Mirolli Marco, Tomassino Ferrauto, Stefano Nolfi (2010): Categorisation through Evidence Accumulation in an Active Vision System, Connection Science, 22(4): 331-354.
doi: 10.1080/09540091.2010.505976

Categorisation through Evidence Accumulation in an Active Vision System

Marco Mirolli, Tomassino Ferrauto, Stefano Nolfi

Istituto di Scienze e Tecnologie della Cognizione, CNR,
Via San Martino della Battaglia 44, I-00185 Roma, Italy
{marco.mirolli,tomassino.ferrauto,stefano.nolfi}@istc.cnr.it

Authors' information

Marco Mirolli (corresponding author)

Affiliation: Istituto di Scienze e Tecnologie della Cognizione, CNR

Postal address: Via San Martino della Battaglia 44, I-00185 Roma, Italy

Telephone: +39 06 44595231

Fax: +39 06 4459 5243

e-mail: marco.mirolli@istc.cnr.it

Tomassino Ferrauto

Affiliation: Istituto di Scienze e Tecnologie della Cognizione, CNR

Postal address: Via San Martino della Battaglia 44, I-00185 Roma, Italy

Telephone: +39 06 44595255

Fax: +39 06 4459 5243

e-mail: tomassino.ferrauto@istc.cnr.it

Stefano Nolfi

Affiliation: Istituto di Scienze e Tecnologie della Cognizione, CNR

Postal address: Via San Martino della Battaglia 44, I-00185 Roma, Italy

Telephone: +39 06 44595233

Fax: +39 06 4459 5243

e-mail: stefano.nolfi@istc.cnr.it

Categorisation through Evidence Accumulation in an Active Vision System

Marco Mirolli, Tomassino Ferrauto, Stefano Nolfi

Istituto di Scienze e Tecnologie della Cognizione, CNR,
Via San Martino della Battaglia 44, I-00185 Roma, Italy
{marco.mirolli,tomassino.ferrauto,stefano.nolfi}@istc.cnr.it

Abstract. In this paper we present an artificial vision system that is trained with a genetic algorithm for categorising five different kinds of images (letters) of different sizes. The system, which has a limited field of view, can move its eye so as to visually explore the images. The analysis of the system at the end of the training process indicates that correct categorisation is achieved by (1) exploiting sensory-motor coordination so as to experience stimuli that facilitate discrimination, and (2) integrating perceptual and/or motor information over time through a process of accumulation of partially conflicting evidence. We discuss our results with respect to the possible different strategies for categorisation and to the possible roles that action can play in perception.

Keywords: Active vision; categorisation; neural networks

Vision is a palpation with the look
Merleau-Ponty (1973)

1 Introduction

1.1 Active categorical perception

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 (the two slices of bread) are considered as peripheral processes separated from each other and from the cognitive processes (the meat), which represent the central core of intelligence. The assumption that perceptual, cognitive, and motor processes are fundamentally independent from each other implies that they can (have to) be studied separately, and, in particular, that cognition can be identified with the reasoning/planning processes and has little to do with perception and action. It is within this theoretical framework, for example, that Marr's theory of vision (Marr, 1982), which can be considered as one of the most systematic, influential, and paradigmatic approaches to vision and perception in general, has been elaborated.

The severe criticisms raised to this general view during the last two decades, however, has led to the development of an alternative framework according to which perception, action, and cognition are deeply intermingled processes, which 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; Nolfi, in press).

With respect to the relation between perceptual, cognitive, and motor processes, vision represents a paradigmatic case since: 1) it is the most important perceptual modality in humans; 2) it is the one that has been most extensively studied; 3) it is intuitively conceptualised as a passive process. To use the words of Alva Noë: "When we try to understand the nature of sensory perception, we tend to think in terms of vision, and when we think of vision, we

tend to suppose that the eye is like a camera and that vision is a quasi-photographic process. To see, we suppose, is to undergo snapshot-like experiences of the scene before us." (Noë, 2004, p. 35).

Theoretical and experimental evidence collected by studying vision in both natural and artificial systems (Yarbus, 1967; Ballard, 1991; Churchland et al., 1994; O'Regan and Noë, 2001; Findlay and Gilchrist, 2003; Noë, 2004) demonstrate instead that vision is an active process in which the actions performed by the agent (e.g. the eye movements) play a fundamental/constitutive role. From the empirical point of view, a clear demonstration of the active nature of vision comes from the seminal work of Yarbus (1967), who registered the eye movements of subjects that were asked to look at a picture while executing different tasks. Yarbus showed two fundamental things: 1) humans move their eyes continuously even when they look at static pictures; 2) movements are task-specific, that is functional to the (cognitive) task at hand. Thanks to a considerable amount of psychological and neuroscientific research, today we have gained a lot of knowledge both on the neural circuits that control eye movements (Wong, 2008) and on the role of these movements in orienting attention, visual search, and other activities such as reading (see Findlay and Gilchrist, 2003). Despite all this knowledge, however, we still do not have a clear understanding of the ways in which eye movements are shaped in order to enable or facilitate visual perceptual processes in general, and categorisation in particular.

Similar considerations can be done with respect to the study of active vision through a synthetic methodology (i.e. through the development of artificial systems displaying skills similar to those possessed by natural organisms). Indeed, although some recent studies have demonstrated how artificial systems can exploit the possibility to actively explore the scene to solve specific tasks in simple and robust ways (Aloimonos et al., 1988; Bajcsy, 1988; Ballard, 1991; Harvey et al., 1994; Floreano et al., 2004, 2005; de Croon et al., 2006; Suzuki and Floreano, 2008), we still know very little on how eye movements should be shaped in relation to the current situation/task and on how the visual information sensed during scene exploration can be integrated over time to serve a given function.

In this paper we will investigate how an agent can exploit the possibility of co-determinining its own sensory states through its actions and the possibility of integrating the experienced sensory-motor states over time to solve a categorisation problem. Before describing the experimental scenario and the obtained results, we first illustrate in the next section the relationship between the work presented in this paper and the most relevant literature on active perception in artificial systems.

1.2 Related previous work

Categorisation is one of the most fundamental capacities displayed by natural organisms and represents a prerequisite for the exhibition of several other cognitive skills (Harnad, 1987). Not surprisingly, categorisation has been extensively studied (Cohen and Lefebvre, 2005) both in the natural sciences (e.g. Psychology, Philosophy, Ethology, Linguistics, and Neuroscience) and in the artificial sciences (e.g. Artificial Intelligence, Neural Network research, and Robotics).

Recent work performed by evolving artificial embodied agents to produce different behaviours in different environmental contexts has provided interesting insights on how the coordination between sensory and motor processes can be exploited for categorisation. In particular, Beer (1996); Nolfi (1997, 2002b); Beer (2003) have demonstrated how agents that have to produce different behaviours in different contexts might succeed in doing so without internally discriminating the current context. For example, in the experiments performed by Nolfi (1997, 2002b) a wheeled robot provided with a simple neural controller with 6 sensory neurons (that encode the state of 6 corresponding infrared sensors) directly connected to two motor neurons (encoding the desired speed of the two corresponding robot's wheels) has been evolved (Nolfi and Floreano, 2000) to find and remain close to cylindrical objects located inside an arena surrounded by walls. The evolved robots display an ability to differentiate their behaviour in different contexts by avoiding walls and approaching and remaining close to cylinders. The analysis of the sensory states experienced by the robots situated in their environment and the lack of any internal states clearly demonstrate that the robots have no internal 'knowledge' of whether they are currently located in front of a wall or of a cylindrical object. The behaviours of remaining close to cylinders and of avoiding walls, in fact, are the result of the dynamical interactions between the agent and the environment and cannot be explained by considering the characteristics of the agent alone. The problem is solved by regulating how the agent reacts to different sensory states (i.e. by exploiting sensory-motor coordination) so that the dynamical system constituted by the agent and the environment

converges on a limit cycle dynamics (consisting in oscillating back and fourth and left and right while remaining in the same relative position) close to cylinders and not close to walls.

Other authors (Harvey et al., 1994; Scheier et al., 1998; Kato and Floreano, 2001; Nolfi and Marocco, 2001; Nolfi, 2002b; Nolfi and Marocco, 2002; Floreano et al., 2004) have demonstrated how the possibility to influence sensed stimuli through actions can be used to find (or even build) discriminative stimuli (i.e. stimuli which can be unambiguously associated to the current context). In the experiments described in Kato and Floreano (2001), for example, an agent placed in front of a white board containing a black isosceles triangle or a black square has been evolved for the ability to visually categorise the objects' shape. The analysis of adapted individuals indicates that they solve the problem by exploring the image so as to find a portion of the image which can be associated unambiguously with the shape of the object. More specifically, the evolved strategy consists in finding the black object and then moving toward one of its vertical edges. Reaching one of the edges, in fact, ensures that the agent senses two different types of patterns (corresponding to a diagonal or a straight edge) which can be associated to the corresponding category (triangle or square, respectively).

Still other recent research work has demonstrated how the strategies summarised above (i.e. categorising on the basis of behavioural attractors or of self-selected discriminative stimuli) can be applied successfully also to tasks in which the contexts to be categorised are apparently identical from the point of view of the agent's perceptual system. An illustrative case is constituted by the work reported in Nolfi and Marocco (2001), in which a wheeled robot provided only with short-range infrared sensors is asked to find and remain in the north-west or south-east corners of the rectangular arena in which it is situated (thus discriminating these locations from the north-east and south-west corners). At the end of the adaptive process the robot displays an ability to solve the task despite the fact that the four corners are perceptually identical and the sizes of the walls and the proportion between the north/south and east/west sides vary randomly in each trial (but the north and south walls are always longer than the east and west walls). The analysis of the evolved strategies indicates that the robot solves the problem by reaching one of the four corners, leaving the corner at an angle of about 45^o, and then turning left and following the wall up to the next corner, when they encounter a wall on their left side. Leaving a corner at such an angle, in fact, ensures that the robot will only encounter the north or the south walls on its left side. Being located in front of these two walls, in turn, implies that the correct corners (i.e. the north-west and south-east corners) are necessarily located at the end of the left side of the wall. The apparent paradox constituted by the ability to discriminate between perceptually identical locations can be explained by considering that embodied and situated agents always perceive a subset of all possible sensory states, which depends on their own behaviour. By selecting the appropriate behaviours, a situated agent can manage to select a sub-set of sensory states that is not perceptually ambiguous. More generally, all these experiments point to the fact that the stimuli sensed by an embodied agent depend on the environment, on the agent's sensory system, and on the agent's behaviour (and not only on the first two components as we often tend to assume implicitly). As stressed by Nolfi (2005), this means that in embodied and situated agents sensory stimuli are always action-mediated (i.e. are always influenced by the agent's actions).

