Evolutionary active vision system

Lanihun, Olalekan; Tiddeman, Bernie; Shaw, Patricia; Tuci, Elio

Published in:
Adaptive Behavior

DOI:
10.1177/1059712319874475

Publication date:
2019

Document Version
Peer reviewed version

Evolutionary Active Vision System: From 2D to 3D

Olalekan Lanihun†, Bernie Tiddeman†, Patricia Shaw† and Elio Tuci‡
† Abeystwyth University, United Kingdom — ‡ University of Namur, Belgium

October 31, 2019

abstract

Biological vision incorporates intelligent cooperation between the sensory and the motor systems, which is facilitated by the development of motor skills that help to shape visual information that is relevant to a specific vision task. In this paper, we seek to explore an approach to active vision inspired by biological systems, which uses limited constraints for motor strategies through progressive adaptation via an evolutionary method. This kind of approach gives freedom to artificial systems in the discovery of eye movement strategies that may be useful to solve a given vision task, but are not known to us. In the experiment sections of this paper, we use this type of evolutionary active vision system for more complex natural images in both 2D and 3D environments. To further improve the results, we experiment with the use of pre-processing the visual input with both Uniform Local Binary Patterns (ULBP, Ojala et al. , 2002) and Histogram of Oriented Gradients (HOG, Dalal and Triggs, 2005) for classification tasks in the 2D and 3D environments. The 3D experiments include application of the active vision system to object categorisation and indoor vs outdoor environment classification. Our experiments are conducted on the iCub humanoid robot simulator platform.

1 Introduction

Active vision is the process of exploring a visual scene to obtain relevant features for subsequent meaningful and intelligent processing. Such visual systems require a form of control, and are intelligently guided to only those areas that have relevant and valuable information to the task at hand. Vision is not a passive process as has been known in conventional computer vision (see Ojala et al. , 2002; Belongie et al. , 2002), but is action dependent (see Avraham and Lindenbaum , 2010; Kagan and Hafed , 2013). In most traditional computer vision, the local image sample does not guide the scanning process, but instead use an exhaustive search (e.g window sliding method, Osuna et al. , 1997). However, research shows that the use of action in perception can reduce the computational cost of vision tasks, and at the same time simplify very difficult tasks (see Nolfi , 1998; Tsotsos , 1992; Mirolli et al. , 2010; Kato and Floreano , 2001). Consequently, as action has been shown to be an integral part of perception, the challenge in developing active vision models is finding intelligent action strategies that will enhance the vision task at hand (Croon , 2008).

In some models the assumption made is that vision is an iterative process of state estimation and the selection of relevant actions (Denzler and Brown , 2002; Borotschnig et al. , 1999). However, in this work we present an active vi-

sion system that has the following properties: (i) it uses limited assumptions or constraints for its action strategy (eye movement); and (ii) it does not need any kind of ground truth for the eye control. This is because such assumptions or ground truth may not allow the model to discover strategies that are not known to the designer but may exist in biological agents. We have therefore chosen an evolutionary adaptive model used in the field of evolutionary robotics for the control of active vision (see Tuci , 2014; Marocco and Floreano , 2002, for similar methods). This technique delegates the strategies used for eye movement to the adaptation process of the evolutionary method. Also, given the strong dependency between eye movements and perception, we also investigated two pre-processing techniques (ULBP, Ojala et al. , 2002) and (HOG, Dalal and Triggs, 2005). We have chosen ULBP and HOG because they are simple to implement as well as their usefulness as feature descriptors in many computer vision applications, such as face recognition (Ahonen et al. , 2006) and object detection (Stefanou and Argyros , 2012). The novelty of our framework is not only in the problems that are solved and methods that are used (pre-pocessing techniques), however, it is the novelty of these problem domains in the investigated active vision systems (evolutionary active systems); and the originality in the combination of existing pre-processing methods (ULBP and HOG) with the active vision system.

Therefore our research objectives are:

(i) To use evolutionary active vision systems in more complex scenes and environments for categorisation tasks.

(ii) To improve the performance of the categorisation tasks through pre-processing techniques.

We further list the contributions in this paper below.

