on which the agents associate images and source-appropriate behavior, 20 000 randomly sampled agent positions and orientations (called *poses* in computer vision and robotics) were co-registered with the images encountered by an agent at that pose in each of the six environments in Figure 1 (by co-registered we mean that the image and pose are matched together as a pair). A given image typically corresponded to a number of poses in each environment, and only a small subset of images was uniquely associated with a particular pose. Thus multiple poses ($\sim$3 on average; range = 2.76–3.41) provided agents with the same image.

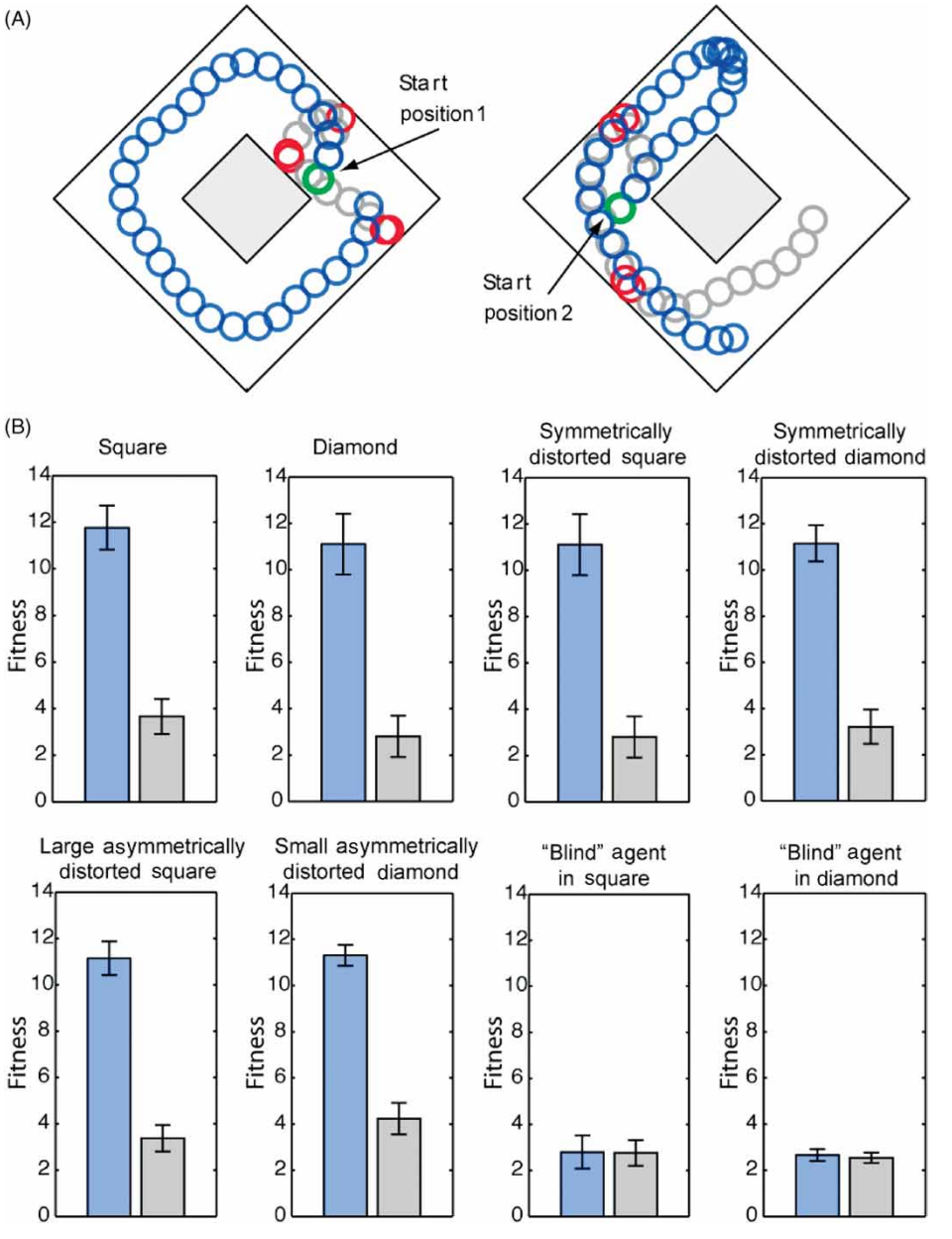

Figure 6. Behavior of evolved agents with and without visual input appropriate to the source of the images confronting them. (A) Copies of the same evolved agent were placed at two random locations (1, 2) in the standard arena (bird's eye view); red circles indicate collisions and blue circles the sequential positions of the agents in response to a stream of 30 visual images appropriate to the generative sources in the environment during one of the 10 epochs comprising a single lifetime. The gray circles illustrate the behavior of the same agents in receipt of inappropriate visual information (see text). (B) The fitness of the agents, including "blind agents," tested over a complete lifetime. The blue bars illustrate mean fitness of evolved agents exposed to images derived from the source confronting them; the grey bars show the mean fitness of agents exposed to images that were inappropriate to the actual image source (mean fitness of 50 agents, 30 responses each; bars show SE).