The simple control policies described above are not always sufficient for producing optimal (categorisation) behaviour. In some cases it has been demonstrated that the agent might need to complement its sensory-motor activities with additional processes that operate by integrating sensory-motor information over time in internal states, for example, by extracting and using simple information regarding the time spent by the robot in performing a given behaviour. An example of the use of such temporal information is constituted by an extension of the experiment just described (Nolfi and Marocco, 2001). The sensory-motor strategy described in the previous paragraph, in fact, allows the robot to avoid the wrong corners but is not conductive to the robot remaining in one of the two correct corners since the stimuli experienced near corners do not provide any cue about whether the present corner belongs to the good or bad category. As a consequence, robots provided with simple reactive controllers that cannot hold any information about previously experienced sensory states solve the problem sub-optimally by repeatedly moving back and forth between the two correct corners. Only robots provided with internal neurons and recurrent connections are able to stop in one of the two correct corners. Interestingly, this optimal solution is based on the combination of the same sensory-motor strategy described above with a simple timing mechanism that keeps track of time elapsed from the start of the trial. The evolved robots use this information to remain in the current corner after the robot interacted with the environment for a sufficient amount of time (Nolfi and Marocco, 2001). After a given time, in fact, experienced corners necessarily belong to the correct category. For other examples of how the time duration of

a given sensory state can be used to discriminate functionally different contexts see, Nolfi (2002a) and Suzuki and Floreano (2006).

All the above-mentioned studies involved the discrimination between only two different categories of stimuli and resulted in very interesting but also very simple strategies. In this paper we investigate a richer scenario in which an agent has to discriminate between *five* different categories and in which simple strategies based on the selection of discriminative stimuli do not suffice because of: (a) the considerable number of categories, (b) the possibility to sense only a limited part of the object to be categorised, (c) the variations occurring between the items of the same category, and (d) the fact that sensors provide noisy information. The obtained results demonstrate that the agents succeed in developing an ability to solve the categorisation problem also in this case. Furthermore, the analysis of adapted individuals indicates that the problem is solved by selecting, through sensory-motor coordination, stimuli that provide cues for categorising which, nevertheless, represent partially conflicting evidence. The problem due to the presence of partially ambiguous evidence is solved primarily through a simple process of evidences accumulation.

The rest of the paper is structured as follows: in section 2 we describe our experimental set-up; in section 3 we report the obtained results and the analyses of the evolved solutions; finally, in section 4 we discuss the implications of our work for the understaning of categorisation and active perception.

2 Experimental set-up

To investigate whether an artificial agent provided with a moving 'eye' and with foveal and peripheral photoreceptors can categorise objects with different shapes we devised the simple experimental scenario described in 2.1. The agent is provided with a neural controller that regulates how the eye moves and how the experienced sensory stimuli are used to discriminate the category of the object (2.2). Given the strong interdependences between the eye's motor control and visual perception, in order to investigate how such an agent can exploit its eye movements to enable categorisation, we trained the robot controller through an evolutionary method in which the fine-grained parameters that regulate the agent-environment interactions and the agent's categorisation responses are encoded in free parameters that are varied randomly, and in which variations are retained or discarded on the basis of their effect on the overall ability of the agent to perform the categorisation task (2.3).

2.1 The agent and the environment

The experimental scenario involves a simulated agent provided with a moving eye located in front of a screen (of 100 x 100 pixels) that is used to display the objects to be categorised (one at a time). The eye includes a fovea, constituted by 5 x 5 photoreceptors distributed uniformly over a square area located at the centre of the eye's 'retina', and a periphery, constituted by 5 x 5 photoreceptors distributed uniformly over a square area that covers the entire retina of the eye (for a similar approaches, see Schmidhuber and Huber, 1991; Kato and Floreano, 2001). Each photoreceptor detects the average grey level of an area corresponding to 1 x 1 pixel or to 10 x 10 pixels of the image displayed on the screen, for foveal and peripheral photoreceptors, respectively (see Figure 1b). The activation of each photoreceptor ranges between 0 and 1, with 0 representing a fully white and 1 representing a fully black visual field. The eye can explore the image by moving along the up-down and left-right axes (the maximal displacement along each axis in each time step is 12 pixels). The screen is used to display five types of italic letters ('l', 'u', 'n', 'o', 'j') of five different sizes (with a variation of $\pm 10\%$ and $\pm 20\%$ with respect to the intermediate size: see Figure 1a, for the letter 'l'). The letters are displayed in black/gray over a white background. As shown in Figure 1b, the eye can perceive only a tiny part of a letter with its foveal vision and a much higher 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. The small resolution and size of the visual field and the low number and variability of the visual stimuli can not permit this. Rather, the characteristics of the set-up have been chosen so as to allow us to study how an active vision system can categorise stimuli through the exploitation of its eye movements and, possibly, to the integration of the perceived information over time. An analysis of the characteristics of the agent's sensory and control system that represent the minimal pre-requisites for solving the chosen task is reported in section 3.7.

[Fig. 1 about here.]

2.2 The neural network controller

Agent controllers are constituted by neural networks with the architecture shown in Figure 1c. The controller includes 57 input neurons that encode the current state of the 25 foveal and 25 peripheral photoreceptors and the efference copies of the two motor neurons and of the 5 categorisation units (i.e. the state of these units at time t-1). 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 in order to take into account the fact that the grey level measured by the photoreceptor is subject to noise. The two motor neurons determine the eye movements, that is the variation of the eye position over the x and y axes, respectively, within a range corresponding to $[-12; 12]$ pixels of the image. The five categorisation units allow the agent to label the categories of the five corresponding letters. [1] The input neurons are simple relay units which are set to the current value of the corresponding sensor (in the case of peripheral and fovea sensors) or to the previous activation state of the corresponding neuron (in the case of the efference copies of movement and categorisation units). The output of the 5 leaky internal neurons depends on the input received from the input neurons through the weighted connections and on their own activation at the previous time step, and is calculated as follows:

$$O_i^t = \tau_i O_i^{t-1} + (1 - \tau_i) A_i^t \tag{1}$$

where O_i^t is the output of unit i at time t, A_i^t is the activation of unit i at time t as given by the standard logistic function (eq. 2), and τ_i is the time constant of unit i, in $[0; 1]$.

The output of the motor and categorisation unit is calculated through a logistic function:

$$O_i = \frac{1}{1 + e^{-(\sum_j O_j w_{ji} + b_i)}} \tag{2}$$

where O_i is the output of unit i, w_{ji} is the weight of the connection from unit j to unit i, and b_i is the unit's bias. The output of the two motor neurons is then linearly normalised in the range $[-12; 12]$ and used to vary the position of the eye along the x and y axes of the image, respectively.

The fact that we included direct connections between the peripheral photoreceptors and the two motor neurons while we did not include connections between these receptors and the internal neurons on which categorisation depends, represents a very crude abstraction of the functional organisation of the human visual system, in which eye movements seem to be driven primarily by the periphery while visual recognition seems to be based primarily on the information provided by the fovea (Findlay and Gilchrist, 2003; Wong, 2008).

2.3 The task and the adaptive process

Agents are evaluated over 50 trials lasting 100 time steps each. At the beginning of each trial: (a) one of the five different letters in one of the five different sizes is displayed at the center of the screen (each letter of each size is presented twice to each individual), (b) the state of the internal neurons of the agent's neural controller is initialised to 0.0, and (c) the eye is initialised in a random position within the central third of the screen (so that the agent can always perceive 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 as soon as the agent loses visual contact with the letter (i.e. if it does not perceive any part of the letter through its peripheral vision for three consecutive time steps). The task of the agent consists of labelling the category of the current letter correctly during the second half of the trial (i.e. after the agent has had enough time to visually explore the image). More specifically, the agents are evaluated on the basis of the following equation (fitness function):

[1] As in other studies on active perception in evolutionary artificial systems (e.g. Kato and Floreano, 2001; Nolfi, 2002a), for the sake of simplicity, the task of our system is to explicitly categorize its stimuli through a dedicated set of output units. For an alternative way of encoding categories that is independent from their number, and which may be used in future work, see Tuci et al. (in press). Still another, even more interesting, possibility to be explored in future work would be to remove categorisation units and ask the system to perform *behavioural* categorization (i.e. different sensorymotor behaviours for different categories) as, for example, in Nolfi and Marocco (2001); Beer (2003); Mirolli (submitted).

$$F = \frac{\sum_{t=1}^{nT} \sum_{c=sFC}^{nC} \left(0.5 \cdot 2^{-rank} + 0.5 \cdot \left(O_r^t \cdot 0.5 + \sum_{O \in O_w^t} (1 - O) \cdot \frac{0.5}{nL - 1} \right) \right)}{nT \cdot (nC - sFC)} \quad (3)$$

where nT is the number of trials (i.e. 50), nC is the number of steps in a trial (i.e. 100), sFC is the step in which we start to calculate fitness (i.e. 50), $rank$ is the ranking of the activation of the categorisation unit corresponding to the correct letter (from 0, meaning the most activated, to 4, meaning the least activated), O_r^t is the activation of the output corresponding to the right letter in trial t, O_w^t is the set of activations corresponding to the wrong letters for trial t, and nL is the number of letters (i.e. 5). The first component of the fitness function (2^{-rank}) rewards the agent's ability to activate the categorisation unit corresponding to the current category more than the other units; the second component ($O_r^t \cdot 0.5 + \sum_{O \in O_w^t} (1 - O) \cdot \frac{0.5}{nL - 1}$) rewards the ability to maximise the activation of the correct unit while minimising those of the other units, with the maximization of the correct unit weighting as much as the sum of the minimization of all other units. This two-components fitness function produced better results than either of the two single components alone (results not shown). Notice that individuals are not rewarded for moving their eyes nor for producing a certain type of exploration behaviour but only for the ability to categorise the current letter.

The free parameters of the agents' neural controller are adapted through an evolutionary method (Nolfi and Floreano, 2000; Floreano et al., 2008). The initial population consists of 100 randomly generated genotypes, each encoding the free parameters of a corresponding neural controller, which include all the connection weights, the biases, and the time constants of leaky neurons. Each parameter is encoded with 8 bits. In order to generate the phenotypes, weights and biases are linearly mapped in the range $[-5; 5]$, while time constants are mapped in $[0; 1]$. Each individual of the current generation is evaluated for 50 trials as described above on the basis of the eq. 3. At the end of such evaluation, the best 20 individuals reproduce by producing five offspring each. One out of the five offspring consists of an exact copy of the reproducing individual. The other four offspring consist of copies with the addition of random mutations (i.e. each bit has a 2% probability of being replaced by a randomly selected value). The evaluation, selection, and reproduction process is repeated for 3000 generations. The whole evolutionary experiment has been replicated 20 times starting with different randomly generated initial genotypes.

3 Results

The best agents of all replications display on average good performance, with the best agent of the best replication reaching close to optimal performance. Furthermore, the evolved agents are able to categorise correctly letters with sizes that differ from those experienced during the adaptive process (section 3.1). The analysis of the behaviours exhibited by the best individuals indicates that they explore the image through the use of eye movements and that they succeed in correctly categorising letters after having explored a tiny part of the image with their fovea. Moreover, this analysis indicates that agents tend to produce different behaviours for letters belonging to different categories and similar behaviours for letters belonging to the same category, independently of the letter size (section 3.2). The analysis of the perceived stimuli indicates that the active exploration of the letters allows the agent to experience regularities in two input channels (i.e. in the fovea and in the motor efference copies), which might be almost equally used for letter discrimination (section 3.3). The analysis of the actual role played in the categorisation behaviour of the best evolved agent by the two channels indicates that, in the case of the basic experiment, the visual channel plays the most significant role (section 3.4). The analysis of the process through which the agent discriminates letters indicates that categorsiation is based on a simple process of evidence accumulation (section 3.5). The analysis of a new set of experiments demonstrates that the relative importance of the two input channels (visual and motor) for the categorisation process depends on the way in which the two kinds of information are encoded (section 3.6). Finally, the analysis of a series of control experiments indicates that the possibility of integrating information over time constitutes a necessary prerequisite for solving the chosen task, while a dual visual system (i.e. a system in which there is a distinction between a fovea and a periphery) is not strictly necessary providing that the field of view is sufficiently wide (section 3.7).

