Two Examples of Active Categorisation Processes Distributed Over Time

Tomassino Ferrauto **Elio Tuci** **Marco Mirolli** **Gianluca Massera** **Stefano Nolfi**
ISTC-CNR, Via San Martino della Battaglia, 44, 00185 Rome, Italy
{tomassino.ferrauto, elio.tuci, marco.mirolli, gianluca.massera, stefano.nolfi}@istc.cnr.it

Abstract

Active perception refers to a theoretical approach grounded on the idea that perception is an active process in which the actions performed by the agent play a constitutive role. In this paper we present two different scenarios in which we test active perception principles using an evolutionary robotics approach. In the first experiment, a robotic arm equipped with coarse-grained tactile sensors is required to perceptually categorize spherical and ellipsoid objects. In the second experiment, an active vision system has to distinguish between five different kinds of images of different sizes. In both situations the best individuals develop a close to optimal ability to discriminate different objects/images as well as an excellent ability to generalize their skills in new circumstances. Analyses of evolved behaviours show that agents are able to solve their tasks by actively selecting relevant information and by integrating these information over time.

1 Introduction

Traditionally, Cognitive Science and Artificial Intelligence tended to view intelligence as the result of a chain of three information processing systems, constituted by perception, cognition, and action. According to this view, the perception system operates by transforming the information gathered from the external world (sensations) into internal representations of the environment itself. The cognitive system operates by transforming these internal representations into plans (i.e. strategies for achieving certain goals in certain contexts). Finally, the action system transforms plans into sequences of motor acts. This is what Susan Hurley has labelled the "Cognitive Sandwich" view of intelligence (Hurley, 1998), according to which perception and action are considered as peripheral processes separated from each other and from cognition, which represents the central core of intelligence.

The criticisms raised to this general view during the last two decades, however, led to the development of a new framework according to which perception, action, and cognition are deeply intermingled processes that cannot be studied in isolation (Clark, 1997; Pfeifer and Scheier, 1999). According to this view, behaviour and cognition should be conceptualised as dynamical processes that arise from the continuous interactions occurring between the agent and the environment (van Gelder, 1998; Beer, 2000).

This new view of cognition led also to a new approach to categorisation. Categorisation represents one of the most fundamental cognitive capacities displayed by natural organisms, being an important prerequisite for the exhibition of several other cognitive skills (Harnad, 1987): for example, it is involved in any task that calls for differential responding, from operant discrimination to pattern recognition to naming and describing objects and states-of-affairs. The "Cognitive Sandwich" view of intelligence tends to look at categorisation by focusing on processes that are passive (i.e., the agents can not influence their sensory states through their actions) and instantaneous (i.e., the agents are demanded to categorise their *current* sensory state). The new paradigm to the study of cognition mentioned above demands to look at categorisation processes that are "active" and possibly distributed over time.

Active perception can be studied by exploiting the properties of autonomous embodied and situated agents, in which perception is strongly influenced by the agent action (on this issue, see also Gibson, 1977; Noë, 2004). Nevertheless, our ability to build artificial systems that are able to exploit sensory-motor coordination is still very limited. This can be explained by considering that, from the point of view of the designer of the robot, identifying the way in which the robot should interact with the environment in order to sense sensory states that might facilitate perception is extremely difficult. One promising approach, in this respect, is constituted by adaptive methods in which the robots are left free to determine how they interact with environment (i.e. how they behave in order to solve

their task). There are several works that successfully employed such methods for the control of embodied agents in categorisation tasks. For example the works described in (Nolfi, 2002) and in (Beer, 2003) demonstrate how categorisation can emerge from the dynamical interaction between the agent and the environment. Other works have shown how an active perception system can act in order to perceive discriminating stimuli that greatly simplify the discrimination task (see, for example Scheier et al., 1998; Nolfi and Marocco, 2002). In some cases, however, sensory-motor coordination is not sufficient to experience well differentiated sensory patterns for different categories. Thus, in these circumstances the agents are required to integrate "ambiguous" sensory-motor states over time. So far, only a few studies have shown evolved agents that are able to cope with this kind of problems (e.g. Gigliotta and Nolfi, 2008; Tuci et al., 2004).