1. This type of active vision system (evolutionary method) is used for more complex images taken from the camera of the iCub robot.

2. We demonstrate the effectiveness of the active vision system in a more realistic setting for 3D object categorisation using the humanoid robot iCub) platform.

3. We extend the applicability of the system to the 3D environment for indoor and outdoor environment classification task using the iCub platform.

4. We extend the system with pre-processing using Uniform Local Binary Patterns (ULBP, Ojala et al. , 2002) in both 2D and 3D environment categorisation tasks.

5. We further extend the system with pre-processing using Histogram of Oriented Gradients (HOG Dalal and Triggs, 2005)) for classification tasks in the 2D and 3D environments.

In the next sections, we first briefly review active vision systems, then describe our system and experimental methods, followed by results, discussion and finally conclusions.

2 Active Vision Models

Various active vision models have been proposed in the literature that select their actions (eye movements) in different ways and mostly for a specific task. For instance, there are models for detecting edges (e.g. Kass et al. , 2008), for controlling the gaze of a simulated fish (e.g. Terzopoulos and Rabie , 1995) and for detecting an object in a visual scene (e.g. Minut and Mahadevan , 2001). However, there are also others that are instances of a more general approach such as the probabilistic approach (see

Vidal-Calleja et al. , 2010; Guerrero et al. , 2010; Dame and Marchand , 2013; Davison , 2005) and adaptive approach (see Mirolli et al. , 2010; Kato and Floreano , 2001; Croon , 2008; Tuci , 2014).

The central aim of the probabilistic models is to reduce uncertainty in the world state. It regards active vision as a series of iterative steps of state estimation and action selection, and therefore uses a pre-determined probabilistic framework for action selection. All the probabilistic models have one thing in common: they take action with the goal of reducing uncertainty in the belief state.

On the other hand, adaptive approaches do not use assumptions or pre-determined framework for their action (eye movement) strategy, but they are progressively adapted in order to optimise the performance of the task at hand. That aside, there are additional predefined attributes which also impose some limitations, such as the choice of the controller (e.g neural network) and the optimisation technique. However, in this model the goal is not to predetermine what the active vision system does internally.

Our approach (evolutionary active vision) is an instance of the adaptive approach that makes use of fewer assumptions for its eye movement, by delegating the matter to the adaptation process of the evolutionary method for neural network control.

2.1 Evolutionary Active Vision System

Various evolutionary active vision systems have been investigated based on the complexity of the controller. For instance, there are those that rely solely on sensory-motor coordination and are also known as reactive systems. These reactive systems are not common in most vision tasks because of their complexity. For example, Nolfi and Marocco (2000) evolved an active vision system in which mobile robots were able to visually discriminate between different landmarks. Similarly, Schembri and Belardinelli (2015) implemented an active vision system using a simple 3-layer feed-forward neural network controller evolved with a genetic algorithm. The goal of the agent was to hit as many small circles as possible and to avoid the big ones over the course of a lifetime. The common features shared by these systems was that, despite their very simple architecture they were able to use their intelligent sensory-motor coordination to select sensory patterns that were favourable to the given vision tasks.

More complex systems have been developed that have a form of memory determined by the recurrent connections or feedback provided in the controllers, that may include hidden layers (see Kato and Floreano , 2001; Marocco and Floreano , 2002). The evolved active vision system described in (Kato and Floreano , 2001) autonomously discriminates between different shapes irrespective of their locations and sizes. The controller of the system has a very simple discrete time recurrent neural network architecture, with no hidden nodes, and was evolved by a genetic algorithm. The system exhibited a behavioural strategy of exploring different areas of the shapes in order to enhance the categorisation task.

In the same vein, Marocco and Floreano (2002) extended the simple active vision model in (Kato and Floreano , 2001) for a navigation problem posed for a mobile robot equipped with a pan and tilt camera. The evolved robots were able to navigate an arena by exhibiting a behaviour where they select simple visual features and maintain the edge between the floor and the wall in sight of the camera. The common theme with these active vision systems is that even though the controllers have very reduced internal states in the form of only recurrent connections or memory feedback, by their dynamic interactions with the environ-

ment, they were able to generate behaviours that allowed them to exploit regularities in ways appropriate to the vision tasks.