3.1 Performance

In order to analyse the ability of the adapted agents to categorise the letters, we measured the percentage of times in which, during the second half of each trial, the categorisation unit corresponding to the current letter is the most activated. Moreover, in order to ensure accuracy and to verify the ability of the agents to generalise their skill to different sizes of letters, we evaluated the best individuals of each replication of the experiment for 10000 trials during which it is exposed 40 times to all possible combinations of the 5 letters with 50 sizes (uniformly distributed over the range $[-20\%, +20\%]$ with respect to the intermediate size). 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 next subsections we will focus our analyses primarily on the best evolved agent, that is the best individual of replication 12.

3.2 Behavioural analysis

By analysing the behaviour displayed by the best individual we can see how the eye quickly moves toward the bottom-left part of the letter and then soon converges either on a fixed point attractor, in which the eye stops moving and keeps foveating a specific point of the letter, or on a limit cycle attractor, in which the eye keeps moving while foveating 2-6 specific parts of the image in sequence. Interestingly, the agent displays the same behaviours in interaction with letters belonging to the same category (independently of letter size and starting position), and different behaviours for letters of different categories (see Figure 2, and the movies available at http://laral.istc.cnr.it/mirolli/activevision.html).

[Fig. 2 about here.]

In particular, the behaviour of the best agent converges on a fixed point attractor in the case of letters 'n' and 'o', and on a limit cycle in the case of the letters 'l', 'u', and 'j'. More specifically, in the case of letter 'n' (Figure 2c), the agent's eye stops while foveating the white background located in the lower-left side of the letter. In the case of letter 'o' (Figure 2d), it stops while foveating a small sector of the arc that forms the lower-left part of the letter. In the case of the letter 'l' (Figure 2a), the eye keeps jumping back and forth between an area in the bottom-left part of the letter and a blank area located in the further bottom-left area. When interacting with a 'j' (Figure 2e), the agent keeps jumping back and forth between an area at the bottom of the letter and a blank area located nearby on the top-left with respect to the former area. Finally, in the case of the letter 'u' (Figure 2b), the agent produces an almost-circular, periodic trajectory by making counter-clockwise circular jumps during which it tends to foveate the blank background twice and a specific point of the left side of the letter once.

These patterns of behaviour imply that in the case of letter 'o' the agent converges on a limit state behaviour in which it experiences only fully discriminative stimuli (i.e. stimuli that are experienced only in that categorical context), a strategy already observed in some of the experiments reviewed above (e.g. Scheier et al., 1998; Nolfi and Marocco, 2002; Floreano et al., 2004). However, in all other cases, the system converges on limit state behaviours in which the experienced stimuli provide partially conflicting evidences (i.e. the same stimuli can be experienced in different categorical contexts). For example, the stimulus experienced when the agent foveates a fully white area is experienced in interaction with all letters but the 'o' (even though with different probabilities).[2] In order to ascertain whether the fact that the behavioural attractors are all displaced in the bottom left area of the images was contingent to the specific sensory-motor interactions produced with the letters to be categorised or whether it is something that depends on the general strategy found by our system, we tested the behaviour of the system with other images, never experienced during the evolutionary process. The behaviours of the system with the new images are analogous to those exhibited with the images used during the training process: in particular, after a very short initial phase the system falls in a limit state (fixed point or limit cycle) behaviour during which the eye keeps

[2] Obviously, the details of the strategy that we are discussing hold only for the best individual that we are analysing. By looking at the behaviour displayed by the best individuals of the other replications of the experiment, however, we observed that all evolved individuals converge either on fixed point attractors or on limit cycles, depending on the category of the letter. On the other hand, the specific areas visited by the agent and the type of limit state behaviour produced in interaction with each type of letter vary in different replications. Similarly, not all other replications produced agents that exhibit a general preference for a specific area of the image, as it happens for the agent we analyse in detail here.

looking to an area located on the bottom-left part of the image (see, figure 3). Hence, the causes that bring the system towards behavioural attractors located in the bottom-left part of the image are general, and do not depend on the details of the images with which the system interacts.

[Fig. 3 about here.]

To understand how the agent manages to move its eye toward the south-west portion of the visual stimulus by converging on different types of limit state behaviours for stimuli belonging to different categories, we first tried to disentangle the roles played in eye movements by the recurrent internal neurons and by the peripheral vision photoreceptors, which are directly connected to the motor neurons. By freezing the state of the internal neurons to a null value, we observed that the eye moves toward the south-west portion of the visual stimuli independently of its original position, and keeps moving toward the same direction indefinitely, thus abandoning the visual stimulus. However, by freezing the state of the internal neurons to a vector of values corresponding to the average state of the same neurons recorded during 10000 trials performed in a normal condition, we observed that the agent produces behaviours that are qualitatively similar to that produced in normal condition (i.e. it moves towards the south-west portion of the image and then converges on the same limit-state behaviours observed in normal conditions). These tests indicate that the internal neurons only play the role of biasing the activation state of the motor neurons: in other words, the eye-movements produced by the agent are not (much) influenced by the way in which the state of the internal neurons vary over time. The characteristics of the behaviours produced by the agents, therefore, fundamentally depend on: (a) the way in which the agent reacts to the visual patterns experienced on the peripheral vision (which in turn depends on the weights of the connections that project directly from the photoreceptors of the periphery to the motor neurons), and (b) the effects that eye-movements have on the next experienced sensory states. By measuring the direction and length of the movements that would be produced in case only a single photoreceptor is activated, we observed that the agent tends to move up or down when the upper or lower parts of the visual field, respectively, are activated, and left or right when the left or right parts of the visual field, respectively, are activated (figure 4a). The fact that the agent always ends up in the south-west part of the image is explained by the fact that, averaging over all the photoreceptors, the components of the movements towards west and towards south are stronger than those pointing towards east and towards north, respectively. This is demonstrated by the fact that the agent reacts to homogeneous peripheral stimuli, in which all photoreceptors are equally activated, by moving towards south-west (figure 4b). Moreover, this explanation is also supported by the analysis of the movements produced by the agent in the first step of each of the 10000 test trials (40 replications with 5 letters by 50 dimensions), i.e. when the eye is randomly positioned on the images. In fact, in the majority of the cases (74.55%) the agent produces a movement toward south-west, and only in 23.22%, 2.22%, and 0.0001% of the cases, the eye moves toward south-east, north-west, and north-east, respectively.

[Fig. 4 about here.]

The fact that the eye-movements converge towards limit state behaviours and towards different kinds of limit state behaviours for letters of different categories can be explained by analyzing the way in which the agent reacts to different sensory patterns together with the sensory effects that the eye-movements produce when the agent is situated in an environment containing a letter of a certain category. For exemple, the limit state behaviours in which the agent oscillates between two specific points of a letter (independently of the letter size), can be explained by the fact that the reaction triggered by the visual perception of a specific portion of the visual stimulus triggers a movement that produces a sensory effect which in turn triggers a movement that brings the agent back to the original position. This is the case, for example, for letter 'j' (figure 5): when the agent foveates a point in the lower-left part of the letter (figure 2e) it produces a jump towards north-west (figure 5a, black arrows) that will make it foveate a blank point on the left of the letter (figure 2e); this position produces peripheric patterns that trigger movements towards south-east (figure 5a, gray arrows), thus bringing the agent back to its previous location. Similarly, fixed point behavioural attractors can be explained by considering that the visual stimuli sensed by the agent once it looks at the bottom-left part of the letter trigger either very short-range movements that make the agent experience new stimuli that in turn trigger opposite short-range movements or a stay-still action which, by leading to the same visual input, causes the indefinite reiteration of the same stay-still action (see, for example, the case of letter 'n' (figure 5b).

[Fig. 5 about here.]

3.3 On the role of behaviour in facilitating categorisation

To verify whether the agent's behaviours just described facilitate the categorisation process we analyzed the separation of the stimuli belonging to different categories in the two input spaces (of the fovea and of the motor copies) while the agent interacts with the images. In order to do that, we used a modified version of the Geometric Separability Index (GSI) proposed by Thornton (1997), which computes the percentage of times in which the nearest stimulus of each experienced stimulus belongs to the same category. More specifically, we used a more demanding measure, which we call the Modified Geometric Separability Index (MGSI), which takes into account not only the nearest neighbour, but all the stimuli belonging to same category. For each category k, the MGSI is defined as the average, over the $|C^k|$ patterns belonging to k, of the proportion of the $|C^k|-1$ stimuli belonging to the same category that are in the $|C^k|-1$ nearest stimuli, where $|C^k|$ is the cardinality of the set of patterns belonging to k. More formally, the MGSI is calculated as follows:

$$MGSI(C^k) = \frac{\sum_{x \in C^k} \frac{\sum_{n \in N_x} I_{C^k}(n)}{|C^k|-1}}{|C^k|} \tag{4}$$

where N_x is the set of the $|C^k|-1$ patterns nearest to pattern x and $I_{C^k}(n)$ is the indicator function of set C^k, which returns 1 if n is in C^k, 0 otherwise.

We calculated the MGSI on the stimuli experienced both through the foveal photoreceptors and through the efference copy of the motors during 250 trials in which the agent experiences the 5 letters of the 5 different sizes for 10 times each. For each type of input information and for each category, the MGSI has been calculated for each of the 100 steps composing a trial, so that we could observe the evolution of the MGSIs during the agent's interactions with the images.

Figure 6a and b show the evolutions of the MGSIs for the visual and motor stimuli, respectively. The first thing to be noted is that, in general, the separability of the stimuli increases during the first part of the trials for all the letters and in both the visual and the motor-copy channels (only the visual stimuli related to the 'u' maintain a separability close to chance level). This clearly confirms that the behaviour exhibited by the agent allows it to get to perceive more discriminative stimuli, thus facilitating the categorisation process. On the other hand, only the MGSI of the visual stimuli of the 'o' reaches a high value (around 0.8), with all other MGSIs remaining at intermediate values (in between 0.2 and 0.6). The fact that the visual patterns perceived with the 'o' are easily separable from all the others is explained by the fact that, as we have seen in the previous section, in presence of an 'o' the agent ends up in a fixed point attractor in which it foveates a specific part of the letter, which in turn makes the agent experience a visual pattern that is specific to that letter. On the other hand, the fact that all other visual patterns are poorly separable is explained by the fact that the limit state behaviours exhibited with the other letters all lead to experiencing similar patterns for a considerable amount of time: in particular, the stimulus that corresponds to the foveation of a white portion of the image is experienced once every other cycle in case of 'l' and 'j', around two thirds of the time with 'u', and al the time with 'n'.