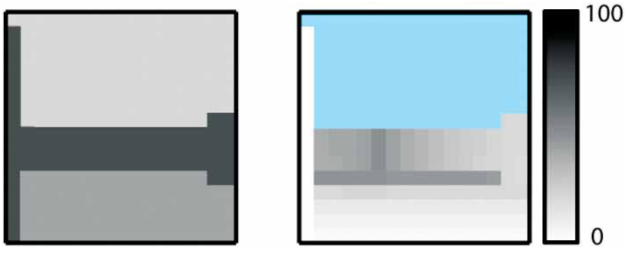

Figure 7. Example of a depth map. The image on the left is a typical luminance pattern projected on an agent's sensor array. The corresponding depth map on the right shows the distance of the generative sources of the image; the lighter elements are closer, as indicated by the calibration bar (arbitrary units); the sky is shown in blue since its direction and distance are undefined, and not considered. The mean distance to surfaces described in the text is the mean value of all the defined distances in each depth map.

To evaluate how the agents might be dealing with these challenges, each of the several poses associated with an image was translated into a depth map (Figure 7). The mean distance to the surfaces in the depth map was then used as an overall measure of the local geometry for a particular pose. Given this information we could quantify the distribution of source distances associated with the images generated in any one of the environments in Figure 1.

The uniform construction of the two standard environments ensured that the distribution of sources associated with an image had a small standard deviation, signifying similar average surface distances, and thus relatively similar underlying geometrical sources (Figure 8A and Table I). In the distorted environments, however, the standard deviation was 4–5 times higher, indicating statistical dispersion of the sources underlying an image; as a result, the same images mapped onto a range of geometrically different configurations (Figure 8B and C). Thus agents in distorted environments clearly contend with an inverse mapping problem that has no logical solution.

Evidence from agent behavior in novel environments. If agent behavior is indeed determined by the relationship between images and generative sources in the environment in which they evolved, then their behavior in a novel environment should deviate predictably in accord with the relationships between images and their sources in the new environment. Of the 20 000 images randomly sampled in each environment, an image exactly matched one or more images sampled from one of the other environments ~75% of the time, showing that the environments generate broadly similar images. Because of the different geometrical configurations of the environments, however, stimuli encountered by the agents could be generated by very different sources in one environment compared to another (Figure 8). To understand how the distribution of stimuli and sources changes from environment to environment, each distinct image sampled from the standard square environment was assigned a label λ defined as

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} \left|(|x_i| - |z_i|)\right|,$$