This paper presents two experiments that aim to extend the current state of the art to more complex scenarios. The rationale behind the decision to investigate more complex scenario is twofold. On one side we wanted to verify whether the adaptive techniques used in previous related works scale to more challenging problems. On the other side we wanted to ascertain whether more complex problems would lead to solutions that are qualitatively similar to those observed in previous research or not. The first experiment consists of a simulated anthropomorphic robotic arm with coarse grained tactile sensors that is asked to discriminate between spherical and ellipsoid objects. The high number of Degrees of Freedom (DoFs), the necessity to master the effects of gravity, inertia, and collisions, and the high similarity between the two objects make this problem rather challenging. The second experiment consists in an active vision system that has to correctly recognise five different letters of different sizes. In this case the difficulty lies in the number of categories (almost all previous works use only two classes) and in the variability *within* elements of the same category. Despite the two setups are quite different, we show that the principles that underlie the behaviour of successful agents in the two cases are the same. In particular, successful agents are able to obtain close to optimal performance by (a) *actively selecting* sensory stimuli so to reduce perceptual ambiguities as much as possible, and (b) *integrating perceived sensory-motor states over time.*

2 Experiment 1

2.1 Methods

The first experimental setup consists of a simulated anthropomorphic robotic arm and hand with tactile sensors which is asked to discriminate between spherical and ellipsoid objects (see Fig. 1a and 1b). The experiment presented here is an extension of the work described in Tuci et al. (2009): please refer to that paper for additional information.

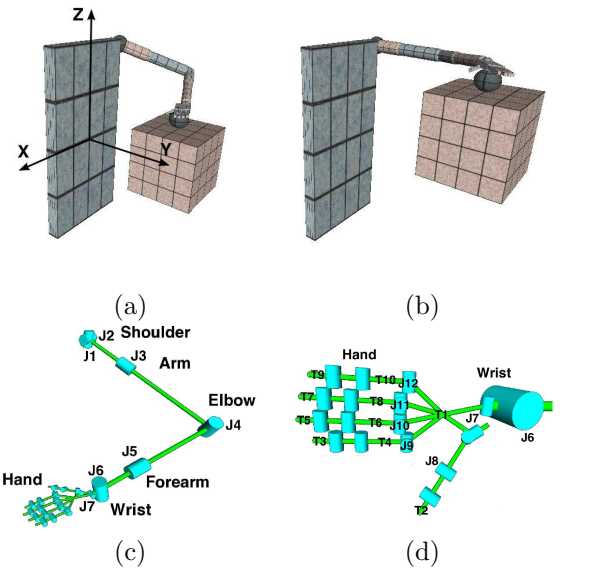


Figure 1: The simulated robotic arm (a) in position A, and (b) in position B. The kinematic chain (c) of the arm, and (d) of the hand. In (c) and (d), cylinders represent rotational DoFs; the axes of cylinders indicate the corresponding axis of rotation; the links among cylinders represents the rigid connections that make up the arm structure. T_i with $i = 1, ..., 10$ are the tactile sensors.

The robot and the robot/environment interactions are simulated using Newton Game Dynamics (NGD), a library for accurately simulating rigid body dynamics and collisions (`www.newtondynamics.com`). The arm has 7 actuated DoFs while the hand has 20 actuated DoFs. Fig. 1c shows the kinematic chain for the arm, the forearm and the wrist, with labels from J_1 to J_7 indicating rotational joints with the rotation axis along the axis of the corresponding cylinder. The robotic hand is composed of a palm and fourteen phalangeal segments that make up the digits (two for the thumb and three for each of the other four fingers) connected through 15 joints with 20 DoFs (see Fig. 1d). (See Massera et al., 2007, for a detailed description of the structural properties of the arm). Tactile sensors (indicated by the labels T_1 to T_{10} in Fig. 1d) return 1 if the corresponding part of the hand is in contact with any other body (e.g., the table, the sphere, the ellipsoid, or other parts of the arm), 0 otherwise.

The agent controller consists of a continuous time recurrent neural network (CTRNN, see Beer and Gallagher, 1992) with 22 sensory neurons, 8 internal neurons, 16 motor neurons, and 2 categorization neurons. The first 7 input neurons are updated on the basis of the state of the proprioceptive sensors on

joints J_1 to J_7 respectively (angles are linearly scaled on the range $[-1, 1]$), other 10 input neurons are updated accordingly to the state of tactile sensors T_1 to T_{10} respectively, and the remaining 5 input neurons are updated on the basis of the state of the hand proprioceptive sensors on joints J_8 to J_{12} respectively (angles are linearly scaled in the range $[0, 1]$, with 0 for a fully extended and 1 for a fully flexed finger). In order to take into account the fact that sensors are noisy, 5% uniform noise is added to proprioceptive sensors, while tactile sensors have a 5% probability of returning the wrong value. For all input neurons the activation value is computed by multiplying the corresponding sensory input by a gain factor g.

