

# 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 through a genetic algorithm for categorising five different kinds of images (letters) of different sizes. The visual system is composed of a periphery (large visual field with low resolution) and a fovea (small visual field with high resolution) that can freely move over the images. As a consequence, the system solves its task by exploiting its sensory-motor interactions with its (visual) environment. The analysis of the behaviour of the evolved system indicates that categorisation is achieved by: (1) acting so to fall in behavioural attractors that are specific for each category, and (2) integrating perceptual and/or motor information over time through a process of accumulation of partially conflicting evidences. We discuss our results with respect to the general theory of categorisation and, in particular, 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 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, 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 Noe: “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.” (Noe, 2004, page 35).

Theoretical and experimental evidences 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; Noe, 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 following 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 to co-determine its own sensory states through its actions and the possibility to integrate the experienced sensory-motor states over time to solve a perceptual categorisation problem. Before describing the experimental scenario and the obtained results, we first illustrate in the next section the relationship between the present work presented in this paper and the most relevant literature of 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 (the Artificial Intelligence, Neural Network, and Robotics).

Recent work performed by evolving artificial embodied agents to produce different behaviours in different environmental contexts 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 behaviors 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) 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 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 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 in contexts 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 robots display an ability to solve the task despite the four corners are perceptually identical and despite 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 robots solve the problem by reaching one of the four corners, leaving the corner with an angle of about  $45^\circ$ , 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 with such 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 good 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 from the agents' behaviour. By selecting the appropriate behaviours, the agent can manage to select a sub-set of sensory states that it is not perceptually ambiguous. More generally, all these experiments point to the fact that the stimuli sensed by an embodied agent depend from the environment, from the agents' sensory system, and from the agent's behaviour (and not only from 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 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 producing 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 bad corners but does not allows to remain on one of the two good 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 keep moving back and fourth between the two right corners. Only robots provided with internal neurons and recurrent connections are able to stop in one of the two good 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 states 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 categorizing as well as partial contradictory evidences. The problem due to the presence of partially ambiguous evidences 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 understanding of categorization 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). In order to analyse how such an agent can exploit its eye movements to enable the categorisation process, and given the strong interdependence between eye’s motion control and visual categorisation, we decided to train 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 25 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 to do this. Rather, the characteristics of the set-up have been chosen so 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.

[Fig. 1 about here.]

## 2.2 The neural network controller

Agents are provided with a neural network controller with the architecture shown in Figure 1c. The controller includes 57 sensory 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 photoreceptors of the fovea in each time step in order to take into account the fact that the grey level measured by the photoreceptor is subjected 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  $[-25; 25]$  pixels of the image. The five categorisation units allow the agent to label the category of the five corresponding letter (see below). The sensory 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 from the input received from the sensory neurons through the weighted connections and by their own activation at the previous time step, and is calculated as follow:

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

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  $\tau_i$  is the time constant of unit  $i$ , in  $[0; 1]$ .

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

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

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  $[-25; 25]$  and used to vary the position of 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 for 50 trials lasting 100 time steps each. At the beginning of each trial: (a) the screen is set so to display one of the five different letters in one of the five different sizes (each letter of each size is presented twice to each individual), (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 in 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), which measures (a) the agents' ability to activate the categorisation unit corresponding to the current category more than the other units and (b) the ability to maximise the activation of the right unit while minimising those of the other units:

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

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). This two-components fitness function produced better results than a single component fitness function rewarding the first or the second component alone (results not shown). Notice that individuals are not rewarded for moving their eyes or 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 100 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 consists of copies with the addition of mutations (i.e. with 2% of the bits replaced by randomly selected values). Such 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 categorize correctly also 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 also that agents tend to produce different behaviours for letters belonging to different categories and similar behaviours for letters belonging to the same category, independently from the letters' 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 both sensory channels (i.e. in the fovea and in the motor efference copies), which might be almost equally used for letter discrimination 3.3. On the other hand, the analysis of the actual role played in the categorisation behaviour of the best evolved agent by the two channels indicates that the visual channel plays the most significant role (section 3.4). The analysis of the process through which the agent discriminates letters indicates that categorization is based on a simple process of evidences accumulation (section 3.5). Finally, the analysis of a new set of experiments demonstrates that the relative importance of the two sensory channels (visual and motor) for the categorisation process depends on the way in which sensory information is encoded (section 3.6).

#### 3.1 Performance analysis

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 for 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). As indicated in figure 2, 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.

[Fig. 2 about here.]