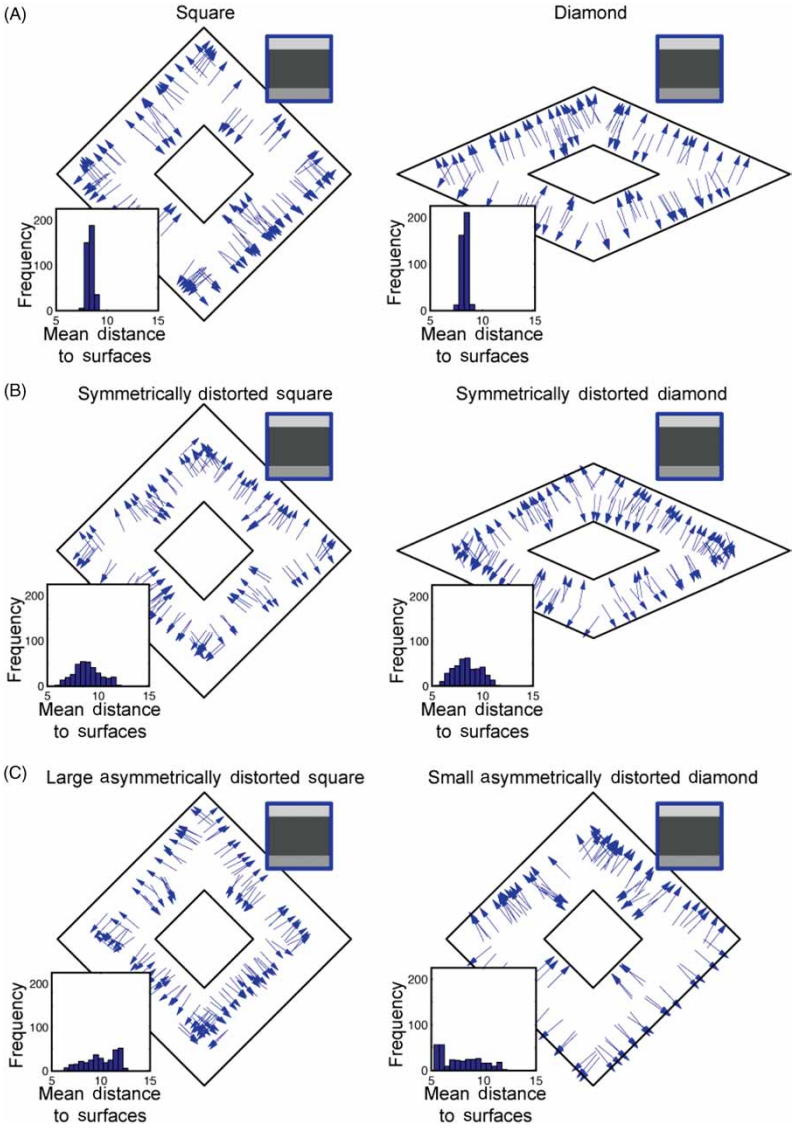

Figure 8. The distribution of poses co-registered with a particular image in each of the simulated environments (arrows are a random subset of this distribution; agent location is specified by the arrow tail, and orientation by the arrow direction). The corresponding image is outlined in blue; inset graphs show the frequency distribution of poses as a function of mean distance to surfaces underlying the image. (A) The two standard environments. The poses co-registered with the image for both the square and diamond arenas form symmetrical, peaked distributions; the most frequently sampled mean distance from surfaces is ~9 units. (B) The symmetrically distorted environments. The several poses co-registered with the same image in the symmetrically distorted square and diamond environments form wider distributions, indicating a larger range of possible geometrical configurations underlying the image. The most frequently sampled mean distance is again ~9 units. (C) The asymmetrically distorted environments. The poses co-registered in these environments generate wide, asymmetric distributions. The most frequently sampled mean distance from surfaces is ~12.5 units for large asymmetrically distorted square environment, and ~5.5 units for the small asymmetrically distorted square environment.

Table I. Characteristics of the distributions of mean distances to surfaces underlying images in different environments.

Environment	Mean SD	Mean skewness
Square	0.06	0.01
Diamond	0.06	0.00
Symmetrically distorted square	0.22	−0.02
Symmetrically distorted diamond	0.21	−0.01
Large asymmetrically distorted square	0.30	−0.06
Small asymmetrically distorted square	0.33	0.04

Table shows statistical dispersion and asymmetry of the frequency distributions of the mean distance to surfaces underlying the images generated in each environment.

where n is the number of poses associated with the particular image, and $x_{1:n}$ and $z_{1:n}$ are the coordinates that specify the 2-dimensional location of each pose. Thus, λ is a measure of where the mean of a distribution of the poses associated with a particular image is positioned in the eight symmetrically identical sectors in the standard square environment, as illustrated in Figure 9A. Each pose was assigned the same label as its corresponding image, and the average values of the labels at each position plotted. The same procedure was carried out for the 20 000 poses sampled from each of the other five environments, each pose being assigned the label appropriate to the corresponding image as it appeared in the square environment (Figure 9B–F). In this way, the average change in projected images as a function of agent position could be evaluated across environments.

Figure 9 shows that the distribution of images changes predictably in these simple and largely symmetrical environments as a function of distortion: similar images were further from tall walls and closer to short walls in the distorted environments compared to their distances in the standard square environment (cf. Figure 9E and F to Figure 9A). If the behavioral success of the evolved agents is based on the relationships between images and the sources of those images in the environment experienced during evolution, then behavior in their native environment should reflect the distributions of sources underlying images (see Figure 8 and Table I). Moreover, behavior in a novel environment should be predicted by the relationship of the native environment to the novel one determined by the analyses in Figure 9. To evaluate these suppositions, the fittest 10% of agents in each population were placed in each of the six different environments in Figure 1 for a single lifetime and their behavior analyzed.

Behavior of agents evolved in the standard square environment. The agents evolved in the square environment tended to circle the arena either clockwise or counter-clockwise, maximizing fitness by staying as close to the outer walls as possible while minimizing collisions (Figure 10A, first column; see also Figure 4A). These agents followed the same strategy when placed in the diamond environment (Figure 10A, second column), occasionally colliding with the walls in the acute corners of the diamond, presumably because this geometry had not been encountered in the square environment. In the symmetrically distorted square environment (Figure 10A, third column), the agents followed patterns of movements more