Internal neurons are fully connected to each other, and each receives one incoming synapse from each sensory neuron. Each motor and categorization neuron receives one incoming synapse from each internal neuron while there are no direct connections between sensory and motor neurons. The state of both internal, motor and categorization neurons is updated using the following equations:

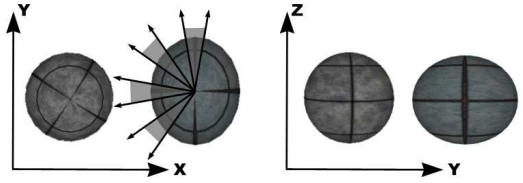
$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (1)$$

$$\tau_i \dot{y_i} = -y_i + \sum_{j \in N_i} \omega_{ji} \sigma(y_j + \beta_j) \quad (2)$$

where y_i is the state for neuron i, $\sigma(y_j + \beta_j)$ is the output of neuron j and N_i is the set of index of neurons with connection to neuron i. All time constants τ_i, biases β_i, network connection weights ω_{ij}, and all the input gains are genetically specified networks' parameters. There is one single bias for all the sensory neurons.

The activation values of motor neurons determine the state of the simulated muscles of the arm. Each joint in the arm is moved by an antagonist pair of muscles, so two neural outputs are associated with each joint (in total 14 neurons). For a complete description of the muscle model used in this work, see Massera et al. (2007). The joints of the hand are actuated by a limited number of independent variables through velocity-proportional controllers: the neural network has 2 output neurons for hand movements, one to set all desired thumb angles, the other to set the desired angles for all other fingers. The DoFs relative to joints J_9 to J_{12} are not actuated. Finally, the activation values of the two categorization neurons are used to categorize the shape of the object (see below).

A generational genetic algorithm is employed to set the parameters of the networks (see Goldberg, 1989; Nolfi and Floreano, 2000). The initial population contains 100 genotypes, represented as vectors of 420 parameters, each encoded with 16 bits. Generations following the first one are produced by a combination of selection with elitism and mutation:



(a) (b)

Figure 2: (a) The sphere and the ellipsoid of the first experiment viewed from above and (b) from west. The radius of the sphere is 2.5 cm. The radii of the ellipsoid are 2.5, 3.0 and 2.5 cm. In (a) the arrows indicate the intervals within which the initial rotation of the ellipsoid is set in different trials.

for each new generation, the 20 highest scoring individuals ("the elite") from the previous generation are retained unchanged, while the remainder of the new population is generated by making 4 mutated copies of each of the 20 highest scoring individuals with 1.5% mutation probability per bit.

During evolution, each genotype is translated into an arm controller and evaluated 8 times in position A and 8 times in position B (see Fig. 1); for each position, the arm experiences 4 times the ellipsoid and 4 times the sphere. Moreover, the rotation of the ellipsoid with respect to the z-axis is randomly set in different ranges for each trial (see Fig. 2a). At the beginning of each trial, the arm is located in the corresponding initial position (i.e., A or B), and the state of the neural controller is reset. It is then left free to interact with the object (e.g. by sliding the hand above it so to make it slightly roll) for 4 simulated seconds (400 time steps) but the trial is terminated earlier if the object falls off the table.

In each trial, an agent is rewarded by an evaluation function that seeks to assess its ability to recognise and distinguish the ellipsoid from the sphere. Rather than imposing a representation scheme in which different categories are associated with *a priori* determined states of the categorization neurons, we leave the robot free to determine how to communicate the result of its decision, while requiring that objects' categories are well represented in the categorization-output space. More precisely, at each time step, the output of the two categorization neurons is a point in the bi-dimensional Cartesian space $C = [0, 1] \times [0, 1]$. Given a set of such points, one can build the AABB (Axis-Aligned Bounding Box), which is the minimum rectangle containing all points in the set such that its edges are parallel to the coordinate axes. The idea is that of scoring agents on the basis of the extent to which the AABBs associated to different categories are non-overlapping. During each trial, we collect the categorization output produced by the agent during the last 20 steps. We consider the sphere category (referred to as C^S) as the minimum bounding box

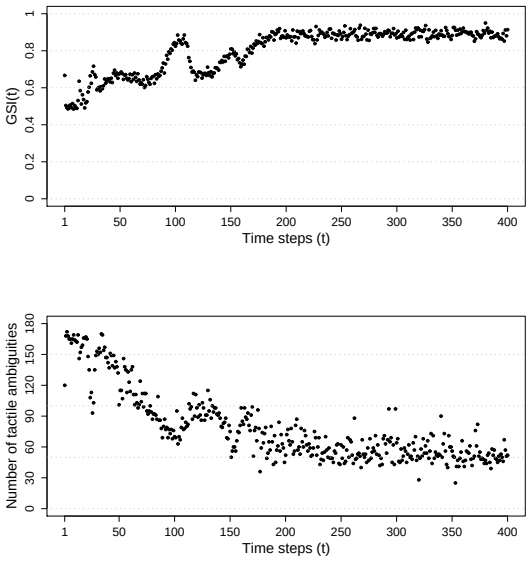


Figure 3: (a) The Geometric Separability Index (GSI). (b) Number of tactile ambiguities.

of all the categorization output collected while the agent was interacting with the sphere, and the ellipsoid category (referred to as C^E) as the minimum bounding box of all the categorization output collected while the agent was interacting with the ellipsoid.