There are also active vision systems that have more complex internal states, such as those that are provided by Continuous Time Recurrent Neural Networks (CTRNN, see Mirolli et al. , 2010; Croon , 2008; Lanihun et al. , 2014). In this case, in addition to the recurrent connections, the neurons also have some dynamics that realises internal states. For instance, Mirolli et al. (2010) used an active vision system with a 3-layer CTRNN, which was evolved by a genetic algorithm. The active vision system was given the task of categorising 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. The movement of the artificial eye was controlled by motor neurons of the output units, which determined the eye displacement per time step, in order to capture relevant input features for the neural network controller. The system was rewarded only for its ability to discriminate between the shapes of the letter and left free to determine how to explore the visual scene. Subsequent analysis based on the best individual of all replications of the evolutionary run showed that the agent was able to solve the problem by: (i) using sensory-motor co-ordination to generate behaviours that allowed the agent to experience visual regularities in different categorical contexts; and (ii) integrating perceptual and motor information over time.

By way of further example, Croon (2008) developed an active vision model that uses CTRNNs for a car-driving simulation. Unlike the active vision system in (Mirolli et al. , 2010), the system had a modular structure of two CTRNNs, i.e. one controlling the eye movement and the other for controlling the movement of a simulated car. The output units of the eye controller determined the visual features that were extracted as the car moved through a simulated road per time step, which formed the corresponding inputs to the two controllers. The task of the agent was to drive over a simulated track as quickly as possible, while avoiding various obstacles on the way. The controller parameters were optimised with a genetic algorithm. Subsequent analysis showed that the system used the gaze shifts: (i) to find relevant features that contributed to successful driving; (ii) to keep relevant features in sight; and (iii) to avoid disruptive visual inputs while driving.

The common trend among these systems that used more complex internal states was that the increased complexity helped the system to generate more complex dynamics for integrating sensory-motor information over time.

However, our work is different from the previously mentioned evolutionary approaches in the following respects:

1. We aim to show the plausibility of biological active vision systems in complex artificial systems using our evolutionary method for categorisation tasks. As such, we have extended our method for categorisation to more realistic natural 2D images and to 3D environments using a humanoid robot platform.

2. We investigated two pre-processing techniques commonly used in computer vision, i.e. HOG (Dalal and Triggs, 2005) and ULBP (Ojala et al. , 2002), so as to show how active vision can be enhanced by low level processing (Magnussen , 2000; Le Meur et al. , 2004; Diamant , 2008).

3 Methods

The active vision framework is inspired by the model in (Mirolli et al. , 2010) and is supplemented with the continuous neural network up-

date equations illustrated in (Tuci , 2014). We have built our framework on their periphery-only architecture, Fig. 1, which gave the best performance among all the architectures experimented with in (Mirolli et al. , 2010). We extend this framework for classification in 2D and 3D using the iCub humanoid robot platform (Tsagarakis et al. , 2007), and further used pre-processing to enhance the categorisation tasks.

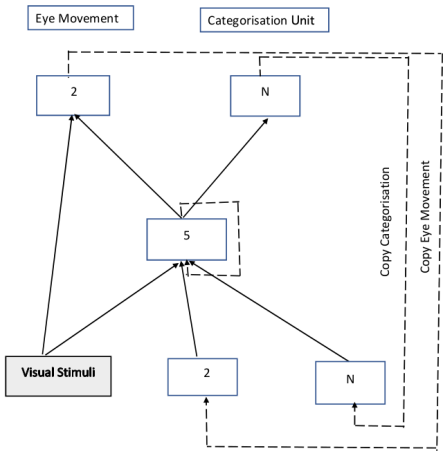


Figure 1: The neural network architecture. The number in the boxes is the number of neurons and N is the number of object categories.