### 3.2 Behavioural analysis

By analysing the behaviour displayed by the best individual we can see how, after an initial phase lasting 5-30 time steps (in which the behaviour varies significantly for different initial positions of the eye and for different letter sizes), the behaviour of the agent converges either on a fixed point attractor (i.e. the eye stops moving after having reached a particular position of the letter) or on a limit cycle attractor (i.e. the eye keeps moving 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 different behaviours for letters of different categories (see Figure 3, and the movies available from <http://laral.istc.cnr.it/mirolli/activevision.html>).

[Fig. 3 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 the letter 'n' (Figure 3c), the agent's eye stops while foveating the white background located on the lower-left side of the letter. In case the letter 'o' (Figure 3d), it stops while foveating a small sector of the arc that form the lower-left side part of the letter. In the case of these two letters, therefore, the agent/environment system converges on fixed point attractors that produce specific visual stimuli that represent discriminative stimuli only on the basis of the behavior consistently exhibited by the agent, a strategy already observed in some of the experiments reviewed above (Scheier et al., 1998; Nolfi and Marocco, 2002; Floreano et al., 2004). However, these stimuli are not fully discriminative by themselves. For example, as we will see below, the stimulus corresponding to a full white area on the fovea is repeatedly experienced also with other letters. But only with letter 'n' it is the only stimulus that is perceived after the initial phase.

In fact, in the case of the other three letters, the agent converges on three different limit cycles. In the case of the letter 'l' (Figure 3a), the eye keeps jumping back and forth between an area in the bottom-left of the letter and a blank area located in the further bottom-left area. When interacting with a 'j' (Figure 3e), 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 3b), the agent keeps producing a circular trajectory by keep 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.

By observing the behaviour displayed by the best individuals of the other replications of the experiment we observed that all evolved individuals converge either on fixed point attractors or on limit cycles, depending on the category of the letter. However, the specific areas visited by the agent and the relation between the category of the letter and limit state that is reached (fixed point or limit cycle) vary in different replications. Similarly, other replications of the experiment do not exhibit a general preference for a specific area of the image overall, like it happens for the agent just analysed, who tends to end up always in the lower-left part of the images.

### 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 stimuli belonging to different categories in the two sensory spaces (of the fovea and of the motor copies) while the agent interacts with the environment. 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 takes into account not only the nearest neighbour, but all the stimuli belonging to same category. Hence, we devised what we call the Modified Geometric Separability Index (MGSI), which is defined as the average, over all patterns, of the proportion of the  $N^k - 1$  stimuli belonging to the same category that are in the  $N^k - 1$  nearest stimuli, with  $N^k$  representing the number of patterns of category  $k$ . More formally, the MGSI is calculated as follows:

$$MGSI(P) = \frac{\sum_{x \in P} \frac{\sum_{n \in N_x} I_{C_x}(n)}{|C_x|}}{|P|} \quad (4)$$

where  $|S|$  indicates the cardinality of the set  $S$ ,  $P$  is the set comprising all the patterns,  $C_x$  is the set of all patterns belonging to the same category as pattern  $x$  ( $x$  doesn't belong to  $C_x$ ),  $N_x$  is the set of the  $|C_x|$  patterns nearest to pattern  $x$  and  $I_{C_x}(n)$  is the indicator function of set  $C_x$ , which returns 1 if  $n$  is in the set  $C_x$ , 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 in the 5 different sizes for 10 times each. For each type of sensory information, the MGSI index has been calculated for each of the 100 steps composing a trial, so that we could observe the evolution of the MGSI during the agent's interactions with the images.

As shown in figure 4, the separability of the stimuli for both sensory channels (visual and motor) significantly increases at the beginning of the trial, stabilizing on intermediate values after about 30-40 time steps. The fact that the MGSI index increases confirm that the behaviour exhibited by the agent allows it to perceive more discriminative stimuli, thus facilitating the categorisation process. On the other hand, fact that the MGSI index does not reach very high values (i.e. values around 1.0) implies that sensory-motor coordination does not lead to a situation in which the stimuli belonging to different categories are fully separated, as was already demonstrated by the observation, reported in the previous section, that some stimuli are experienced in interaction with different categorical contexts.

[Fig. 4 about here.]

### 3.4 On the role of different sensory channels

The analysis reported in the previous section demonstrates that both sensory channels provide cues for discriminating the current category. To verify whether the best adapted agent exploits both types of regularities or whether it use 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 sensors 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 time 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 none of the two (figure 5, white, grey, and black histograms, respectively), we can see that the categorisation behavior depends almost exclusively on the visual input. Indeed, the average probabilities that the agent's categorization output corresponds to the category of the data experienced in the foveal photoreceptors, in the motor-copy sensors, or to another category are 0.8874, 0.0255, and 0.0871, respectively. Furthermore, the analysis of performance for each letter (figure 5) 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 sensory 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 sensory channels depends on the way in which sensory 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 to overcome the problem that the experienced stimuli are not fully discriminative.