The final fitness FF attributed to an agent is the sum of two fitness components: F_1 rewards the robots for touching the objects, and corresponds to the average distance over a set of 16 trials between the hand and the experienced object; F_2 rewards the robots for developing an unambiguous category representation scheme on the basis of the position in a two-dimensional space of C^S and C^E. F_1 and F_2 are computed as follows:

$$F_1 = \frac{1}{16}\sum_{k=1}^{16}\left(1-\frac{d_k}{d_{max}}\right) \quad (3)$$

$$F_2 = \begin{cases} 0 & \text{if } F_1 \neq 1 \\ 1-\frac{area(C^S \cap C^E)}{\min\{area(C^S),area(C^E)\}} & \text{if } F_1 = 1 \end{cases} \quad (4)$$

with d_k the euclidean distance between the object and the centre of the palm at the end of the trial k and d_{max} the maximum distance between the palm and the object when located on the table. $F_2 = 1$ if C^S and C^E do not overlap (i.e., if $C^S \cap C^E = \emptyset$).

2.2 Results

Eight evolutionary simulations, each using a different random initialisation, were run for 500 generations. Results of post-evaluation tests illustrated in (Tuci et al., 2009) shows that the best evolved agent (hereafter, A_1) possesses a close to optimal ability to discriminate the shape of the objects as well as an excellent ability to generalize their skill in new circumstances. Moreover, in (Tuci et al., 2009) it is shown that A_1, for one of the two positions experienced during evolution (i.e., position A, angle of joints $J_1, ..., J_7$ are $\{-50°, -20°, -20°, -100°, -30°, 0°, -10°\}$), exploits only tactile sensation to categorise the objects. In this Section, we take advantage of this latest result by running tests that further explore the dynamics of the decision of A_1 in position A, beyond the qualitative description illustrated in (Tuci et al., 2009). In particular, our interest is in finding out whether the discrimination process occur at a specific moment, as a response to a sensory pattern that encode the regularities which are necessary for discriminating, or if it occurs over time by integrating the information contained in several successive sensory states. Movies of the best evolved strategies can be found at `http://laral.istc.cnr.it/esm/active_perception`.

To answer this question we use a slightly modified version of the Geometric Separability Index (hereafter, referred to as *GSI*) originally proposed in (Thornton, 1997). GSI represents an estimate of the degree to which tactile sensor readings experienced during the interactions with the sphere or with the ellipsoid are separated in sensory space. We built four hundred data sets, one for each time step with the ellipsoid (i.e., $\{\tilde{I}_k^E(t)\}_{k=1}^{180}$), and four hundred data sets, one for each time step with the sphere (i.e., $\{\tilde{I}_k^S(t)\}_{k=1}^{180}$). Where, $\tilde{I}_k^E(t)$ is the tactile sensor readings experienced by A_1 while interacting with the ellipsoid at time step t of trial k; and $\tilde{I}_k^S(t)$ is the tactile sensor readings experienced by A_1 while interacting with the sphere at time step t of trial k. Trial after trial, the initial rotation of the ellipsoid around the z-axis changes of 1°, from 0° in the first trial to 179° in the last trial. Each trial is differently seeded to guaranteed random variations in the noise added to sensors readings. At each time step t, the *GSI* is computed as follows:

$$\begin{aligned} GSI(t) &= \frac{1}{180}\sum_{k=1}^{180} z_k(t) \\ z_k(t) &= \begin{cases} 1 & \text{if } m_k^{EE}(t) < m_k^{ES}(t) \\ 0 & \text{if } m_k^{EE}(t) > m_k^{ES}(t) \\ \frac{u_k(t)}{u_k(t)+v_k(t)} & \text{otherwise} \end{cases} \\ m_k^{EE}(t) &= \min_{\forall j \neq k}(H(\tilde{I}_k^E(t), \tilde{I}_j^E(t))) \\ m_k^{ES}(t) &= \min_{\forall j}(H(\tilde{I}_k^E(t), \tilde{I}_j^S(t))) \\ u_k(t) &= |\{\tilde{I}_j^E(t) : H(\tilde{I}_k^E(t), \tilde{I}_j^E(t)) = m_k^{EE}(t)\}_{\forall j \neq k}| \\ v_k(t) &= |\{\tilde{I}_j^S(t) : H(\tilde{I}_k^E(t), \tilde{I}_j^S(t)) = m_k^{ES}(t)\}_{\forall j}| \end{aligned} \quad (5)$$