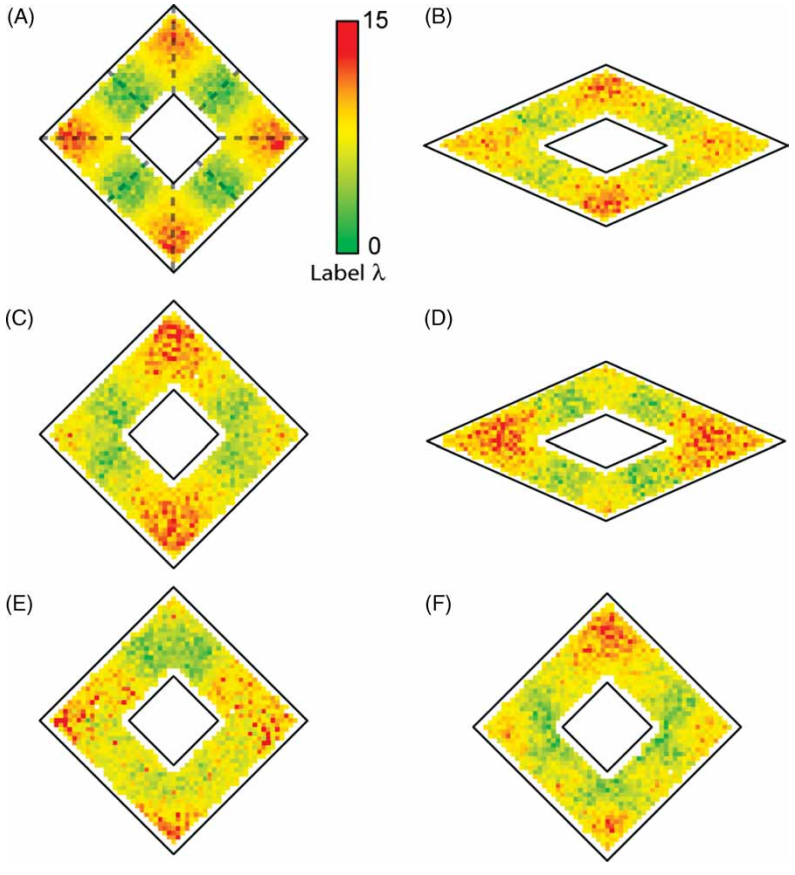

Figure 9. The relationships between images and their sources in different environments. (A) For this analysis the standard square environment was divided into eight symmetrically identical sectors (dotted lines). Colored pixels represent the mean λ value of all poses for each 1×1 unit location in the environment, thus tracking the distribution of visual stimuli (see text). (B) The diamond environment is similar to the square environment in the uniformity of its construction. As a result, many of the same images are presented to agents in comparable positions throughout the environment. (C) The symmetrically distorted square environment. The sloping walls and floors change the distribution of images such that images generated by the shorter walls are associated with surfaces that are closer to the agent than the sources of same stimuli in the square environment. Likewise, the images falling on the agents' sensors from larger walls are farther from the underlying surfaces than in the square environment. (D) The symmetrically distorted diamond environment. As with the symmetrically distorted square, the unusual geometry changes the relationship of stimuli and their generative sources compared to the same images in the square environment. (E, F) The large and small asymmetrically distorted square environments. In these cases, the distribution of stimuli is skewed, the mean distance to the generative source for majority of images being either closer or farther compared to images and their sources in the square environment.

appropriate to the diamond environment, in agreement with the shifted distribution of stimuli with respect to their native square environment (Figure 9C). Thus they overestimated the distance from the walls when these surfaces were taller and the floor sloped downwards, and underestimated the distance to the walls when the