[Fig. 5 about here.]

### 3.5 On the dynamics of the categorisation process

The results described above demonstrates that the categorization process is based primarily on the stimuli experienced through the foveal photoreceptors experienced as a result of the execution of specific behaviours. Moreover, the analysis reported in the previous section indicates that the categorization process should also involve an ability to integrate the experience sensory-motor information over time since the stimuli belonging to different categories

are not well separated in the input space and since, in some cases, the same stimuli are experienced in different categorical contexts.

In order to verify 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 sensors has been set to  $[0.5; 0.5]$  (that represents a stay-still action) and the state of the foveal photoreceptors has been 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 output of the categorisation units converge on a fixed value (i.e. it changes less than 0.01 for 5 contiguous time steps).

The fact that the correct categorization output unit results the most activated also in the average categorization patterns produced during these tests (Figure 6a) for all letters but the l, together with the fact that the average patterns produced during this test are similar to the average patterns produced in normal conditions (Figure 6b), indicates that the order with which stimuli are experienced plays a minor role and that categorization process is determined primarily on the basis of type and frequency of the stimuli and that are experienced.

[Fig. 6 about here.]

Finally, we compared categorisation performance (i.e. the percentage of time in which the most activated categorisation unit corresponds to the perceived letter) in three conditions: (1) the normal condition (i.e. when the agent is let free 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 motor-copy is kept fixed at  $[0.5; 0.5]$  and the foveal photoreceptors are fed, for 100 consecutive cycles, with one visual pattern chosen randomly between all the 2500 patterns recorded in normal conditions (i.e. during the last 50 cycles of 50 trials performed with letters of 5 different dimensions).

As shown in figure 7, the randomisation of the order of the visual stimuli produces a very limited degradation of agent’s categorisation performance (i.e. from 0.94 to 0.84: black and grey bars in figure 7, respectively). On the contrary, the average performance when the categorisation behaviour depends on a single visual pattern drops down to 0.51 (white bar). This further test demonstrates that single sensory patterns do not provide enough information to discriminate the category in about half of the cases. Moreover, this test demonstrates that the order with which stimuli are experienced plays a minor role.

To summarize, the categorisation problem is solved through a process of accumulation of partially conflicting evidences. Evidences correspond to the strength of the association between each stimulus and each corresponding categorical output, while this association in turn can be considered as a rough estimation of the probability that that stimulus is experienced in the corresponding categorical context. During such process, some of the experienced stimuli produce a relative increase of the wrong categorisation units. Overall, however, the summed contribution of the evidences provided by all experienced sensory stimuli ensures that the correct categorisation unit reaches the highest activation level.

[Fig. 7 about here.]

### 3.6 On the role of visual versus motor information

In the previous section we demonstrated 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 ascertaining 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 sensory channel tends to have a

quantitative 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 run a new set of experiments in which the number of sensory 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  $[0; 1]$ .

First of all, the comparison between the performance obtained in the first and second set of experiments (Figure 8a) indicates that the new encoding of motor information leads to slightly better results. This trend is confirmed also by the comparison of the performance obtained in each replication of the experiment (see Figures 2 and 8b, for the first and second experiment respectively). 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 achieve a performance higher than 0.9 are 2 out of 20 and 8 out of 20 in the first and second set of experiments, respectively.

[Fig. 8 about here.]

In order to assess the relative role of the two sensory channels in the new set of experiments, we subjected the best evolved individual 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 9a). 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 9b).

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. 9 about here.]

## 4 Discussion

### 4.1 Categorising by integrating information through time

In this paper we presented an active vision system that is trained through 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. In other words, the system is let 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 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 conflicting evidences in 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 kind of strategy seems to be the simpler that is sufficient for solving the given task. In fact, it seems that, on the one hand, finding a way of coordinating the sensory-motor process so to experience fully discriminative stimuli is not possible, while, on the other hand, taking into account the order with which stimuli are experienced is not necessary (at least in most of the cases). This kind of strategy, based on the accumulation of partially conflicting perceptual evidences, thus represents an extension of the sensory-motor strategies reviewed at the beginning of the paper that can be successfully applied in cases in which stimuli corresponding to different categories cannot be fully separated.

An important topic for future research will be to study under which conditions categorization might require to take into account also the order with which stimuli are experienced. Some evidences in this direction might also

be gathered by further analyses to be conducted on the experiments described in section (sec. 3.6, in which the order with which stimuli are experienced seems to play a more important role. Another interesting issue for future research consists in the study of the conditions that might lead to still more complex categorization forms involving the exploitation of sensory-motor contingencies, that is, the anticipation of the sensory consequences of agents's own actions (Noton and Stark, 1971; O'Regan and Noë, 2001; Noe, 2004).