The active vision system autonomously takes an input from a visual scene restricted by the active window. The visual stimuli are processed by a visual extraction method and are mapped by an evolved neural network controller to gaze shifts and classification units. In the output layer, 2 of the neurons determine the movement of the eye per time step either in x and y directions in the 2D experiments, or pan and tilt in the 3D experiments. The other output neurons are for labelling the N possible categories. It also has 5 internal neurons, 2 neurons representing copies of activation values of the gaze shifts and N classification units from the previous time step. The visual extraction module is processed by either a grey-scale averaging method as used in (Mirolli et al. , 2010) or pre-processing techniques such as HOG (Dalal and Triggs, 2005) or ULBP (Ojala et al. , 2002) are adopted. The gaze shifts which enhance the performance of the task are determined by the visual features, previous gaze shifts/categorisation outputs at time $t-1$, and/or the internal state of the controller.

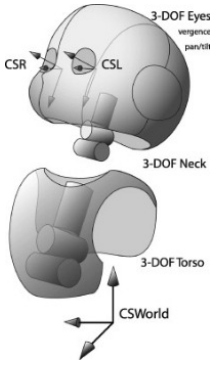


Figure 2: A simple illustration of the iCub vision kinematics (image from (Leitner et al. , 2017)).

3.1 The Robot

We use the iCub humanoid robot platform (Tsagarakis et al. , 2007) for the implementation in the 3D experiments. However, we use a simple iCub simulator described in (Tuci , 2016). This is to minimise the computational overhead that would have been involved in using the original iCub simulator for our evolutionary method. The iCub eye control has 3 degrees of freedom (DOF), for vergence, pan and tilt (Fig. 2). However, because of the computational complexity, we only make use of 2 degrees of freedom for the right eye (pan and tilt). Also for simplicity and because of the computational overhead, we exclude head, neck and other proprioceptive in-

formation from our experiments.

In each time step of every trial in the evolutionary run, we calculate the tilt ($Tilt_{step}$, see Eq. 1) and pan (Pan_{step}, see Eq. 2) by mapping the outputs O_1 and O_2 (with range 0 to 1) to the range $[-2.5°, +2.5°]$ as follows:

$$Tilt_{step} = (O_1 - 0.5) * MAX_{step} \quad (1)$$

$$Pan_{step} = (O_2 - 0.5) * MAX_{step} \quad (2)$$

where $MAX_{step} = 5°$ is the maximum step for the pan and tilt. The pan and tilt are then updated using:

$$Tilt_{new} = Tilt_{new-1} + Tilt_{step} \quad (3)$$

$$Pan_{new} = Pan_{new-1} + Pan_{step} \quad (4)$$

The updated pan and tilt are then normalised back to the range $[0, 1]$, representing the range from the lower and upper limit for each angle, and fed back into the network using:

$$Tilt_{input} = \frac{Tilt_{new} - Tilt_{low\_limit}}{Tilt_{high\_limit} - Tilt_{low\_limit}} \quad (5)$$

$$Pan_{input} = \frac{Pan_{new} - Pan_{low\_limit}}{Pan_{high\_limit} - Pan_{low\_limit}} \quad (6)$$

To map the output pan and tilt onto the iCub we use the Denavit-Hartenberg convention, with the tilt being link 6 and the pan being link 7 in the kinematic chain (i.e. $Tilt_{new} = \theta_6$ and $Pan_{new} = \theta_7$). The full set of link parameters (Table 1) are used to calculate the forward kinematics for the iCub right eye.

Table 1: Table showing the link parameters a, d, α, θ of the iCub right eye (for the tilt i=6 and pan i=7), where a and d are in millimetres, and α and θ are in radians.

Link (i)	a_i (mm)	d_i (mm)	α_i (radian)	θ_i (radian)
i=6	0	34	-pi/2	Θ_6
i=7	0	0	pi/2	Θ_7 - pi/2

3.2 Adaptive Tasks

Here we present the details of the tasks solved by the evolved active vision system. First, we investigate categorising natural images of objects taken from the iCub camera, and evaluate the impact of the pre-processing techniques. Secondly, we move to a 3D iCub simulator and evaluate the system for 3D object categorisation. In 3D object categorisation, the visual field often covers much of the object in a single time-step, making the active behaviour less essential. So in our final experiment we investigate indoor/outdoor categorisation, where scene exploration is essential to achieving the task.