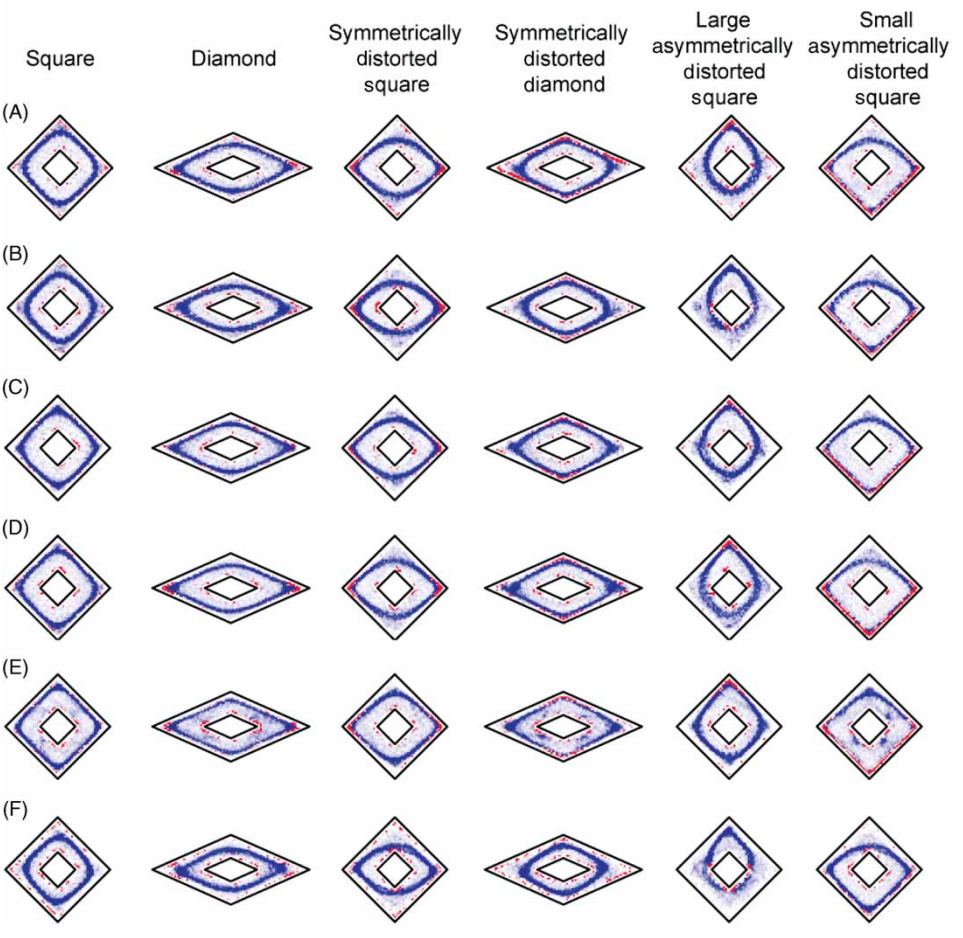

Figure 10. Density of positions occupied by 10 agents tested over a single lifetime in each of the six environments. Blue pixels represent the frequency that the center of an agent occupied for each 1×1 unit location in the environment (darker blue represents higher density, defined by the number of positions in corresponding grid squares overlain on each environment; red indicates collisions). Agents evolved in the square environment (A), the diamond environment (B), the symmetrically distorted square environment (C), the symmetrically distorted diamond environment (D), the large asymmetrically distorted square environment (E), and the small asymmetrically distorted square environment (F) were tested in each of the novel environments.

floor sloped upwards and the walls were shorter (Figure 1C). In the symmetrically distorted diamond environment (Figure 10A, fourth column), the agents again moved closer to the walls when the walls were shorter and the floor sloped upwards, and stayed farther away when these surfaces were taller and the floor sloped downwards (Figure 1D). This behavior can also be understood in terms of the change in relationships between images and their sources of the symmetrically distorted diamond environment relative to the relationships the agents experienced in their native square environment. The behavior was much the same in the asymmetrically distorted environments (Figure 10A, fifth and sixth columns;

see Figure 1E and F), the differences in the agent's positions again being predicted by the changes in image–source relationships.

Behavior of agents evolved in the standard diamond environment. The results for agents evolved in the diamond environment were similar to those for agents evolved in the square environment (Figure 10B; see also Figure 4A), as expected from the fact that the relationships between images and their sources in the square and diamond environments are typically similar (see Table I and Figures 8 and 9). However, the maximum density of positions occupied by the evolved agents moved slightly inward from the walls in every environment tested, indicating that although the majority of image–source relationships are identical in these tests, the behavior is not quite the same as that of the agents that had evolved in the square environment. A reasonable explanation of this difference would be that optimizing fitness in the diamond environment requires that agents evolve to keep farther away from the potentially dangerous acute angles in the environment. In any event, the behavior of agents evolved in the two standard environments (Figure 10A and B) indicates that the control systems link stimuli to behavior appropriate to the generative sources of stimuli in their native environments. This result is not surprising, since images in the standard environments typically correspond to a particular source with a high probability (see Figure 8A and Table I).

Behavior of agents evolved in symmetrically distorted environments. It was therefore of particular interest to examine in the same way the behavior of agents that had evolved in distorted environments (Figure 1C–F). Surprisingly, the density of positions occupied and the locations of collisions for agents evolved in both the symmetrically distorted square and diamond environments were similar to the behavior of agents evolved in the standard square environment (Figure 10C and D; cf. Figure 10A and B). The positions occupied by these agents were not optimally distributed; that is, rather than circling the arena while staying close to the outer walls, the agents moved away from the tall walls and sometimes collided with the shorter walls. Recall, however, that each image in the symmetrically distorted environments could have been generated by multiple sources with different underlying geometries (Table I). For many images in these environments, the distribution of possible sources underlying a given image formed a symmetrical distribution with the most frequently encountered underlying source being identical to the underlying sources typically found in the two standard environments (see Figure 1C and D for geometries, Figure 8B for an example of the distribution of poses underlying an image, and Table I for the statistical of distribution of sources underlying images in the symmetrically distorted environments). Evidently, these agents also evolved control systems that linked images to behavior appropriate to the most probable geometries underlying the images that confronted them.

Behavior of agents evolved in asymmetrically distorted environments. The two asymmetrically distorted environments (Figure 1E and F) are similar to the symmetrically distorted square, except that the large asymmetrically distorted square has a higher frequency of tall walls, and the small asymmetrically distorted square has a higher frequency of short walls. These environments thus change the

distribution of possible sources of images such that the more frequently encountered geometries are skewed towards taller or shorter walls, respectively (see Figure 8C and Table I).

Agents evolved in the large asymmetrically distorted environment exhibited behavior different from that of the other evolved agents (Figure 10E). Although most of the positions in the large asymmetrical distorted square environment were well dispersed and avoided walls, the remaining positions were more peripheral in the arena, leading to a higher frequency of collisions. Thus these agents appeared to have associated images with behaviors appropriate to the most frequently occurring and thus probable geometries underlying images in their native environment. Agents evolved in the small asymmetrically distorted square environment behaved in a similar way (Figure 10F). In this case, agents again followed well-dispersed paths most of the time, but tended to keep away from the outer wall.

In summary, evolution in the two asymmetrical environments can also be understood in terms of a strategy in which agents contend with the inverse optics problem by associating images with their probable underlying sources.