Finally, another interesting topic for future research is the investigation of how different source of information (e.g. motor and visual) can be fused so to produce better performance with respect to those that can be obtained on the basis of either source of information in isolation. Still a further issue to be investigated is whether different sensory 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, such as the accumulation of partially conflicting evidences, which might not suffice for motor information.

## 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 organisms' 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 of 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 and mirror neurons: e.g. Rizzolatti et al., 1988; Gallese et al., 1996; Rizzolatti et al., 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 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 our last experiment, the one in which we changed the encoding of the motor input (sec. 3.6), 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, than the categorisation process can be based not only on the information gathered by the external environment but also on the information regarding agents own behavior. 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 result 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 input.

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. Interestingly, very recently Hafd and Krauzlis (2006) have re-vitalised this idea by showing through behavioural studies that eye movements can significantly improve 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.

## References

- Aloimonos, J., Bandopadhyay, A., and Weiss, I. (1988). Active vision. *International Journal of Computer Vision*, 1(4):333–356.
- Bajcsy, R. (1988). Active perception. 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*. Cambridge 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.
- 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 Psychology of 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.

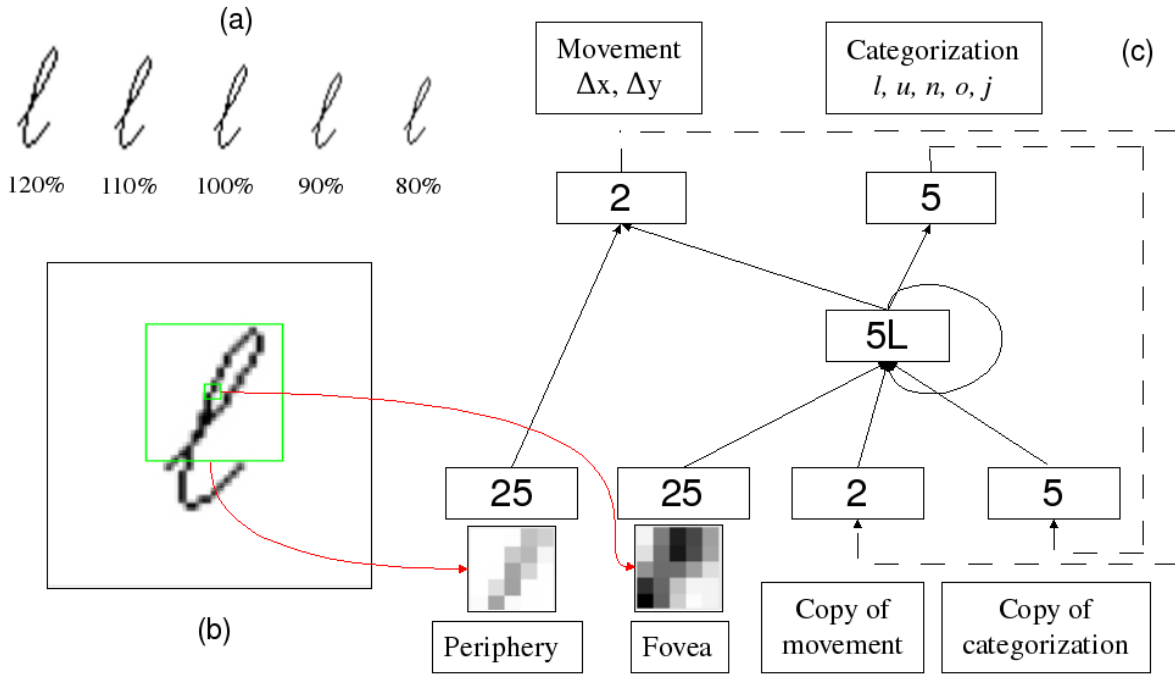
- Floreano, D., Kato, T., Marocco, D., and Sauser, E. (2004). Coevolution of active vision and feature selection. *Biological Cybernetics*, 90(3):218–228.
- Floreano, D., Suzuki, M., and Mattiussi, C. (2005). Active Vision and Receptive Field Development in Evolutionary Robots. *Evolutionary Computation*, 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., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119:593–609.
- 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.
- 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.
- Noe, 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 Kuniyoshi, 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.
- 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.
- 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.