3.2.1 iCub images Categorisation

In this experiment we use the grey-scale averaging for the more complex natural images taken from the camera of the iCub, as compared to artificially generated hand-written italic-letter images used in (Mirolli et al. , 2010). We further tested the proposed feature extraction methods, i.e. ULBP (Ojala et al. , 2002) and HOG (Dalal and Triggs, 2005) to investigate the impact on the performance of the active vision system.

The original images are coloured, and of size 320 × 240 pixels of five different objects, namely: *soft toy*, *TV remote control*, *microphone*, *board wiper*, and *hammer*. The data-set consists of 350 images divided into two folds for training and validation. The first fold of 7 different sizes for each object varying between $[-20\%, 20\%]$ with respect to the original size; and each of these is given 5 different orientations varying between $[-4, 4]$ degrees with respect to the original orientation. The second

fold also of 7 different sizes varying between $[-30\%, 30\%]$ of the original size; and each of these rotated by 5 different orientations varying between $[-9, 9]$ degrees with respect to the original orientation. We used a larger range of scale and orientation in the second fold so as to make the categorisation task more challenging.

The original coloured images are first converted into grey-scale. We then evaluate the agent for 350 trials, and at the beginning of each trial: (i) one of the 175 images (in a fold) is presented to each individual (i.e. network weights set according to the genes); (ii) the state of the internal neurons of the agent's controller is initialised to 0.0; and (iii) the eye is initialised in a random position within the central third of the image. During the 100 time steps of each trial, the agent is left free to explore the image.

Also, in order to terminate trials when the active window of $(50 \times 50$ pixels) no longer includes any part of the object for three consecutive time steps, we use a Canny Edge Detector (Canny , 1986) to detect the edges in each image presented. A rectangular mask is set on the object in the image, and every white (edge) pixel outside the boundary of the rectangular mask are set to black. Through this means, we are able to get edge images that are black outside object boundaries, and objects of white and black. Fig. 3 shows the grey-images, and the images after setting the rectangular masks on the Canny Edge Detector processed images. It should be noted that the above processing is only used to terminate each trial after the active window has lost focus on the object for more than 3 consecutive time steps and as a result time is saved during training. The input vector into the neural network is obtained from the grey-images processed by the visual extraction methods (grey-scale, ULBP or HOG), and the copies of the movement and categorisation units at previous time step $t-1$.

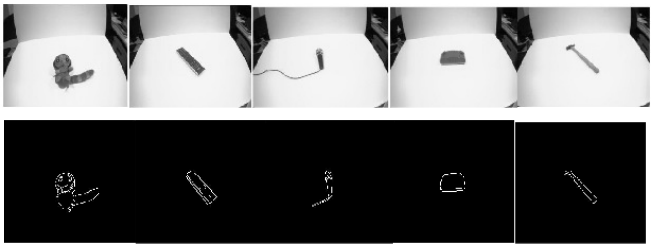


Figure 3: Pictures showing the images of the objects, from left to right: soft toy, TV remote control, microphone, board wiper and hammer. Top row, the greyscale images. Bottom row, the images after processing with Canny edge detection and a masking rectangle.

3.2.2 3D Object Categorisation

This experiment is designed to investigate how a simulated agent (the iCub) can exploit its eye movement to improve object categorisation and how this categorisation capability can be further improved with pre-processing techniques. The agent is situated in a 3D environment in-front of a coloured object on a coloured table against a black background (Fig. 4).

We chose four objects; a sphere, cube, cone and torus, in which the stimuli have similar appearance, rendering the categorisation task more challenging. The four different coloured objects are presented to the agent for categorisation one at a time. (Fig. 4).

The agent is evaluated for 48 trials in which each of the four objects (sphere, cube, cone and torus) is presented to the iCub agent 12 times; and each trial lasted 100 time steps.