where $H(x, y)$ is the Hamming distance between tactile sensor readings. $|x|$ means the cardinality of the set x. *GSI*=1 means that at time step t the closest neighbourhood of each $\tilde{I}^E_k(t)$ is one or more $\tilde{I}^E_k(t)$. *GSI*=0 means that at time step t the closest neighbourhood of each $\tilde{I}^E_k(t)$ is one or more $\tilde{I}^S_k(t)$.

As shown in Fig. 3a, the *GSI(t)* tends to increase from about 0.5 at time step 1 to about 0.9 at time step 200, and to remain around 0.9 until time step 400. This trend suggests that during the first 200 time steps, the agent acts in a way to bring forth those tactile sensor readings which facilitate the object identification and classification task. In other words, the behaviour exhibited by the agent allows it to experience two classes of sensory states, rather well separated in the sensory space, which correspond to objects belonging to two different categories. However, the fact that the *GSI* does not reach the value of 1.0 indicates that the two groups of sensory patterns belonging to the two objects are not fully separated in the sensory space. In other words, some of the sensory patterns experienced during the interactions with an ellipsoid are very similar or identical to sensory patterns experienced during interactions with the sphere and vice versa. This is confirmed by the graph shown in Fig. 3b, which refers to the number of tactile ambiguities at each time step.

A tactile ambiguity is defined as a condition in which at least some of the patterns are experienced during interactions with both an ellipsoid and a sphere. If there are tactile ambiguities, then the agent cannot determine the category of the object solely on the basis of the single sensory stimuli. The fact that the number of tactile ambiguities never reaches zero while the agent gets an almost optimal performance implies that the agent's categorization strategy involves an ability to integrate sequences of experienced sensory states over time.

3 Experiment 2

3.1 Methods

The second experimental scenario involves a simulated agent provided with a moving eye located in front of a screen that is used to display images to be categorized (one at a time). The eye includes a fovea constituted by 5×5 photoreceptors distributed uniformly over a square area located at the centre of the eye's 'retina', and a periphery constituted by 5×5 photoreceptors distributed uniformly over a square area that covers the entire retina of the eye. Each photoreceptor detects the average grey level of an area corresponding to 1×1 pixel or to 10×10 pixels of the image displayed on the screen, for foveal and peripheral photoreceptors, respectively (see Fig. 4b). The activation of each photoreceptor ranges between

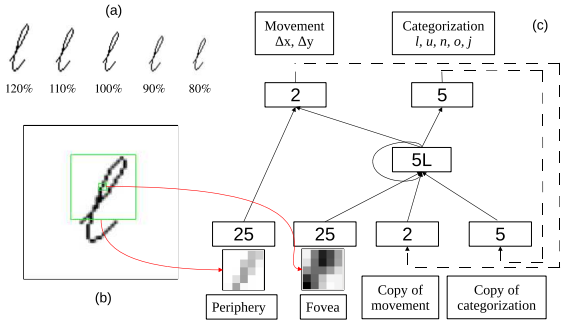


Figure 4: (a) Letter 'l' shown in the 5 different sizes used in the experiment. (b) The screen displaying the letter 'l' in its intermediate size and an exemplification of the field of view of the foveal and peripheral vision (smaller and larger squares, respectively). (c) The architecture of the neural controller. The number inside the each rectangle indicates the number of neurons, the letter L in a box indicates that these neurons are leaky integrators. Solid arrows between two boxes indicate all-to-all connections between neurons of those boxes, while dashed arrows indicate that the activation of the output units at time t is copied in the respective input units at time $t + 1$.

0 and 1 and is given by the average gray level of the pixels spanned by its receptive field (where 0 and 1 represent a fully white and a fully black visual field, respectively). The eye can explore the image by moving along the up-down and left-right axes up to a maximum distance corresponding to 25 pixels of the image. The screen, located in front of the agent's eye, is used to display five types of italic letters ('l', 'u', 'n', 'o', 'j'), each of which can be of 5 different sizes (with a variation of $\pm 10\%$ and $\pm 20\%$ with respect to the intermediate size: see Fig. 4a, for the letter 'l'). The letters are displayed in black/gray over a white background. As shown in Fig. 4b, the eye can perceive only a tiny part of a letter with its foveal vision and a much larger but still incomplete part of the letter with its peripheral vision. It is important to clarify that this set-up is not intended to model how humans actually recognize letters; rather, the characteristics of the set-up have been chosen so to allow us to study how an active vision system can categorize stimuli through the exploitation of its eye movements and, possibly, to the integration of the perceived information over time.