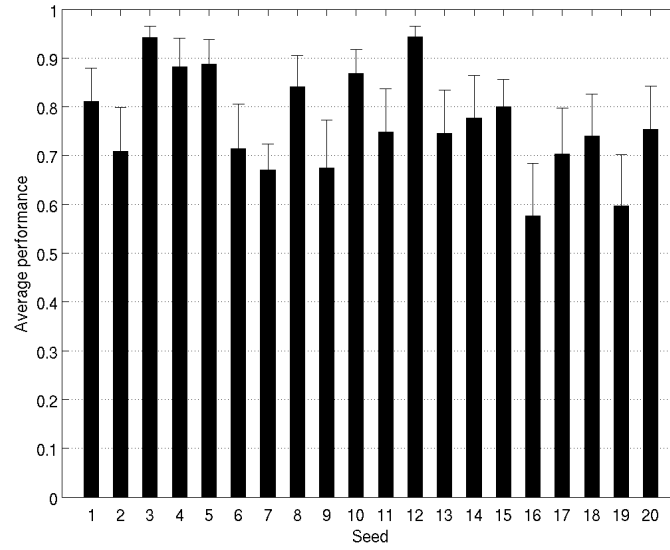
## List of Figures

- 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 sensors 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  $t + 1$ . . . . . 19
- 2 Percentage of correct responses for the best individual of each replication of the experiment. Data obtained by testing each individual for 10000 trials during which it is exposed for 40 times to 5 letters with 50 different dimensions) The lines at the top of each histogram represent the standard error. . . . . 20
- 3 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. . . . . 21
- 4 Modified Geometric Separability Index (MGSI) of the stimuli provided by the foveal photoreceptors and by the efference copy of the motors (thick and thin lines, respectively). Each point along the x axis represents the value of the MGSI calculated over the stimuli experienced during the corresponding time step. . . . . 22
- 5 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 sensors (‘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 sensory states corresponding to all possible combinations of categories over the two sensory channels. Bars represent standard error. . . . . 23
- 6 (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. (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 histograms of a group represents the average activation of the corresponding categorisation output unit (i.e. ‘l’, ‘u’, ‘n’, ‘o’, and ‘r’). . . . . 24
- 7 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 let free to autonomously interact with the images). . . . . 25
- 8 (a) Comparison between the average performance of the best individuals of all replications in the experiments in which the state of the motor sensors is normalised in [0; 1] (M1) and in [0; 10] (M10), respectively. (b) Percentage of correct responses for the best individual of each replication of the second experiment in which the state of the efference copies of the motor neurons is normalised in the range [0; 10]. Data are obtained by testing each individual for 10000 trials during which it is exposed for 40 times to the 5 letters with 50 different dimensions. . . . . 26

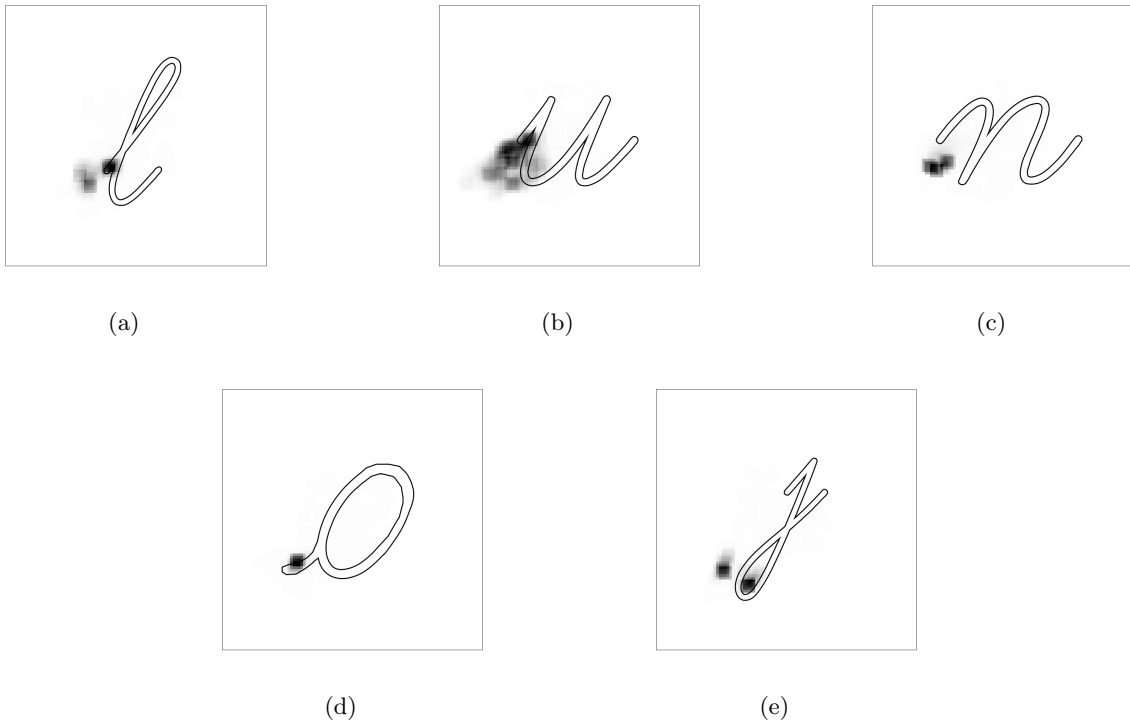
- 9 Percentage of times in which the categorisation answers produced by the best controller corresponds to the letter presented in the fovea (white histograms), to the letter presented in the motor copy (grey histograms), or to another letter (black histograms). (a) Average results for each letter and 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)..... 27