The MGSIs of the motor-copy stimuli are more homogeneous, and, on average, a little bit higher then the ones of the foveal stimuli (the average MGSI in the last cycle over all letters are 0.4 and 0.45 for visual and motor stimuli, respectively). The reason is the following. In the motor space, 'o' and 'n' lead to the same fixed point attractor, corresponding to roughly the same 'stay still' (0.5; 0.5) pattern. This explains the fact that the MGSIs of these two letters fluctuate around 0.5 (each motor pattern belonging to one of the two has roughly half of the nearest patterns belonging to the same letter and half belonging to the other). Even the limit cycles of period 2 (i.e. for 'l' and 'j') tend to produce a MGSI near 0.5, since the motor patterns produced in such cases are quite well clustered in two different sets, each corresponding to one of the opposite movements that generate the limit cycle (south-west to north-east and viceversa for the 'l' and north-west to south-east and viceversa for the 'j'). Finally, the 'u' leads to a less clean, almost circular, behavioural attractor which is produced by motor patterns that are less clustered and more widespread in the motor space. This explains the lower value of the MGSI of the 'u'.

Notwithstanding the differences in the MGSIs both between input channels and between letters within a channel, it is clear that: (1) the behavioural strategy of the system is able to increase the separability of the input patterns that facilitates categorisation; (2) this active perception process does not lead to the perception of fully discriminative stimuli, which implies that the close to optimal categorisation ability exhibited by the system must also involve some form of integration of perceptual information through time which resolves the partially conflicting evidence provided by the experienced stimuli. What is not clear is: (1) whether the information that is used for categorsation is the visual one, the motor one, or both; and (2) which are the characteristics of the process that allows the agent to resolve the ambiguities provided by the experienced sensory states. The next two sub-sections are devoted to clarifying these two points.

[Fig. 6 about here.]

3.4 On the role of different input channels

The analysis reported in the previous section demonstrates that both input channels provide cues for discriminating the current category. To verify whether the best adapted agent exploits both types of regularities or whether it uses one of the two sources of information only, we performed a series of tests in which the state of the foveal photoreceptors are set to the registered sequence experienced in interaction with a specific letter of a specific size and the state of the motor-copy inputs are set to the registered sequence experienced in interaction with another letter of a specific size. In particular, each of the 50 registered sequences of foveal stimuli of each letter (5 dimensions for 10 repetitions) has been tested in combination with each of the 200 registered sequences of motor-copy stimuli of the different letters (4 letters x 5 dimensions x 10 repetitions), for a total of 10000 trials for each letter. During these tests the motor outputs produced by the neural controller are ignored.

By analysing the percentages of times in which the most activated categorisation output corresponds to the stimuli experienced through the foveal photoreceptors, to the stimuli experienced through the efference copy of the motors, or to neither of the two (figure 7, white, grey, and black histograms, respectively), we can see that the categorisation behaviour depends almost exclusively on the visual input. Indeed, the average probabilities that the agent's categorisation output corresponds to the category of the data provided by the foveal photoreceptors, by the motor-copy inputs, or to another category are 0.8874, 0.0255, and 0.0871, respectively. Furthermore, the analysis of performance for each letter (figure 7) indicates that visual information plays the predominant role for all letters.

The analysis of the other replications of the experiment indicates that the relative importance of the two input channels varies overall and for different categories (data not shown). In all replications, however, the visual channel plays the most significant role. We will come back on this issue in section 3.6, where we will see how the relative importance of the two input channels depends on the way in which input information is encoded. Before doing that, however, we will focus our attention on how the information provided by the foveal photoreceptors is integrated over time so as to solve the problems caused by the fact that experienced stimuli provide partially conflicting evidence.

[Fig. 7 about here.]

3.5 On the dynamics of the categorisation process

The results described above demonstrate that the categorisation process is based primarily on the stimuli experienced through the foveal photoreceptors as a result of the execution of specific behaviours. Moreover, the analysis reported in the previous section indicates that the categorisation process should also involve an ability to integrate the experienced sensory-motor information over time since the stimuli belonging to different categories are not well separated in the input space.

In order to assess the role played by the order with which stimuli are experienced, we analysed the average categorisation pattern produced by the agent in a test condition in which, during each trial, the foveal photoreceptors have been set to the value corresponding to each visual stimulus experienced in normal conditions and frozen for the entire duration of the trial. More precisely, for each type of letter, the state of the motor-copy input is set to $[0.5; 0.5]$ (representing a stay-still action), and the state of the foveal photoreceptors is set to one of the 2500 sensory stimuli experienced in normal conditions during the last 50 time steps of 50 trials (performed with 5 different sizes

for 10 times). Each test is terminated as soon as the pattern of activations of the categorisation units converges on a fixed value (i.e. it changes less than 0.01 for 5 contiguous time steps).[3] For each of these tests, the final activations of the 5 categorisation units can be conceived as representing the probabilities, for the network, that the presented pattern is associated with each of the corresponding categories. For each of the five letters, we record the 2500 output vectors resulting from these tests (one for each foveal sensory stimulus) and then calculate the average pattern over them.

The fact that the correct categorisation output unit is the most activated also in the average categorisation patterns produced during these tests (Figure 8a) for all letters but the 'l', together with the fact that the average patterns produced during these tests are similar to the average patterns produced in normal conditions (Figure 8b), seems to suggest that the order in which stimuli are experienced plays only a minor role: in other words, the categorisation process seems to be determined primarily on the basis of the type and frequency of the stimuli that are experienced, and not so much on their sequence.

[Fig. 8 about here.]

Furthermore, we compared categorisation performance (i.e. the percentage of time in which the most activated categorisation unit corresponds to the perceived letter) under three testing conditions: (1) the normal condition (i.e. when the agent is allowed to autonomously interact with the images), (2) the condition just described (i.e. when the motor-copy is kept fixed at $[0.5; 0.5]$, the visual input is kept fixed at one of the patterns belonging to a letter, and the categorisation output has converged on a fixed point), and (3) a condition in which the foveal photoreceptors are fed with visual patterns that are randomly chosen within the ones belonging to the same category. More precisely, for this third condition we ran 2000 trials for each letter, each of which lasted 100 cycles. During each trial we kept the motor copy fixed to $< 0.5, 0.5 >$, and, in each cycle, we fed the network with one visual input that is randomly chosen within the 2500 visual patterns recorded for that given letter in the previous test (as in condition 2, we used only the patterns recorded during the last 50 cycles of the 50 trials: 5 dimensions by 10 repetitions with random starting point).

Figure 9 shows the result of this comparison. The first thing that can be noted is that in the case of 'n' and 'o' the performance is almost optimal in all the three conditions. This is not surprising, since these are the two letters for which the agent reaches fixed point behavioural attractors, which produce specific visual stimuli that are constant and discriminative by themselves. In the case of the other three letters, the performance when categorisation depends on single visual patterns (white bars) is very low, around chance level (which is 0.2). As a result, the average performance in this condition drops down to 0.51, which clearly confirms that, generally speaking, single sensory patterns do not provide enough information for the agent to correctly discriminate the letters. Even the randomisation of the order of the visual stimuli (grey bars) leads to a decrease in performance in the case of letters 'l', 'u', and 'j', but this decrease is much smaller than in the 'single patterns' condition. In fact, in this condition the agent performs much better than chance with all letters, thus reaching an average performance of 0.84, which is not much lower than in the normal condition (i.e. 0.94, black bars). Hence, this result demonstrates that though the order with which stimuli are experienced does plays some role, in particular in the cases of 'l' and 'u', the most substantial part of the categorisation performance depends on the distribution of the stimuli perceived with each letter, with their sequential order playing only a minor role.

To summarize, the categorisation problem is solved through a process of accumulation of the partially conflicting evidence provided by the experienced stimuli (which in turn are selected by the agent itself through sensory-motor coordination). This process of evidence accumulation roughly consists in using each experienced stimulus to increase the activation states of the categorisation outputs proportionally to the probability with which that stimulus is experienced in interaction which the corresponding categorical context. During this process, some of the experienced stimuli produce a relative increase in the wrong categorisation units. However, the overall summed contribution of the evidence provided by all experienced stimuli ensures that the correct categorisation unit reaches the highest activation level. For related accumulator models used to make decisions by integrating over time partial evidence provided by noisy stimuli, see Usher and McClelland (2001).

[3] The reason we use only the stimuli recorded during the second half of the trial is that it is during this period that performance is calculated in normal conditions. This assures that the agent has already reached one of its behavioural attractors, and thus that the stimuli that it receives are the ones that it self-selects through its active perceptual strategy.

[Fig. 9 about here.]

3.6 On the role of visual versus motor information

In the previous sections we have shown how the behaviour exhibited by the adapted agents ensures that both the visual and motor patterns experienced by the agents provide the regularities that can be used to categorise the correct context. This notwithstanding, our analyses showed that adapted individuals tend to rely primarily on the information provided by the visual channel. In this section we report the results of a series of additional experiments that aimed to ascertain the reasons that determined the supremacy of visual over motor information observed in this particular experimental setting.

More specifically, we investigated whether the supremacy of visual information over motor information can be explained simply by the fact that the information provided by the former input channel tends to have a quantitatively stronger impact than the latter (since visual information is encoded over 25 neurons, while motor information over only 2 neurons). In order to do that, we ran a new set of experiments in which the number of input neurons was kept fixed but the relative impact of the efference copy of the motor neurons was magnified by 10 times by normalizing their activation state within $[0; 10]$ instead than within $[0; 1]$.

The comparison between the performance obtained in the first and second set of experiments indicates that the new encoding of motor information leads to slightly better results. Indeed, the best individual over all replications of the first and second set of experiments reached a performance of 0.94 and 0.98, respectively, and the average performance of the best individuals of all replications are, respectively, 0.7692 and 0.8463. Finally, the number of replications that achieved a performance higher than 0.9 (our criterion for success) are 2 out of 20 and 8 out of 20 in the first and second set of experiments, respectively.

In order to assess the relative role of the two input channels in the new set of experiments, we subjected the best evolved individuals to the same analysis described in section 3.4. The obtained results demonstrate that the increase of the relative impact of the efference copy of the motor neurons over the visual neurons (obtained by re-scaling their activation range) leads to solutions in which motor information tends to play the main role. In fact, contrary to the situation observed in the first experiment, in the second experiment the categorisation behaviour of the best individual depends primarily on the motor information and only secondarily on the visual information (Figure 10a). The inversion of the relative importance of the two channels is also confirmed by the average results obtained by comparing the overall performance of the best individuals of all replications of the two sets of experiments (Figure 10b).

Overall, these results indicate that motor information can indeed be exploited for categorisation, and that this did not happen in the first set of experiments only because of the limited impact that motor information could have on internal neurons in that experimental set-up.

[Fig. 10 about here.]

3.7 Analysis of Minimal Requirements

To identify which are the characteristics of the agents neural architecture and sensory system which represent prerequisites for solving the task, we ran a further series of control experiments. The first set of experiments that we conducted aimed at identifying whether the presence of both a fovea and a peripheral vision (that differ in terms of resolution and visual field) was strictly necessary to solve the task. In particular, we ran two control experiments in which the agents were provided with only one set of 5 x 5 visual sensors that are connected both directly to the two motor neurons (as for our peripheral vision) and to the five internal neurons (as for our foveal vision). In one experiment the size and resolution of the visual input are the ones that we used for the fovea, in the other experiment they are the ones that we used for the periphery. Moreover, we ran two other control experiments in which the agents were provided only with peripheral vision but in which, as in the basic architecture, the photoreceptors were linked with the motor neurons only and not with internal neurons. Hence, in these control experiments the visual input directly determines only eye movements, while categorisation is based only on the efference copy of the motor neurons. These last two experiments differ with respect to whether the state of the efference copy of the motor neurons is normalized in $[0, 1]$ or in $[0, 10]$, respectively. Results are reported in Table 1 under the Fovea-Only,

Periphery-Only, No-Fovea, No-Fovea-Gain-10 columns. The obtained results indicate that the peripheral visual channel is sufficient to produce close to optimal performance providing either that the visual photoreceptors are linked to the internal neurons (Periphery-Only) or that the range of the motor copy neurons is sufficiently high (No-Fovea-Gain-10). This observation confirms the fact that our categorisation task can be successfully accomplished either by integrating the sensory information supplied by the visual photoreceptors or by integrating the information supplied by the efference copy of the motors (provided that the range of variation of these neurons is sufficiently high so that small differences are amplified). The fact that architectures with only peripheral vision can produce close to optimal performance, while those with only foveal vision cannot (Fovea-Only), indicates that in order to solve the task the visual field must be sufficiently large, in particular for permitting appropriate exploration of the image. It should be noted, however, that the need of an additional fovea vision (with a higher resolution with respect to the peripheral vision) might turn out to be necessary in cases in which the stimuli to be discriminated are smaller than in our set-up, or in cases in which the range of variation of the stimuli, with respect to their size, is larger.