Discussion

The purpose of these experiments was to assess the idea that biological agents have evolved successful visually guided behavior solely on the basis of feedback that promotes neural associations between images and behavior. The reason this approach must be considered as a basis for the emergence of animal vision is the fundamental challenge presented by the inverse problem, i.e. the inability of retinal images to unambiguously specify their generative sources in the environment. Since this problem has no analytical solution, and since the success of visual animals depends on dealing effectively with image sources, an empirical solution seems inevitable. The results here demonstrate that evolving autonomous agents – simulated in this case – can indeed improve their fitness by instantiating the relationship between images and behavior appropriate to the most probable underlying geometry.

Relevance to human perception

The strategy evolved agents use to respond successfully to inherently ambiguous images is especially pertinent to explaining the anomalous way that humans (and presumably other visual animals) perceive the geometry of their environment. Image analysis and estimation techniques applied to natural image databases have already shown that many otherwise puzzling aspects of animal vision can be explained by the empirical relationship between images and their real-world sources. For example, the grouping of visual contours (Geisler et al. 2001), the perception of line length as a function of orientation (Howe and Purves 2002), the anomalous perception of visual space (Yang and Purves 2003), size contrast effects (Howe and Purves 2004, 2005b), brightness (Yang and Purves 2004) and color contrast and constancy phenomena (Lotto and Purves 2000; Long and Purves 2003; Long et al. 2006) can all be rationalized on the basis of past experience with natural sources. Taken together, these observations suggest that the generation of visual percepts,

and thus visual processing, is fundamentally empirical. By showing that evolved agent behaviors are appropriate to the probable source of the image rather than the particular characteristics of the image on an agent's sensor array, the present results indicate how, in principle, animal vision could have come about on an empirical basis, how it resolves the inverse problem, and why human vision is so often at odds with the metrics of the real world. Evidently the strategy used to generate successful behavior in these experiments is predicated on accumulated statistics about what images have turned out to signify in the environment rather than image features as such.

The evolutionary model of visual behavior in the paradigm we used is particularly relevant to two well-known anomalies in human vision: the specific distance tendency and the perceptions elicited by the Ames room. Studies of perceived distance, perhaps the simplest aspect of visual space, show that the apparent distance of objects bears no simple relation to their physical distances from the observer (Yang and Purves 2003). Thus, when subjects are asked to make judgments with little or no contextual information, objects, whatever their actual distance, are typically perceived to be 2–4 m from the observer, which is the most probable distance of surfaces in natural visual environments. The present results show that agents evolving behavior solely in response to visual inputs are similarly biased toward behavior that reflects the probable distance of the surfaces underlying images.

A probabilistic strategy for visually guided behavior evolved empirically would also explain the anomalous way that human observers experience an Ames room (Figures 2B and C). Ames noted that ''If an observer is given a pointer and asked to touch various parts of the room, he cannot do so accurately and quickly but behaves quite awkwardly, unexpectedly hitting walls, floor or ceiling'' (Ittleson 1952, p. 184). Because agents in the present experiments evolved behavior that corresponds to the likely source of stimuli in their native environment, when placed in an arena similar to an Ames room they also produced inappropriate behavior, evidenced by increased collisions with surfaces and inefficient exploration of the arena.

Implications for the mechanics of vision

Historically, approaches to understanding the mechanisms of vision in either biological or artificial systems fall into two broad categories: (1) physiological and anatomical studies of animal vision, typically in the framework of exploring how neuronal receptive field properties are related to reporting various image features; and (2) algorithmic approaches in machine vision. Exploring visual processing in terms of detecting and extracting features in retinal images has been the dominant theme over much of the last century; more recently, however, this perspective has undergone considerable revision in the face of evidence that neither receptive field properties nor visual percepts can be explained in any simple way using a scheme of feature detection (Knill and Richards 1996; Barlow 2001; Rao et al. 2002; Purves and Lotto 2003; Howe and Purves 2005b).

As a result, the conceptual framework of feature detection has begun to give way to the idea that visual processing must be intimately related to the statistics of natural images. For example, enhanced responses to contrast boundaries

(Hubel and Wiesel 1962) as well as color-opponency responses (Hubel and Wiesel 1968) are now known to be correlated with the basis functions of efficient statistical representations of natural images (Olshausen and Field 1996; Bell and Sejnowski 1997; Caywood et al. 2001; Lee et al. 2002). Moreover, some anatomical characteristics of the primary visual cortex, e.g., preferential horizontal connections between neurons tuned to similar orientations (Bosking et al. 1997), are also consistent with an incorporation in visual processing of natural image statistics (Geisler et al. 2001; Girmin et al. 1999). The present results, however, argue that the behavioral output of visual processing systems are not based on the statistical structure of images *per se*, but on the relationship between images and their possible sources. The statistical structure of images *per se* seems more likely to drive the efficiency of stimulus encoding rather its behavioral or perceptual consequences.