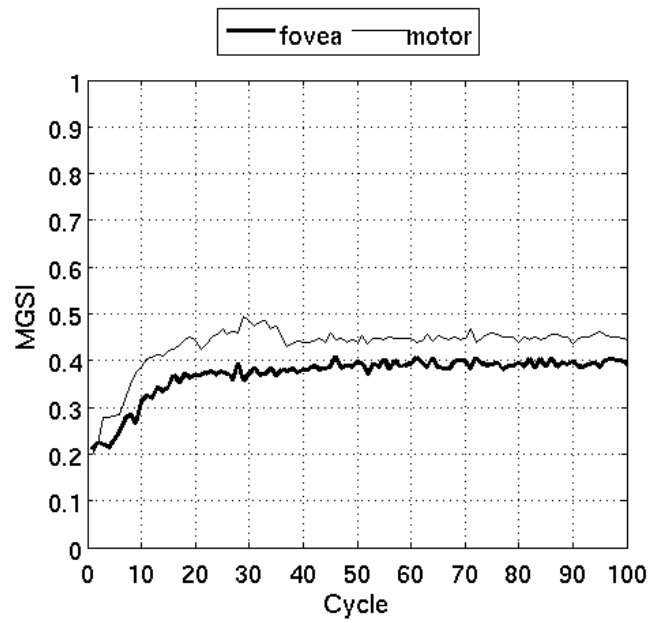
**Fig. 1.** The experimental set-up. (a) Letter ‘l’ shown in the 5 different sizes used in the experiment. (b) The screen displaying the letter ‘l’ in its intermediate size and an exemplification of the field of view of the foveal and peripheral vision (smaller and larger squares, respectively). (c) 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 sensors 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  $t + 1$ .



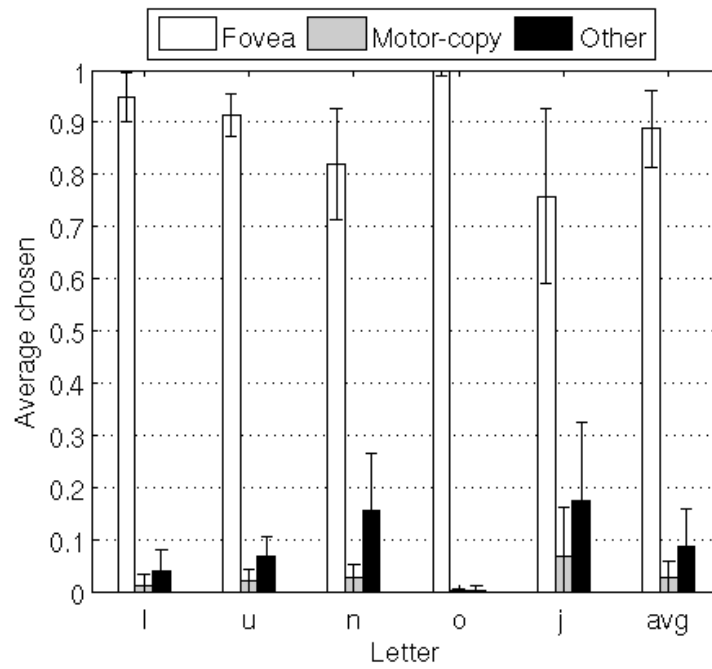
**Fig. 2.** Percentage of correct responses for the best individual of each replication of the experiment. Data obtained by testing each individual for 10000 trials during which it is exposed for 40 times to 5 letters with 50 different dimensions) The lines at the top of each histogram represent the standard error.