A second set of experiments aimed to identify whether the availability of neural mechanisms that allow the agents to integrate sensory-motor information over time represented a pre-requisite for the ability to solve the task. In order to ascertain this point we ran an experiment in which the internal neurons are updated on the basis of a standard logistic function and are not provided with recurrent connections (No-Time-Processing), and an experiment in which the agents were not provided with internal neurons at all (No-Internal-Neurons). The fact that none of these two experiments produced successful agents and that the average and best performances are significantly lower than in the basic set-up indicates that the possibility to integrate sensory-motor information over time represents a necessary prerequisite for solving the task.

[Table 1 about here.]

4 Discussion

4.1 Categorising by integrating information through time

In this paper we presented an active vision system that is trained using an evolutionary method to categorise five types of italic letters of five different sizes. During the training process the system is rewarded only for the ability to discriminate the shape of the letters while it is left free to determine how it should explore the visual scene. The analysis of the best adapted individuals indicated that the strategy used for solving the categorisation task was based on two complementary abilities: (1) the ability to coordinate sensory-motor activity so as to fall in behavioural attractors that are specific for each category and that allow the agents to experience partially different sets of stimuli in different categorical contexts; and (2) the ability to discriminate the current categorical context on the basis of a process of accumulation of partially conflicing evidence for which what matters is the distribution of the perceived stimuli but not (much) their sequential order.

The fact that the system uses this accumulation of evidence strategy can be explained by the fact that this happens to be the simpler strategy that is sufficient for solving the given task. In fact, on the one hand, it seems that finding a way of coordinating the sensory-motor process so as to experience fully discriminative stimuli is not possible in this case because of the complexity of the task. In particular, because of the considerable number of categories, the possibility to sense only a limited part of the image to be categorised, the differences between items of the same category, and the fact that sensors are noisy. On the other hand, in the present task it seems that taking into account the order in which stimuli are experienced is not necessary (at least in most of the cases). Hence, the evolved strategy, based on the accumulation of partially conflicting perceptual evidence, represents an important extension of the purely sensory-motor strategies reviewed in section 1.2 (Harvey et al., 1994; Beer, 1996, 2003; Nolfi, 1997, 2002b,a; Scheier et al., 1998; Kato and Floreano, 2001; Nolfi and Marocco, 2001, 2002; Floreano et al., 2004; Suzuki and Floreano, 2006). This new kind of strategy can be successfully applied in cases in which, for the complexity of the task, it is not possible to behave so as to experience fully discriminative stimuli only (i.e. stimuli that are experienced only in interaction with a single category).

An important topic for future research will be to study under which conditions categorisation might require relying more consistently on the order in which stimuli are experienced. Some evidence in this direction might also

be gathered by further analyses to be conducted on the experiments in which categorisation depends more on the motor efference copy then on the visual input: indeed, preliminary analyses on the experiments described in section 3.6 seem to indicate that in that case the order in which stimuli are experienced plays a more important role. In this respect, it would be interesting to study whether different input channels might provide information which has qualitatively different characteristics. For example, motor information might tend to be more reliable and less noisy than visual information, while visual information might tend to be richer than motor information and might thus allow the exploitation of simpler strategies. Another interesting topic for future research is the investigation of how different sources of information (e.g. motor and visual) can be fused in order to produce better performance with respect to those that can be obtained on the basis of either source of information in isolation. Finally, another important issue for future research consists of the study of the conditions that might lead to still more complex categorisation forms involving the exploitation of sensory-motor contingencies: that is, the anticipation of the sensory consequences of the agent's own actions (Noton and Stark, 1971; O'Regan and Noë, 2001; Noë, 2004).

4.2 The roles of action in perception

The active perception framework stresses the importance of action for perception. But there are at least two different senses in which this importance has been intended so far. The first sense is related to general evolutionary considerations. Organisms must survive and reproduce. For surviving and reproducing what really matters is what one does. All kinds of cognitive processes, including perceptual ones, have evolved for subserving organism behaviour, because it is behaviour that determines whether an animal will reproduce or not. Hence, all cognitive processes should be understood in terms of the kind of behavioural capacities they allow. This kind of action-based view of cognition in general (and perception in particular) has been proposed, among others, by Gibson (1979); Bickhard (2001); Di Ferdinando and Parisi (2004); Gallese and Lakoff (2005). Probably the best known and most influential theoretical proposal in this line is represented by Gibson's notion of *affordance*. Gibson's idea is that what an organism perceives are not the objective properties of the environment, but rather the opportunities for action that the environment affords for that organism. This idea has been gaining increasing empirical support from both psychology (cf. the experiments on the priming of motor responses by visual objects: e.g. Tucker and Ellis, 1998; Borghi, 2005) and neuroscience (cf. the discovery of canonical neurons: e.g. Rizzolatti et al., 1988, 2002).

A second sense in which action can be considered as important for perception is the one most directly related to the active perception framework, and hence to the present work. As discussed in the introduction, the basic idea here is that perception is not a passive process in which an agent analyses the sensory stimuli determined by its environment, but rather an active process, which is constitutively dependent on the sensory-motor interactions between the agent and its environment. The central idea here is that thanks to its own behaviour an agent can co-determine the stimuli that it receives, and that the possibility to influence one's own experienced stimuli is a fundamental, constitutive aspect of most if not all perceptual processes.

We do think that action is important for perception for both these reasons. But the experiment described in section 3.6, in which we changed the encoding of the motor input, suggests still another way in which actions can influence perception: namely, that one's own movements can be used as the input to be categorised. Our experiments demonstrate that if a copy of the movements that the agent produces is suitably encoded as an input for the categorisation system, then the categorisation process can be based not only on the information gathered by the external environment, but also on the information regarding the agent's own behaviour. The reason is that an agent's interactions with different types of object will tend to result not only in different sensory stimuli but also in different movements. And it might turn out to be easier or more effective to classify one's own patterns of movements rather than the stimuli that they determine during the interaction with the environment.

Neuroscientific research assumes the presence in the brain of copies of motor commands (in particular of eye movements commands), and recent empirical evidence has started to reveal the neural bases of this copy, known as *efference copy* or *corollary discharge* (Guthrie et al., 1983; Merriam and Colby, 2005; Sommer and Wurtz, 2006). But the standard view about the functional role that such motor copy plays in vision is that it allows predictions of the sensory consequences of these movements, thus permitting the maintenance of visual stability despite the continuous movements of the eyes (Burr, 2004; Sommer and Wurtz, 2008). Our experiments suggest another possible function that the motor copy of eye movements might play: that of constituting additional inputs for the perceptual interpretation of visual stimuli.

The idea that the categorisation of observed images might be based not only on (sequences of) visual stimuli but also on the movements that the eyes make during visual perception had been proposed in the early '70s by Noton and Stark in their scanpath theory (Noton and Stark, 1971), but has not subsequently received much attention. Recently Hafed and Krauzlis (2006) have re-vitalised this idea by showing through behavioural studies that eye movements can significantly improve performance in visual tasks. In particular, they showed that the coherence of ambiguous and partially occluded visual stimuli is increased when the eyes have to move for visually pursuing the stimulus with respect to a condition in which the eyes perceive the stimulus under passive fixation. And, most importantly, this facilitatory effect of eye movements on perception is found in an experimental set-up in which the retinal stimulations received by subjects under different eye movements conditions are the same, which seems to be a very strong evidence for the hypothesis that the information about ongoing eye movements does in fact play a role in the interpretation of visual stimuli.

Acknowledgements

This work was supported by the European Commission FP7 Project ITALK (ICT-214668) within the Cognitive Systems, Interaction, and Robotics unit. We thank Dario Floreano, Mototaka Suzuki, Guido De Croon, and one anonymous reviewer for useful comments on the manuscript.

References

Aloimonos, J., Bandopadhay, A., and Weiss, I. (1988). Active vision. *International Journal of Computer Vision*, 1(4):333–356.

Bajcsy, R. (1988). Active percetion. In *Proceedings of the Institute of Electrical and Electronics Engineers*, volume 76, pages 996–1005.

Ballard, D. (1991). Animate vision. *Artificial Intelligence*, 48(1):1–27.

Beer, R. D. (1996). Toward the evolution of dynamical neural networks for minimally cognitive behavior. In Maes, P., Mataric, M., Meyer, J., Pollack, J., and Wilson, S., editors, *From animals to animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, pages 421–429, Cambridge, MA. MIT Press.

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, 4(3):91–99.

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

Bickhard, M. H. (2001). Why children Don't have to solve the frame problems: Cognitive representations are not encodings. *Developmental Review*, 21:224–262.

Borghi, A. (2005). Object concepts and action. In Pecher, D. and Zwaan, R., editors, *Grounding Cognition: The role of perception and action in memory, language, and thinking*. Cambrige University Press, Cambridge.

Burr, D. (2004). Eye movements: Keeping vision stable. *Current Biology*, 14(5):195–197.

Churchland, P., Ramachandran, V., and Sejnowski, T. (1994). A critique of pure vision. In Koch, C. and Davis, J. L., editors, *Large scale neuronal theories of the brain*, pages 23–60. MIT Press, Cambridge, MA.

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

Cohen, H. and Lefebvre, C., editors (2005). *Handbook of Categorization in Cognitive Science*. Elsevier, Oxford.

de Croon, G., Postma, E., and van den Herik, H. (2006). A situated model for sensory-motor coordination in gaze control. *Pattern Recognition Letters*, 27(11):1181–1190.

Di Ferdinando, A. and Parisi, D. (2004). Internal representations of sensory input reflect the motor output with which organisms respond to the input. In Carsetti, A., editor, *Seeing, thinking and knowing*, pages 115–141. Kluwer, Dordrecht.

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

Floreano, D., Husband, P., and Nolfi, S. (2008). Evolutionary robotics. In Siciliano, B. and Oussama, K., editors, *Handbook of Robotics*, pages 1423–1451. Springer Verlag, Berlin.

Floreano, D., Kato, T., Marocco, D., and Sauser, E. (2004). Coevolution of active vision and feature selection. *Biological Cybernetics*, 90(3):218–228.

Floreano, D., Suzuki, M., and Mattiussi, C. (2005). Active Vision and Receptive Field Development in Evolutionary Robots. *Evolutionary Computation*, 13(4):527–544. The final version of this article has been published, in Evolutionary Computation (http://www.mitpressjournals.org/loi/evco), Vol. 13, Issue 4, published by The MIT Press.