Agents are provided with a neural network controller with 57 sensory neurons, 5 internal neurons, and 7 output neurons: see Fig. 4c for the network architecture. Notice that sensory neurons relative to the eye periphery are connected only to the two movement output neurons. This connection pattern represents a very crude abstraction of the functional organization of the human visual system, in which eye movements seem to be driven primarily by the periphery while recognition seems to be based pri-

marily on the information provided by fovea (Findlay and Gilchrist, 2003; Wong, 2008). To take into account the fact that sensors are noisy, a random value with a uniform distribution in the range $[-0.05; 0.05]$ is added to the activation state of each photoreceptor of the fovea in each time step.

The output of each of the 5 leaky internal neurons depends on the input received from the sensory and internal neurons through the weighted connections and by its own activation at the previous time step, and is calculated as follow:

$$O_i^t = \tau_i O_i^{t-1} + (1 - \tau_i)\sigma(\sum_{j \in N_i} O_j^{t-1} w_{ji} + b_i) \quad (6)$$

where O_i^t is the output of unit i at time t, τ_i is the time constant of unit i, in $[0; 1]$, w_{ji} is the weight of the connection from unit j to unit i, and b_i is the unit's bias, and $\sigma(x)$ is calculated as in equation 1. The output of the output units is calculated as in equation 6 but the time constant is fixed to 0 (i.e. output neurons do not depend on their previous state). The output of the motor units is then linearly normalized in the range $[-25; 25]$ and used to vary the position of the eye along the x and y axes of the image, respectively.

Free network parameters are learned using a genetic algorithm similar to the one described for the previous experiment. Agents are evaluated for 50 trials lasting 100 time steps each. At the beginning of each trial the screen is set so to display one of the five different letters in one of the five different sizes (each letter of each size is presented twice to each individual), the state of the internal neurons of the agent's neural controller is initialized to 0, and the eye is initialized in a random position within the central third of the screen (so that the agent can always perceive some part of the letter, at least with its peripheral vision). During the 100 time steps of each trial the agent is left free to visually explore the screen. Trials, however, are terminated earlier if the agent does not perceive any part of the letter through its peripheral vision for three consecutive time steps. The task of the agent consists in labelling the category of the current letter correctly during the second half of the trial. More specifically, the agents are evaluated on the basis of the following fitness function FF which comprises two components: the first one measures the agents' ability to activate the categorization unit corresponding to the current category more than the other units; the second one measures the ability to maximize the activation of the right unit while minimizing those of the other units:

$$F_1(t, c) = 2^{-rank(t,c)} \quad (7)$$

$$F_2(t, c) = \frac{1}{2} O_r^{t,c} + \sum_{O \in O_w^{t,c}} \frac{1}{8}(1 - O) \quad (8)$$

$$FF = \frac{\sum_{t=1}^{50} \sum_{c=50}^{100} \left(\frac{1}{2} F_1(t, c) + \frac{1}{2} F_2(t, c) \right)}{50 \cdot 50} \quad (9)$$

where $F_1(t, c)$ and $F_2(t, c)$ are the values of the two fitness components at step c of trial t, $rank(t, c)$ is the ranking of the activation of the categorization unit corresponding to the correct letter (from 0, meaning the most activated, to 4, meaning the least activated), $O_r^{t,c}$ is the activation of the output corresponding to the right letter at step c of trial t and $O_w^{t,c}$ is the set of activations corresponding to the wrong letters at step c of trial t. Notice that, as in the previous setup, individuals are not rewarded for moving their eyes or for producing a certain type of exploration behaviour but only for the ability to categorize (in this case the type of letter).

3.2 Results

Twenty evolutionary simulations were run, each lasting 3000 generations. The best agents of all simulations obtained on the average a good performance, with the best agent of the best replication reaching close to optimal performance. In order to better quantify the ability of the adapted agents to categorize the letters, we measured the percentage of times in which, during the second half of each trial, the categorization unit corresponding to the current letter is the most activated. We evaluated the best individuals of each of the 20 replications of the experiment for 10000 trials during which they are exposed to all possible combinations of the 5 letters with 50 sizes (uniformly distributed over the range $[-20\%, +20\%]$ of the intermediate size), 40 times each for each combination. As a result, we obtained that the average performance over all replications is 76.92% and the performance of the best individual of the best replication is 94.32%. In the remaining part of this section, we will focus our analysis on the best evolved agent, that is the best individual of replication 12.