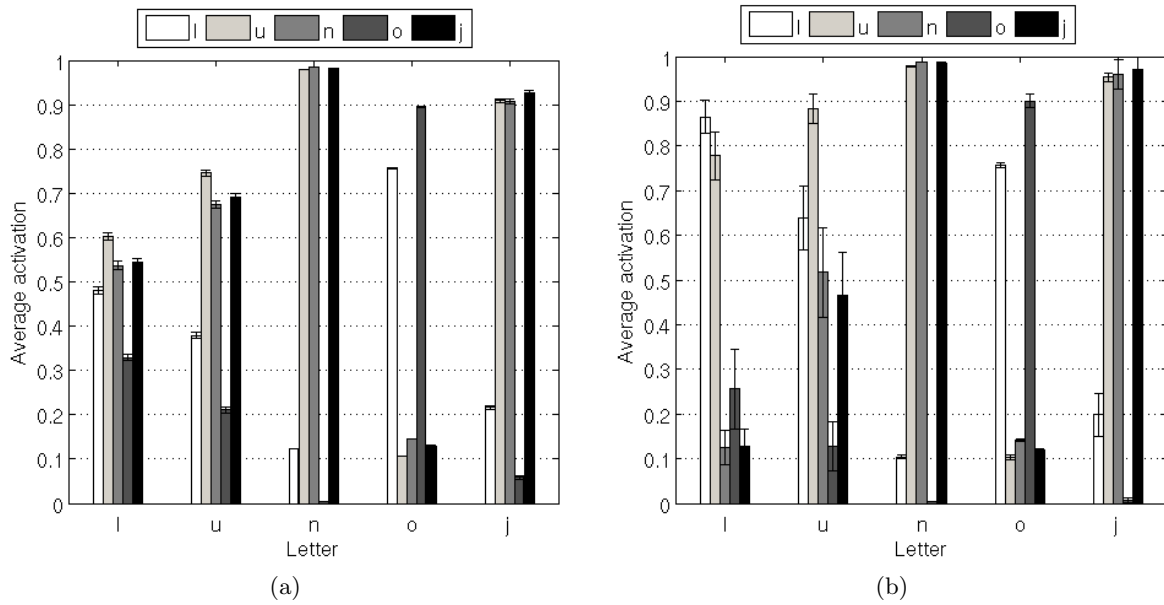
**Fig. 3.** Frequency of observations of different parts of the image for letters of different categories. Each figure plots the data obtained by testing the best evolved individual for 50 trials (10 repetition starting from randomly different initial position for 5 different letter sizes). Gray levels 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.



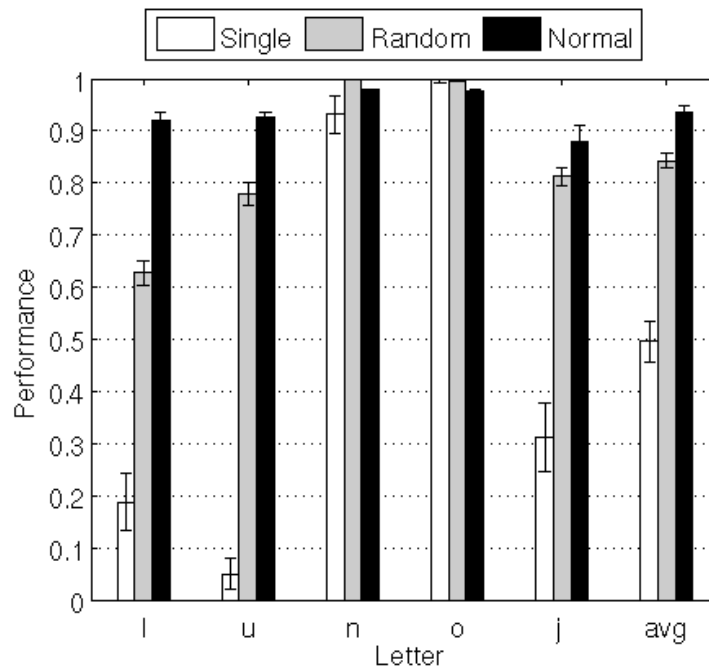
**Fig. 4.** Modified Geometric Separability Index (MGSI) of the stimuli provided by the foveal photoreceptors and by the efference copy of the motors (thick and thin lines, respectively). Each point along the x axis represents the value of the MGSI calculated over the stimuli experienced during the corresponding time step.