Gallese, V. and Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in reason and language. *Cognitive Neuropsychology*, 22:455–479.

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

Guthrie, B. L., Porter, J. D., and Sparks, D. L. (1983). Corollary discharge provides accurate eye position information to the oculomotor system. *Science*, 221(4616):1193–1195.

Hafed, Z. and Krauzlis, R. (2006). Ongoing eye movements constrain visual perception. *Nature Neuroscience*, 9(11):1449–1457.

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

Harvey, I., Husbands, P., and Cliff, D. (1994). Seeing the light: artificial evolution, real vision. In *From animals to animats 3: Proceedings of the third international conference on Simulation of adaptive behavior*, pages 392–401, Cambridge, MA. MIT Press.

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

Kato, T. and Floreano, D. (2001). An Evolutionary Active-Vision System. In *The 2001 Congress on Evolutionary Computation*, volume 1, pages 107–114.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. W. H. Freeman, San Francisco.

Merleau-Ponty, M. ([1948] 1973). *The Visible and the Invisible*. Northwestern University Press, Evanston, IL.

Merriam, E. P. and Colby, C. L. (2005). Active vision in parietal and extrastriate cortex. *Neuroscientist*, 11(5):484–493.

Mirolli, M. (submitted). Representations in dynamical embodied agents: Re-analyzing a minimally cognitive model agent. *Cognitive Science*.

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

Nolfi, S. (1997). Evolving non-trivial behavior on autonomous robots: Adaptation is more powerful than decomposition and integration. In Gomi, T., editor, *Evolutionary Robotics*, pages 21–48. AAI Books, Ontario (Canada).

Nolfi, S. (2002a). Evolving robots able to self-localize in the environment: The importance of viewing cognition as the result of processes occurring at different time scales. *Connection Science*, 14(3):231–244.

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

Nolfi, S. (2005). Categories formation in self-organizing embodied agents. In Cohen, H. and Lefebvre, C., editors, *Handbook of Categorization in Cognitive Science*, pages 869–889. Elsevier Ltd, Oxford.

Nolfi, S. (in press). Behavior and cognition as a complex adaptive system: Insights from robotic experiments. In Hooker, C., editor, *Philosophy of Complex Systems, Handbook on Foundational/Philosophical Issues for Complex Systems in Science*. Elsevier.

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. (2001). Evolving robots able to integrate sensory-motor information over time. *Theory in Biosciences*, 120(3):287–310.

Nolfi, S. and Marocco, D. (2002). Active perception: A sensorimotor account of object categorization. In Hallam, B., Floreano, D., Hallam, J., Hayes, G., and Arcady-Meyer, J., editors, *From Animals to Animats 7: Proceedings of the VII International Conference on Simulation of Adaptive Behavior*, pages 266–271, Cambridge, MA. MIT Press.

Noton, D. and Stark, L. (1971). Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision Research*, 11(9):929–932.

O'Regan, J. K. and Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5):939–1031.

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

Rizzolatti, G., Camarda, R., Fogassi, M., Gentilucci, M., Luppino, G., and Matelli, M. (1988). Functional organization of inferior area 6 in the macaque monkey: Ii. area f5 and the control of distal movements. *Experimental Brain Research*, 71:491–507.

Rizzolatti, G., Fogassi, L., and Gallese, V. (2002). Motor and cognitive functions of the ventral premotor cortex. *Current Opinion in Neurobiology*, 12(2):149–154.

Scheier, C., Pfeifer, R., and Kunyioshi, Y. (1998). Embedded neural networks: exploiting constraints. *Neural Network*, 11(7-8):1551–1569.

Schmidhuber, J. and Huber, R. (1991). Learning to generate artificial fovea trajectories for target detection. *International Journal of Neural Systems*, 2(1-2):135–141.

Sommer, M. A. and Wurtz, R. H. (2006). Influence of the thalamus on spatial visual processing in frontal cortex. *Nature*, 444(7117):374–377.

Sommer, M. A. and Wurtz, R. H. (2008). Brain circuits for the internal monitoring of movements. *Annual Review of Neuroscience*, 31(1):317–338.

Suzuki, M. and Floreano, D. (2006). Evolutionary active vision toward three dimensional landmark-navigation. In Nolfi, S., Baldassarre, G., Calabretta, R., Hallam, J., Marocco, D., Miglino, O., Meyer, J.-A., and Parisi, D., editors, *From Animals to Animats 9: Proceedings of the Ninth International Conference on the Simulation of Adaptive Behavior*, volume 4095 of *LNAI*, pages 263–273, Berlin. Springer-Verlag.

Suzuki, M. and Floreano, D. (2008). Enactive Robot Vision. *Adaptive Behavior*, 16(2-3):122–128.

Thornton, C. (1997). Separability is a learner's best friend. In Bullinaria, J., Glasspool, D., and Houghton, G., editors, *Proceedings of the Fourth Neural Computation and Psychology Workshop: Connectionist Representations*, pages 40–47, Berlin. Springer-Verlag.

Tuci, E., Massera, G., and Nolfi, S. (in press). Active categorical perception of object shapes in a simulated anthropomorphic robotic arm. *IEEE Transaction on Evolutionary Computation*.

Tucker, M. and Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology: Human Perception and Performance*, 24(3):631–647.

Usher, M. and McClelland, J. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review*, 108:550–592.

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.

Yarbus, A. L. (1967). *Eye Movements and Vision*. Plenum Press, New York.

List of Figures

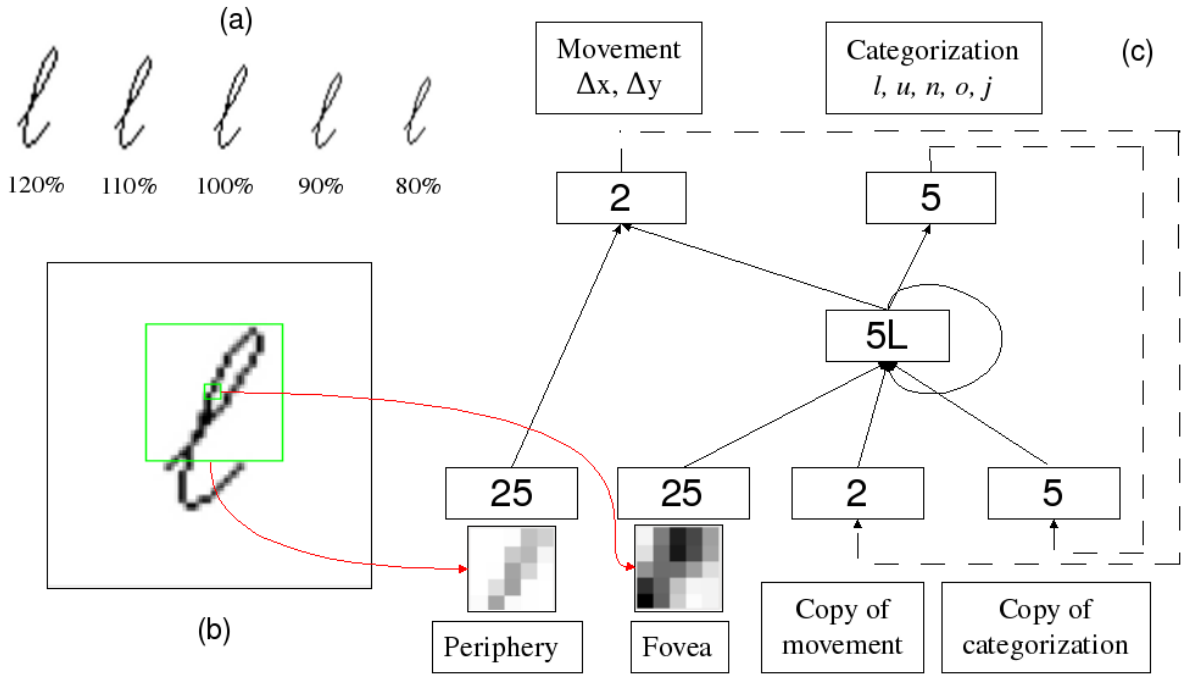
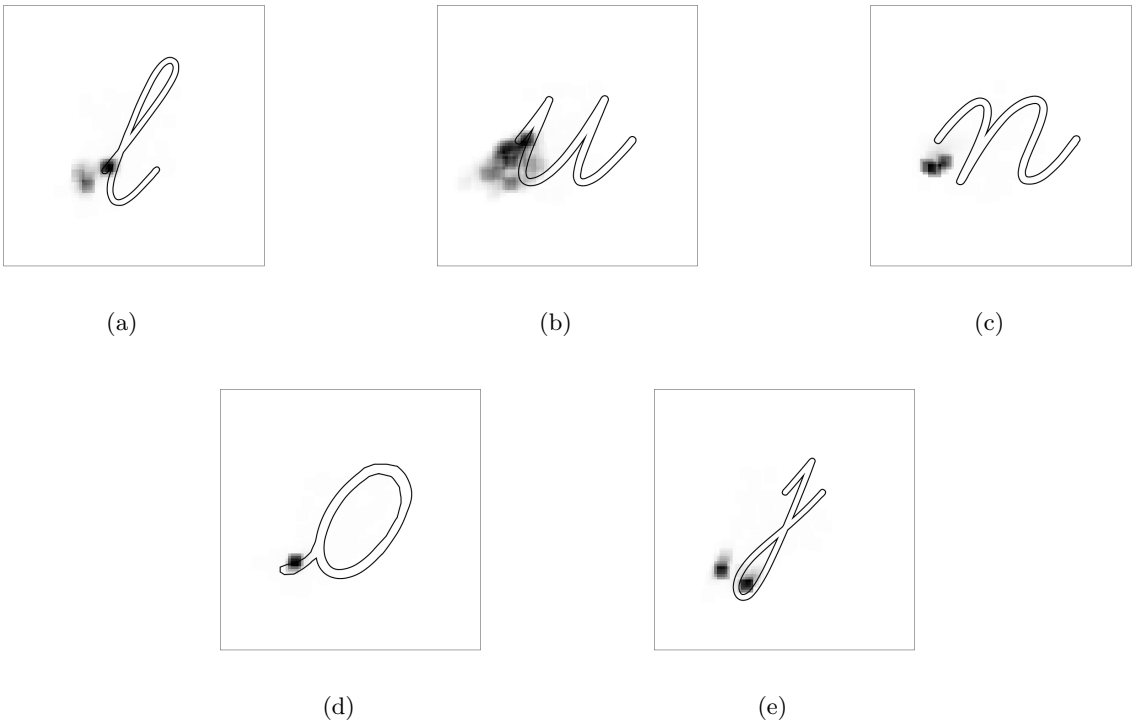


Fig. 1. The experimental set-up. (a) Letter 'l' shown in the 5 different sizes used in the experiment. (b) The screen displaying the letter 'l' in its intermediate size and an exemplification of the field of view of the foveal and peripheral vision (smaller and larger squares, respectively). (c) The architecture of the neural controller. On the bottom, the 5 x 5 periphery sensors encode the average grey level of a square of 10 by 10 pixels each, the 5 x 5 fovea sensors encode the grey level of one pixel each, the other two blocks of two and five input units encode, respectively, the states of the motor and categorisation neurons at the previous time step. Neurons are logically organised in blocks. The number inside the each rectangle indicates the number of neurons. The letter 'L' included in the block of 5 internal neurons indicates that these neurons are leaky integrators (see text). Continuous arrows indicate connections. More specifically, all neurons of the block at the end of the arrow receive connections from all neurons of the block at the beginning of the arrow. Dashed arrows indicate that the activation of the output units at time t is copied in the respective input units at time $t + 1$.