At the beginning of each trial: (i) each object is uniformly randomly scaled with a variation of [-10%,10%] to the original size, and uniformly randomly rotated within the range $[-10°, 10°]$ on the y axis; (ii) the state of the internal neurons of the agent's controller is initialised to 0.0; and (iii) the eye is initialised in each quadrant of the iCub gaze-space, but randomly located in each initialisation within

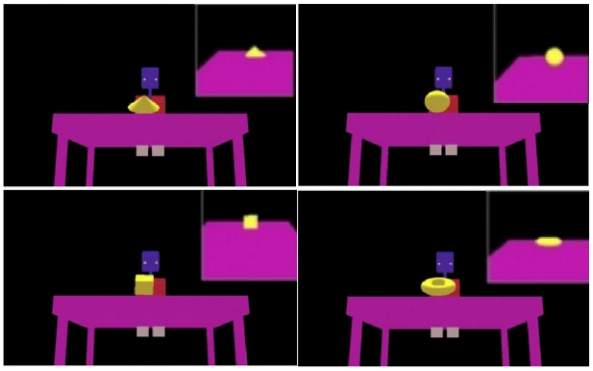


Figure 4: Pictures showing the iCub agent presented with the four objects: **top left**, cone; **top right**, sphere; **bottom left** cube and **bottom right** torus. On the top right of the scene shows the objects from the iCub view.

a quadrant, and with the object within the eye view. During each time step of a trial, we calculate the pan_{step} and $tilt_{step}$ and normalise their updates and input as proprioceptive feedback (pan_{input}, and $tilt_{input}$) along with the categorisation outputs at the previous time step into the network. In each trial the eye is left to freely explore the environment; however, in order to save time and improve exploration, a trial is terminated when the eye (pan or tilt) reached the iCub pan limit ([-0.523616, 0.523616] radians) or tilt limit ([-0.663243, 0.314177] radians) for three consecutive time steps. In each trial, the agent's eye perceives each object presented with visual extraction from grey-scale averaging (Mirolli et al. , 2010), ULBP (Ojala et al. , 2002) or HOG (Dalal and Triggs, 2005).

3.2.3 Indoor-Outdoor Environment Categorisation

In this experiment, the agent is situated in various 3D indoor and outdoor environments.

The environments are represented with 20 texture images, which were downloaded from

Figure 5: Pictures showing the iCub agent in the outdoor environment (left) and outdoor environment (right). Top right shows the environments from the iCub view.

Google's image database (Google Images , 2017). The texture images are dynamically mapped to the interior of a 3D sphere containing the iCub (Fig. 5). Half of the images represented indoor environments and the other half were outdoor environments. The entire dataset of 20 texture images representing the environments are divided into 2-equal halves for training and validation sets for a 2-fold cross-validation. The rotation of the environment ensures that the agent is always seeing different part of the environment in any given trial. The visual information perceive with the retina is processed with one of the visual extraction methods, i.e. grey-scale averaging, ULBP or HOG.

The agent is evaluated for 20 trials, and at the beginning of each trial: (i) the agent is situated in an environment (outdoor or indoor) that is randomly rotated within the range $[-40°, 40°]$ on the z axis with a uniform distribution, and subsequently, the agent uses its pan and tilt movement to explore the environment in each time step.; (ii) the states of the internal neurons of the agent's controller are initialised to 0.0; and (iii) the eye is initialised in each quadrant of the iCub gaze-space, although randomly located in each initialisation within a quadrant. Also, in each time step of a trial, the pan_{step} and $tilt_{step}$ values are calculated and their normalised updates are input as

(pan_{input}, and $tilt_{input}$) as proprioceptive feedback along with the categorisation outputs at previous time step into the network. In each trial, the eye is left to freely explore the environment; however, in order to save time and improve exploration, a trial is terminated when the eye (pan or tilt) reached the iCub pan limit ([-0.523616, 0.523616] radians) or tilt limit ([-0.663243, 0.314177] radians) for three consecutive time steps.

3.3 Neural Network Controller

The gaze control model is a 3-layer continuous neural network architecture: (i) an input layer, whose vector size is determined by the visual feature extraction method, and a copy of the motor/gaze control units and classification units at the previous time step; (ii) recurrent hidden layer units; and (iii) an output layer of motor/gaze control units and classification units. The activations of the input neurons are normalised between 0 and 1, however with 0 representing a fully white visual field, while 1 represents fully black for the 2D grey-scale (as it was done in (Mirolli et al. , 2010)). A random value with a uniform distribution within the range of [-0.05,0.05] is added to the input activation values in each time step, in order to take into account that sensor data are subject to noise.