By the same token, most approaches to machine vision have been based on algorithms that detect, segment, localize, recognize and/or track objects in the image plane. Attempts to recover 3D structure from images have typically employed on two technical strategies: deriving structure from motion and stereopsis (Forsyth and Ponce 2003). Both of these approaches rely on a comparison of features in multiple images. The only well-known method of inferring depth from a single image is Horn's ''shape from shading'' algorithm (Horn 1975), a technique that creates models of image formation and then inverts them, thus solving for depth. As a result of the inverse problem, however, such models are highly under constrained and require simplifying assumptions (e.g., Lambertian surface reflectance) that are not characteristic of images generated by natural environments (Potetz and Lee 2003). It is generally agreed that the efficacy of machine vision to date falls far short of animal vision. The present results suggest why this is the case, namely that accumulated empirical information is necessary for successful vision.

If the visual brain has indeed evolved in more or less the same way as the neural controllers in the simulations we describe, then a deeper understanding of animal vision and improved machine vision will require understanding the empirical relationship between images and sources. The present highly simplified demonstration of the feasibility of vision on a wholly empirical basis should thus encourage more sophisticated simulations, and further exploration of both biological and machine vision in these terms.

Acknowledgements

We are grateful to Beau Lotto, Jim Voyvodic, Debbie Ross, Nestor Schmajuk, Kyongje Sung and Bill Wojtach for many helpful suggestions. This work was supported by the NIH, the AFSOR and the Geller Endowment.

Competing interests statements: The authors declare that they have no competing financial interests.

Appendix A: Tournament selection

In typical evolutionary algorithms, competition is used to determine which chromosomes will contribute to the creation of the individuals that comprise the

next generation. One such approach – the one we used in this work – is tournament selection, an efficient selection method that stochastically samples parents from a population without any global knowledge of the population's fitness distribution. To select each parent, k individuals are randomly picked from the population to form a "tournament." By comparing fitness values, the best of these k individuals is selected and the process is repeated until there are enough parents to sire the next generation (100 in our case). The tournament size, k, can be set to vary the selection pressure in the algorithm; thus the larger the value of k, the higher the probability that highly fit individuals will be selected. In the work reported here we chose $k = 5$, which is considered to be a relatively large value that produces a moderately high selection pressure. For more information on tournament and other selection methods see Eiben and Smith (2003).

Appendix B: Uncorrelated mutation with a single step size

After offspring were produced through single-point crossover, an uncorrelated Gaussian mutation with a single step size was used to introduce variation into the offspring. For each weight w_i in the chromosome, a small value drawn from a Gaussian distribution centered at 0 was added to form a new weight w_i'. The standard deviation of the distribution used to mutate each weight is called the mutation parameter σ. This parameter is an additional variable in each chromosome that is also allowed to evolve, thereby changing the sampling distribution for mutation values. In the present work, the mutation parameter σ is itself mutated each time an individual reproduces through multiplication by a variable e^{Γ}. Γ is a random variable drawn from a Gaussian distribution with mean 0 and standard deviation equal to 1/ sqrt(n), where n is the number of elements in the chromosome. This parameter can be interpreted as a learning rate in evolutionary paradigms of the sort we used (Eiben and Smith 2003; for more information about learning rates in neural networks see Bishop 1995).

The co-evolution of mutation parameters is a biologically plausible way to vary the relative amount that mutation can change a chromosome, and different mutation parameters are more or less well suited to different circumstances. For example, large mutations early in evolution can speed the process by making large changes to a chromosome, while smaller mutations late in evolution can refine a solution without risking a substantial loss of fitness. Empirically, we found that adding a co-evolving mutation parameter increased the speed of evolution in our problem space by a factor of about 10. Thus, maximizing fitness required only $\sim$200 generations compared to $\sim$2000 generations required when we used a fixed mutation parameter.

References

Bäck T. 1996. Evolutionary algorithms in theory and practice: Evolution strategies, evolutionary programming, genetic algorithms. New York: Oxford University Press.

Barlow H. 2001. Redundancy reduction revisited. Network: Computation in Neural Systems 12:241–253.

Bell AJ, Sejnowski TJ. 1997. The 'independent components' of natural scenes are edge filters. Vision Research 37:3327–3338.

Bosking WH, Zhang Y, Schofield B, Fitzpatrick D, et al. 1997. Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. Journal of Neuroscience 17(6):2112–2127.

Bishop CM. 1995. Neural networks for pattern recognition. New York: Clarendon Press.

Caywood MS, Willmore B, Tolhurst DJ, et al. 2001. The color tuning of independent components of natural scenes matches V1 simple cells. Journal of Vision 1:65a.

Cliff D, Harvey I, Husbands P. 1997. Artificial evolution of visual control systems for robots. In: Srinivisan M, Venkatesh S, editors. From living eyes to seeing machines. Oxford: Oxford University Press.

Dain RA. 1998. Developing mobile robot wall-following algorithms using genetic programming. Applied Intelligence 8(1):33–41.

Eiben AE, Smith JE. 2003. Introduction to evolutionary computing (Natural Computing Series). New York: Springer.

Floreano D, Mondada F. 1994. Automatic creation of an autonomous agent: Genetic evolution of a neural-driven robot. From animals to animats. Third International Conference on Simulation of Adaptive Behavior. Cambridge, MA: MIT Press Bradford Books. pp 421–430.