(a) (b) (c)

(d) (e)

Fig. 2. Frequency of observations of different parts of the image for letters of different categories. Each figure plots the data obtained by testing the best evolved individual for 50 trials (10 repetition starting from randomly different initial position for 5 different letter sizes). Gray levels represent the frequency with which each portion of the image has been perceived by one of the photoreceptors of the fovea, with darkness representing high frequencies. To facilitate the interpretation of the data, the pictures show only the contour of the average size letter.

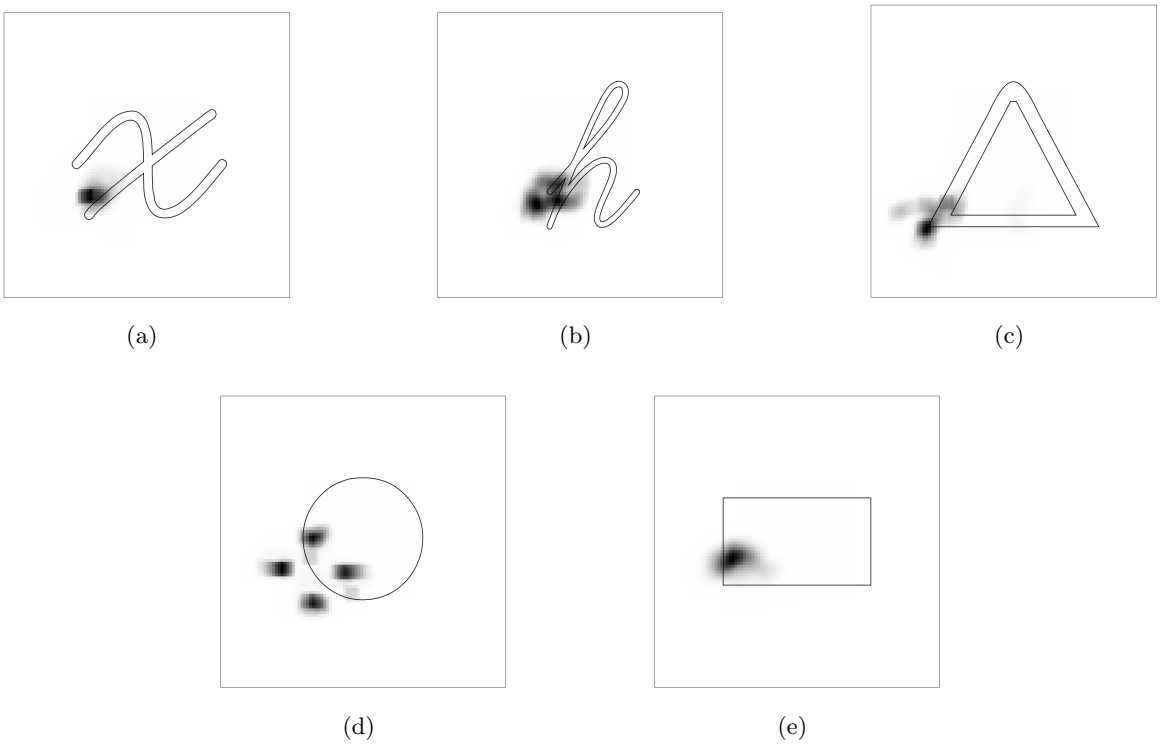
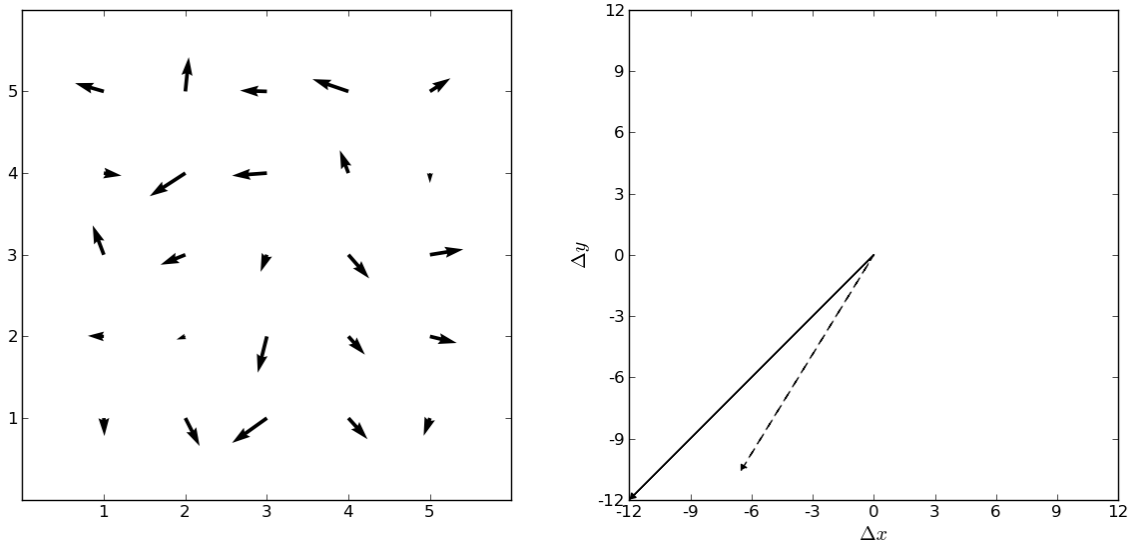


Fig. 3. Frequency of observations of different parts of the image when the agents is tested with new types of images never experienced before. Data obtained by using the procedure described in the caption of figure 2. Average size images are shown. To facilitate the interpretation of the data, pictures (a), (b), and (c) show only the contour of the images. The circle (d) was completely filled with black pixels, while the rectangle (e) was filled with pixels with a random gray level in $[0, 1]$.



(a) (b)

Fig. 4. (a) Motor actions triggered when each of the 5 x 5 photoreceptors of the peripheral vision (distributed over the two dimensional plane corresponding to the x and y axes of the figure) is fully activated while the activation of all other photoreceptors is set to 0. The orientation and length of each arrow represent, respectively, the orientation and amplitude of the eye movement triggered when the corresponding photoreceptor is activated. (b) Eye movements triggered by homogeneous peripheral patterns in which all photoreceptors have the same activation value. The continuous and dashed lines indicate the movement produced when all peripheral photoreceptors are activated, respectively, to 1.0 and 0.05 (which corresponds to the average neuron's activation in ecological conditions). The x and y axes indicate the amplitude of the movement in pixels along the horizontal and vertical and axes, respectively. The orientation of the arrow with respect to the centre of the graph indicates the direction of the movement.

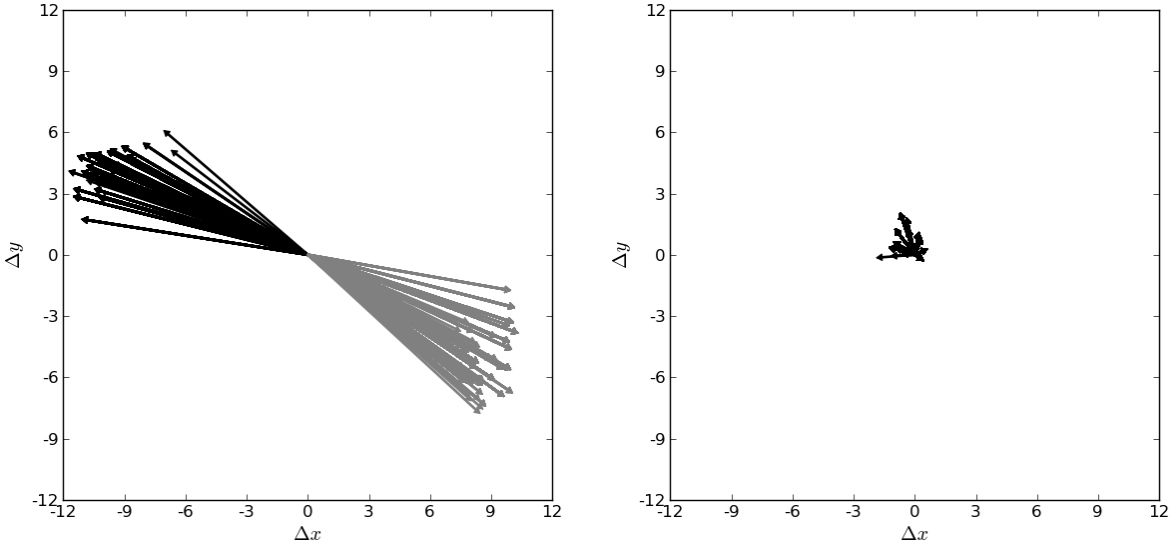


Fig. 5. Eye movements produced by the agent for letter j (a) and n (b) during the last 50 time steps of 50 trials for each letter (10 trials for each of the five letter sizes experienced during the evolutionary process). The x and y axes indicate the amplitude of the movement in pixels along the horizontal and vertical and axes, respectively. In (a), black and gray arrows indicate the actions triggered by two different sets of visual pattern found through a k-means clustering algorithm: each cluster corresponds to the stimuli experienced while the agent foveates one of the the two highly explored areas of letter 'j' shown in figure 2e.

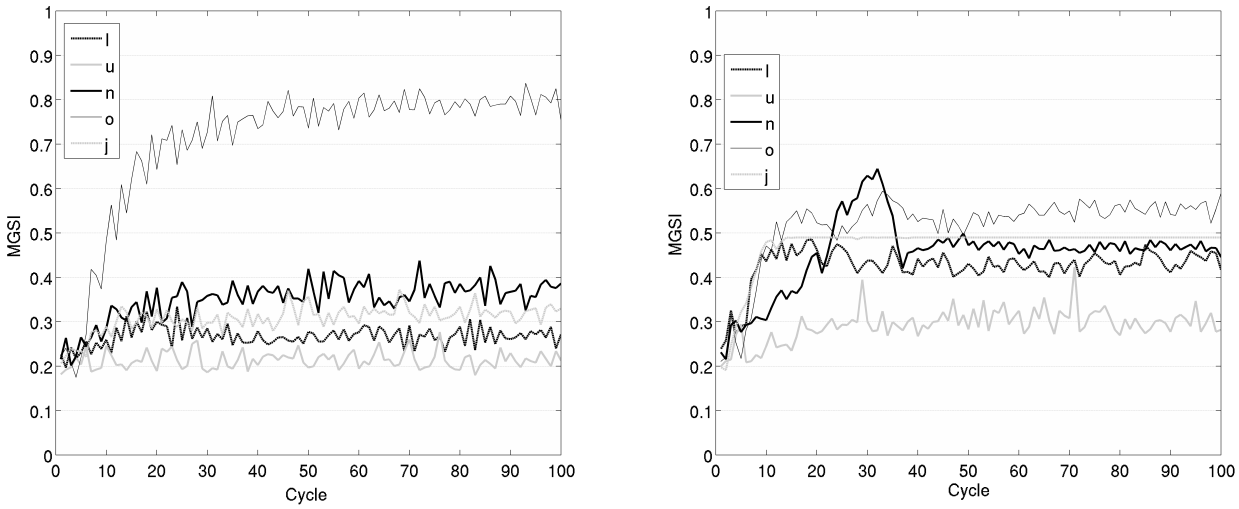


Fig. 6. Modified Geometric Separability Index (MGSI) of the stimuli provided by the foveal photoreceptors (a) and by the efference copy of the motors (b). Each point along the x axis represents the value of the MGSI calculated over the stimuli experienced during the corresponding time step.