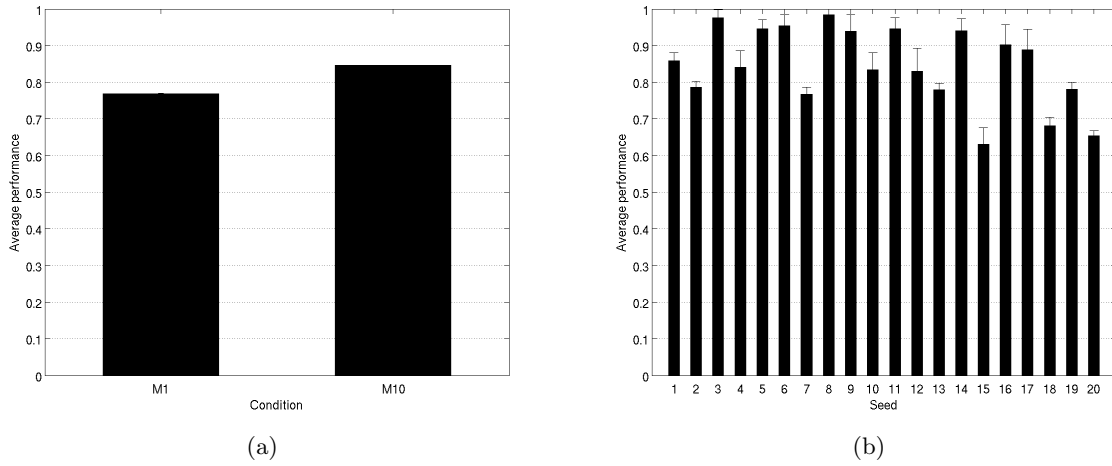
**Fig. 5.** Percentage of times in which the categorisation answers produced by the best controller corresponds to the category of the state of the foveal photoreceptors ('Fovea'), to the category of the state of the motor sensors ('Motor-copy'), or to another category ('Other'). Each group of histogram represents the average performance for each category. Data obtained in a control experiment in which the controller experience pre-recorded sensory states corresponding to all possible combinations of categories over the two sensory channels. Bars represent standard error.



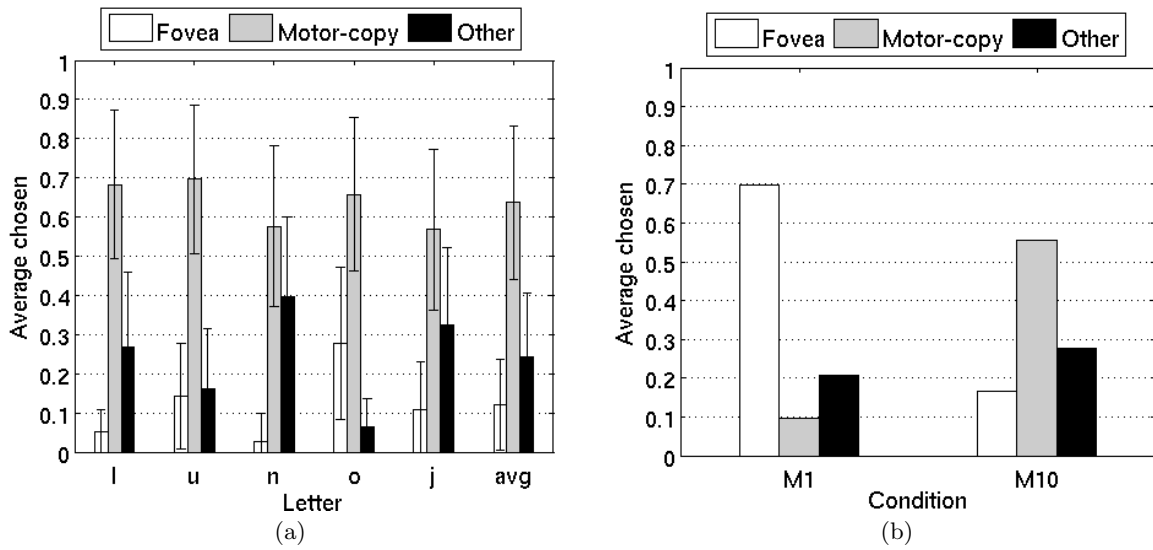
**Fig. 6.** (a) Average categorisation patterns produced in trials in which the agent experienced each possible visual stimulus encountered in natural conditions in a given categorical context for the entire duration of the trial. (b) Average categorisation patterns produced in normal conditions. Each group of histograms represents the average categorisation pattern produced for the corresponding letter indicated in the horizontal axis. Each histograms of a group represents the average activation of the corresponding categorisation output unit (i.e. 'l', 'u', 'n', 'o', and 'j').



**Fig. 7.** 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 let free to autonomously interact with the images).



**Fig. 8.** (a) Comparison between the average performance of the best individuals of all replications in the experiments in which the state of the motor sensors is normalised in  $[0; 1]$  (M1) and in  $[0; 10]$  (M10), respectively. (b) Percentage of correct responses for the best individual of each replication of the second experiment in which the state of the efference copies of the motor neurons is normalised in the range  $[0; 10]$ . Data are obtained by testing each individual for 10000 trials during which it is exposed for 40 times to the 5 letters with 50 different dimensions.



**Fig. 9.** 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).