By analysing the behaviour displayed by the best individual we can see how, after an initial phase lasting typically from 5 to 30 time steps (in which the behaviour varies significantly for different initial positions of the eye and for different letter sizes), the behaviour of the agent converges either on a fixed point attractor (i.e. the eye stops moving after having reached a particular position of the letter) or on a limit cycle attractor (i.e. the eye keeps moving by periodically foveating sequentially 2-6 different specific areas of the image). Interestingly, the agent displays the same type of behaviour in interaction with letters belonging to the same category even if they are of different sizes, and different behaviours for letters of different categories.

As for the previous experimental setup, we wanted to quantitatively ascertain the capacity of evolved

individuals to actively select discriminating stimuli. Apart from the efferent copies that provide as input the categorization output produced by the agent in the previous time step, the categorization answer of our system depends on two sources of information: the visual information provided by photoreceptors of the fovea and the motor information provided by the efferent copies of the motor neurons controlling the eye movements. Starting from the GSI index introduced in the previous experiment, we adapted it to the new setup and then we observed the evolution of the values of this index for both kinds of input (visual and motor) during the interaction of the agent with the images.

More precisely, in this case, the index takes into account all the stimuli experienced in interaction with an object of a given category. Hence, we devised what we call the Modified Geometric Separability Index (*MGSI*), which is defined as the average, over all patterns, of the proportion of the patterns belonging to the same category that are in the $|C_x|$ nearest patterns (using the euclidean distance), with $|C_x|$ representing the total number of patterns in the same category as pattern x. More formally, the *MGSI* is calculated as follows:

$$MGSI(P) = \frac{\sum_{x \in P} \frac{\sum_{n \in N_X} \mathbb{1}_{C_x}(n)}{|C_x|}}{|P|} \qquad (10)$$

where $|S|$ indicates the cardinality of the set S, P is the set comprising all the patterns, C_x is the set of all patterns belonging to the same category as pattern x (x doesn't belong to C_x), N_x is the set of the $|C_x|$ patterns nearest to pattern x and $\mathbb{1}_{C_x}(n)$ is the indicator function of set C_x: it returns 1 if n is in the set C_x, 0 otherwise.

We calculated the *MGSI* of both the visual and motor-copy patterns experienced by the best evolved agent during 250 test trials, ten replications (with different initial positions) for each of the 5 by 5 letter-dimension pairings. More specifically, the two *MGSI*s were calculated for each of the 100 cycles composing trials, so that we could observe their evolution during the agent's interactions with the images. The results are shown in Fig. 5. They show three things. First, the separability of the input patterns in both sensory channels (visual and motor) significantly increase throughout trials, in particular during the first 20 cycles, meaning that the agent's sensory-motor behaviour has evolved so to facilitate the categorization process. Second, the geometric separability of the inputs in the two channels reaches very similar values (with the motor-copy channel being slightly better). Third, the geometric separability of neither of the two channels reaches very high values, meaning that, as in the previous experiment,

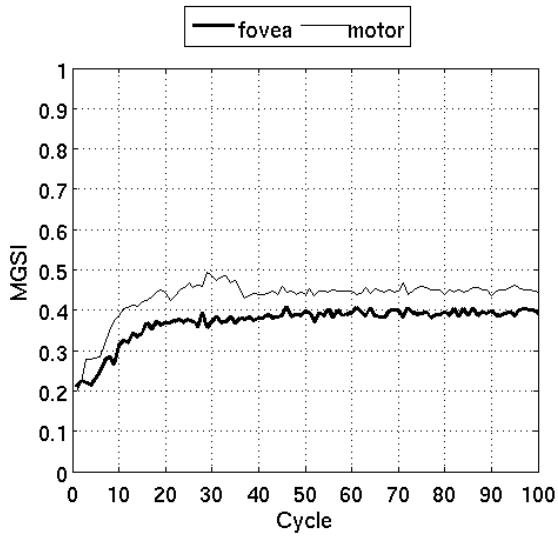


Figure 5: Evolution of the *MGSI* of the fovea and efferent copy of the eye movements inputs during the 100 cycles of the trials. Each point along the x axis represents the value of the *MGSI* calculated by taking all the inputs recorded in 250 trials (5 letters × 5 dimensions × 10 repetitions) during one of the 100 cycles of each trial.

to successfully solve the task the system has to integrate the information collected during different time steps, because each sensory pattern collected in a singular time step does not provide enough information for correct discrimination.