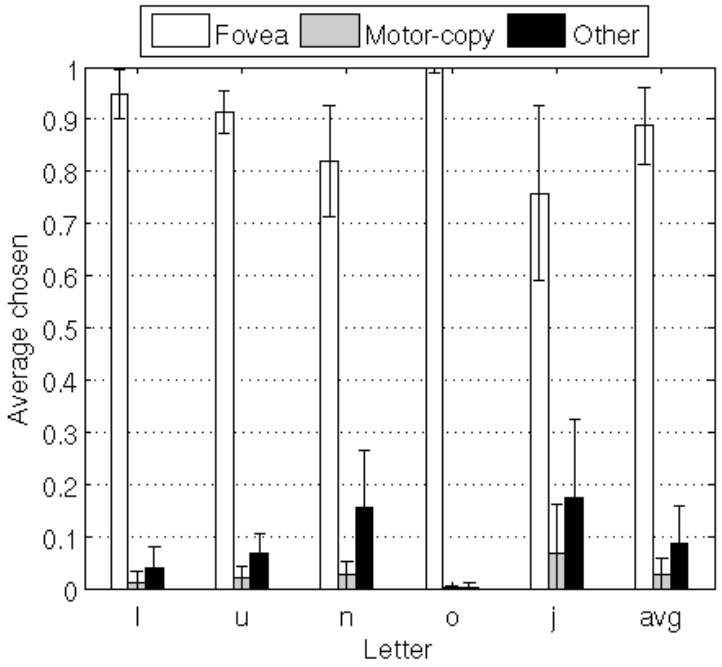
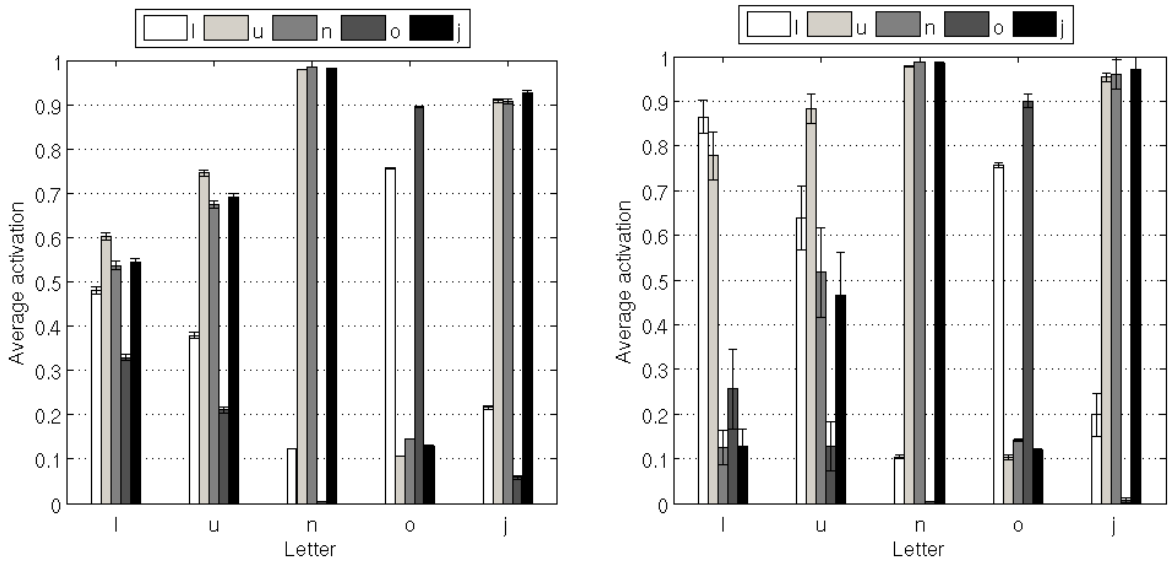


Fig. 7. Percentage of times in which the categorisation answers produced by the best controller corresponds to the category of the state of the foveal photoreceptors ('Fovea'), to the category of the state of the motor-copy neurons ('Motor-copy'), or to another category ('Other'). Each group of histogram represents the average performance for each category. Data obtained in a control experiment in which the controller experience pre-recorded input states corresponding to all possible combinations of categories over the two input channels. Bars represent standard error.



(a) (b)

Fig. 8. (a) Average categorisation patterns produced in trials in which the agent experienced each possible visual stimulus encountered in natural conditions in a given categorical context for the entire duration of the trial (see text). (b) Average categorisation patterns produced in normal conditions. Each group of histograms represents the average categorisation pattern produced for the corresponding letter indicated in the horizontal axis. Each histogram of a group represents the average activation of the corresponding categorisation output unit (i.e. 'l', 'u', 'n', 'o', and 'j').

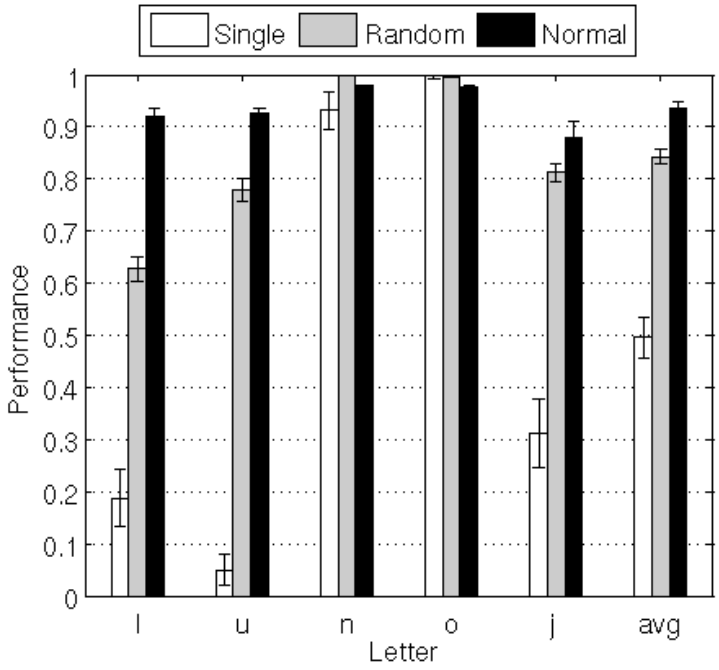
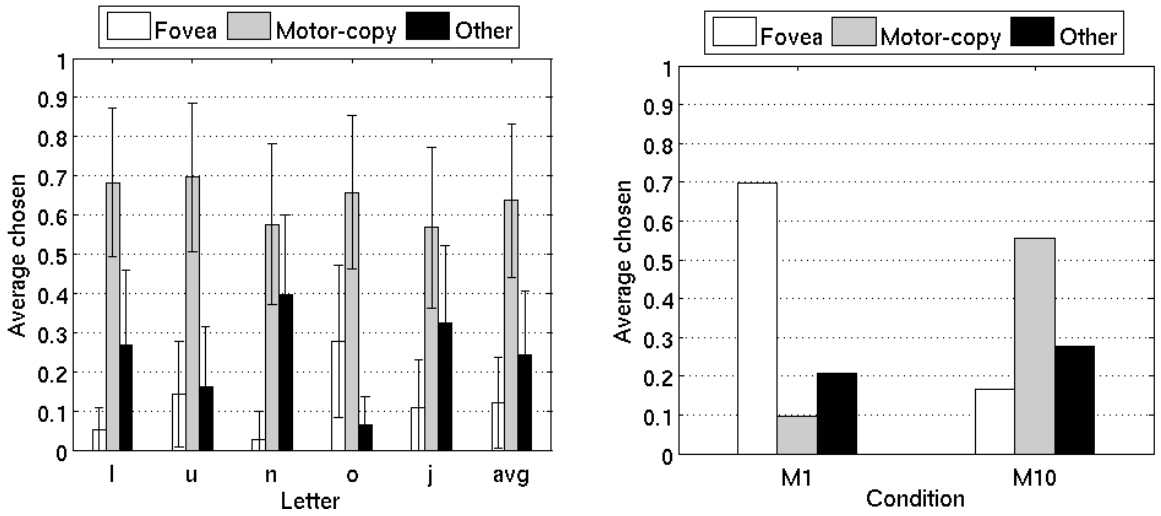


Fig. 9. Comparison of correct responses (i.e. percentage of times in which the most activated categorisation unit corresponds to the correct category) of the best individual of all replications in three conditions. Single (white histograms): the agent perceives a single foveal pattern recorder for a given letter for the entire duration of the trial (the motor-copy input is kept fixed at $[0.5; 0.5]$). Random (grey histograms): the foveal receptors are fed, for 100 consecutive cycles, with randomly chosen visual patterns belonging to a given letter (the motor-copy input is kept fixed at $[0.5; 0.5]$). Normal (black histograms): normal condition (i.e. when the agent is allowed to autonomously interact with the images).



(a) (b)

Fig. 10. Percentage of times in which the categorisation answers produced by the best controller corresponds to the letter presented in the fovea (white histograms), to the letter presented in the motor copy (grey histograms), or to another letter (black histograms). (a) Average results for each letter and over all letters in the case of the best individual. (b) Average results of the best individuals of all replication of the first (M1) and second (M10) series of experiments (in which the efference copy of the motor are normalized in $[0; 1]$ and $[0; 10]$, respectively).

List of Tables

Table 1. Performance obtained in a series of control experiments performed by varying the architecture of the agents neural controller or the agents sensory system. For each experiment we report the average performance of the 20 best evolved individuals obtained in 20 replications of the experiment (*Average*), the performance of the best individual (*Best*), and the number of replications of the experiment in which the evolved individuals reach a performance equal or greater than 0.9 (*Successful*). Performance have been computed by testing each individual for 10000 trials. *Basic*: the experiments performed in the basic experimental conditions described in section 1, *Gain-10*: the experimental condition described in section 3.6, in which the range of the efference copy of the motor neurons is normalized in $[0, 10]$. *Fovea-Only*: the fovea is connected to both the internal and the motor units, and there is no periphery. *Periphery-Only*: the periphery is connected to both the internal and the motor units, and there is no fovea. *No-Fovea*: there is no fovea and peripheral photoreceptors are linked only to the motor neurons. *No-Fovea-Gain-10*: as for No-Fovea, but the efference copy of the motor neurons is normalized in $[0, 10]$. *No-Time-Processing*: the internal neurons are updated on the basis of a standard logistic function and there are no recurrent connections. *No-Internal-Neurons*: the photoreceptors of the fovea and of the efference copy of the motor neurons are connected directly to the categorisation output units.

	Basic	Gain-10	Fovea-Only	Periphery-Only	No-Fovea	No-Fovea-Gain-10	No-Time-Processing	No-Internal-Units
Best	0.9432	0.9831	0.4128	0.9987	0.8040	0.9963	0.8113	0.7270
Average	0.7692	0.8463	0.2903	0.8685	0.5976	0.7497	0.5928	0.5581
Successfull	2	8	0	14	0	5	0	0