Floreano D, Mondada F. 1996. Evolution of homing navigation in a real mobile robot. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics 26:396–407.

Floreano D, Mondada F. 1998. Evolutionary neurocontrollers for autonomous mobile robots. Neural Networks 11:1461–1478.

Forsyth D, Ponce J. 2003. Computer vision: A modern approach. Upper Saddle River, NJ: Prentice Hall.

Geisler WS, Perry JS, et al. 2001. Edge co-occurrence in natural images predicts contour grouping performance. Vision Research 41(6):711–724.

Gibson JJ. 1979. The ecological approach to visual perception. Hillsdale, NJ: Lawrence Erlbaum Associates.

Goldberg DE. 1989. Genetic algorithms in search, optimization and machine learning. Reading MA: Addison Wesley.

Helmholtz HV. 1924. Helmholtz's treatise on physiological optics. New York: Optical Society of America.

Holland JH. 1975. Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. Cambridge, MA: MIT Press.

Horn BKP. 1975. Obtaining shape from shading information. In: Winston PH, editor. The psychology of computer vision. New York: McGraw-Hill.

Howe CQ, Purves D. 2002. Range image statistics can explain the anomalous perception of length. Proceedings of the National Academy of Sciences USA 99:13184–13188.

Howe CQ, Purves D. 2004. Size contrast and assimilation explained by the statistics of scene geometry. Journal of Cognitive Neuroscience 16:90–102.

Howe CQ, Purves D. 2005a. Perceiving geometry: Geometrical illusions explained in terms of natural scene geometry. New York: Springer.

Howe CQ, Purves D. 2005b. Natural scene geometry predicts the perception of angles and line orientation. Proceedings of the National Academy of Sciences USA 102:1228–1233.

Howe CQ, Lotto RB, Purves D. 2006. Comparison of Bayesian and empirical ranking approaches to visual perception. Journal of Theoretical Biology 241:866–875.

Hudel DH, Wiesel TN. 1968. Receptive fields and functional architecture of monkey striate cortex. Journal of Physiology 195(1):215–243.

Hudel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160:106–154.

Ittleson W. 1952. The Ames demonstrations in perception. Princeton, NJ: Princeton University Press.

Knill D, Richards W. 1996. In: Knill DC, Richards W, editors. Perception as Bayesian inference. Cambridge, UK: Cambridge University Press.

Lee TW, Wachtler T, Sejnowksi TJ, et al. 2002. Color opponency is an efficient representation of spectral properties in natural scenes. Vision Research 42(17):2095–2103.

Long F, Purves D. 2003. Natural scene statistics as the universal basis for color context effects. Proceedings of the National Academy of Sciences 100(25):15190–15193.

Long F, Yang ZY, Purves D. 2006. Spectral statistics in natural scene predict hue, saturation, and brightness. Proceedings of the National Academy of Sciences 103:6013–6018.

Lotto B, Purves D. 2000. An empirical explanation of color contrast. Proceedings of the National Academy of Sciences 97:12834–12839.

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

Nolfi S, Marocco D. 2000. Evolving visually-guided robots able to discriminate between different landmarks. In J-A Meyer, A Berthoz, D Floreano, HL Roitblat and SW Wilson (Eds.) From Animals to Animats 6. Proceedings of the VI International Conference on Stimulation of Adaptive Behavior. Cambridge, MA: MIT Press. pp. 413–419.

Nolfi S, Marocco D. 2002. Evolving robots able to visually discriminate between objects with different size. International Journal of Robotics and Automation 4:163–170.

Olshausen BA, Field DJ. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381:607–609.

Pfeifer R, Scheier C. 1999. Understanding intelligence. Cambridge, MA: MIT Press.

Potetz B, Lee T. 2003. Statistical correlations between two-dimensional images and three-dimensional structures in natural scenes. Journal of Optical Society of America 20:1292–1303.

Purves D, Lotto RB. 2003. Why we see what we do: An empirical theory of vision Sunderland, MA: Sinauer Associates.

Purves D, Williams SM, Nundy S, Lotto RB. 2004. Perceiving the intensity of light. Psychology Review 111:142–158.

Rao RPN, Olshausen BA, Lewichi MS. (Eds) 2002. Probabilistic models of the brain: Perception and neural function Cambridge, MA: MIT Press.

Salomon R. 1996. Increasing adaptivity through evolution strategies. Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior. pp 411–420.

Smith TMC. 1997. Blurred vision: Simulation-reality transfer of a visually guided robot. Evolutionary robotics: First European Workshop, EvoRob'98. pp 152–164.

Wertheimer M. 1923/1938. A source book of gestalt psychology. London: Routledge & Kegan Paul.

Wyss R, König P, Verschure PFMJ. 2006. A model of the ventral visual system based on temporal stability and local memory. PLoS Biology 4:836–843.

Yang Z, Purves D. 2003. A statistical explanation of visual space. Nature Neuroscience. 6:632–640.

Yang Z, Purves D. 2004. The statistical structure of natural light patterns determines perceived light intensity. Proceedings of the National Academy of Sciences 101:8745–8750.