4 Conclusions

In this paper we presented two different experimental setups in which embodied agents are asked to categorize various objects by actively selecting their inputs. In the first scenario an anthropomorphic robotic arm equipped with coarse grained tactile sensors has been asked to distinguish between spherical and ellipsoidal objects. The setup is significantly more complex than those used in previous related works due to the high similarity between the objects to be discriminated, the difficulty of controlling a system with so many degrees of freedom, and the need to master the effects produced by gravity, inertia, collisions, etc. Nevertheless the evolved system is able to solve the task and reach close to optimal performance.

The second scenario involves an agent with a simulated moving eye that have to recognize different letters. Whereas work in related literature has mainly focused on experiments comprising only two categories, this setup is more challenging as there are significantly more categories with more variability (five letters of different dimensions). Also in this case the system is able to successfully solve the task with a close to optimal performance.

Both experiments show that active perception systems are indeed able to cope with complex scenarios. The ability to actively select one's own input is exploited by agents by selecting stimuli that provide regularities that can be used to categorize (i.e. stimuli that are often, although not necessarily always, experienced in interaction with objects of the corresponding category). Despite the effectiveness of their actions, however, agents often encounter input patterns associated with more than one category. Thus, evolved agents also show a complementary ability to integrate over time the partially conflicting information provided by the experienced stimuli.

Acknowledgements

This research work was supported by the *ITALK* project (EU, ICT, Cognitive Systems and Robotics Integrating Project, grant n° 214668).

References

Beer, R. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, 4:91–99.

Beer, R. and Gallagher, J. (1992). Evolving dynamic neural networks for adaptive behavior. *Adaptive Behavior*, 1(1):91–122.

Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, 11(4):209–243.

Clark, A. (1997). *Being There: putting brain, body and world together again*. Oxford University Press, Oxford.

Findlay, J. M. and Gilchrist, I. D. (2003). *Active Vision. The Psychology of Looking and Seeing*. Oxford University Press, Oxford.

Gibson, J. J. (1977). The theory of affordances. In Shaw, R. and Bransford, J., (Eds.), *Perceiving, Acting and Knowing. Toward an Ecological Psychology*, chapter 3, pages 67–82. Lawrencw Erlbaum Associates, Hillsdale, NJ.

Gigliotta, O. and Nolfi, S. (2008). On the coupling between agent internal and agent/environmental dynamics: Development of spatial representations in evolving autonomous robots. *Adaptive Behavior*, 16:148–165.

Goldberg, D. (1989). *Genetic algorithms in search, optimization and machine learning*. Reading, MA: Addison-Wesley.

Harnad, S., (Ed.) (1987). *Categorical Perception: The Groundwork of Cognition*. Cambridge University Press.

Hurley, S. (1998). *Consciousness in Action*. Harvard University Press, Cambridge, MA.

Massera, G., Cangelosi, A., and Nolfi, S. (2007). Evolution of prehension ability in an anthropomorphic neurorobotic arm. *Front. Neurorobot.*, 1.

Noë, A. (2004). *Action in Perception*. MIT Press, Cambridge, MA.

Nolfi, S. (2002). Power and limits of reactive agents. *Neurocomputing*, 49:119–145.

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

Nolfi, S. and Marocco, D. (2002). Active perception: A sensorimotor account of object categorisation. In Hallam, B., Floreano, D., Hallam, J., Hayes, G., and Meyer, J.-A., (Eds.), *Proc. of the* 7^{th} *Inernational Conference on Simulation of Adaptive Behavior (SAB '02)*, pages 266–271. MIT Press, Cambridge, MA.

Pfeifer, R. and Scheier, C. (1999). *Understanding intelligence*. MIT Press, Cambridge, MA.

Scheier, C., Pfeifer, R., and Kunyioshi, Y. (1998). Embedded neural networks: exploiting constraints. *Neural Networks*, 11(7-8):1551–1596.

Thornton, C. (1997). Separability is a learner's best friend. In Bullinaria, J., Glasspool, D., and Houghton, G., (Eds.), *Proc. of the* 4^{th} *Neural Computation and Psychology Workshop: Connectionist Representations*, pages 40–47. Springer Verlag, London, UK.

Tuci, E., Massera, G., and Nolfi, S. (2009). Active categorical perception in an evolved anthropomorphic robotic arm. In *Proc. of the IEEE Conference on Evolutionay Computation (CEC '09), Special Session on Evolutionary Robotics*, ISBN: 978-1-4244-2959-2. Draft available at `http://laral.istc.cnr.it/elio.tuci/pagn/pubb.html`.

Tuci, E., Trianni, V., and Dorigo, M. (2004). Feeling the flow of time through sensory-motor coordination. *Connection Science*, 16(4):301–324.

van Gelder, T. J. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, 21:615–665.

Wong, A. M. (2008). *Eye Movement Disorders*. Oxford University Press, Oxford.