The values of the input, hidden, and output neurons are updated using equations 7, 8 and 9 respectively:

$$y_i = gI_i; i = 1, ..., n-1 \tag{7}$$

$$\tau_i \dot{y_i} = -y_i + \sum_{j=1}^{j=k-1} w_{ji}\sigma(y_j + \beta_j); i = n, ..., k-1 \tag{8}$$

$$y_i = \sum_{j=n}^{j=k-1} w_{ji}\sigma(y_j + \beta_j); i = k, ..., u \tag{9}$$

In these equations, using terms derived from an analogy with real neurons, y_i represents the cell potential, g is a gain factor, τ_i the decay constant. I_i (with $i = 1, ..., n-1$) is the activation of the i^{th} input neuron. Neurons $n, ..., k-1$ and $k, ..., u$ are the hidden and output neurons respectively. w_{ji} is the weight of the synaptic connection from pre-synaptic neuron j to post-synaptic neuron i. β_j is the bias term and $\sigma(y_j + \beta_j)$ is the firing rate, where $\sigma(x)$ is the sigmoid function. All input neurons share the same bias β^I, and the same holds for all output neurons β^O. The decay constants, bias terms, weights and gain factor are genetically specified network parameters. We approximated the dynamics of differential equation 8 using the standard forward Euler method with an integration time step $\Delta T = 0.1$. In the next section, we discuss the three methods that are used for processing visual stimuli inputs into the neural network controller.

3.3.1 Visual Feature Extraction

The following methods are used to extract the visual stimuli for the neural network controller in all the experiments in this paper: the grey-scale averaging method (Mirolli et al. , 2010), ULBP (Ojala et al. , 2002) and HOG (Dalal and Triggs, 2005).

We allowed the active vision to dynamically select an area to be processed per time step and then used one of the visual extraction methods to process the pixels within the active window. As such, we still keep to our philosophy of an active vision model that does not process the entire image but instead allows the system to actively select features through the dynamic interaction of sensory-motor components (Croon , 2008).

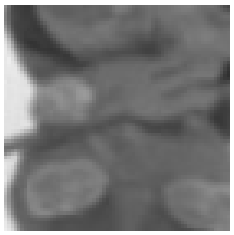


Figure 6: Original active window area of soft-toy object grey-image.

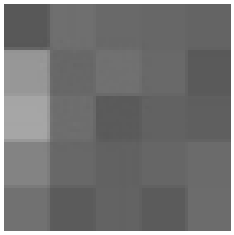


Figure 7: The active window area after grey-scale averaging method was applied.

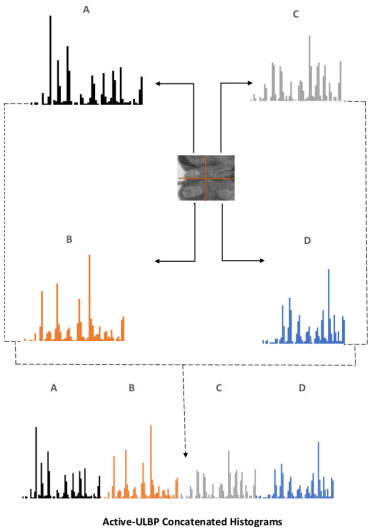


Figure 8: Active-ULBP histograms of the cells of the active window, and the concatenated histograms.

Grey-scale averaging

In the grey-scale averaging method, the coloured image is first converted to a grey-scale image.

The active vision model then takes visual input from a window of $s \times s$ pixels extracted from the grey-image of $m \times m$ in each time step. The window is sub-divided into $k \times k$ input cells and the average value calculated in each cell, resulting in k^2 visual inputs.

Fig. 6 shows an example active window grey-scale image patch (i.e. soft toy image) and Fig. 7 shows the average pixels of the active-window that were input into the neural network.

Active Uniform Local Binary Patterns method

ULBP is an extension of Local Binary Patterns (LBP Ojala et al. , 1996) that considers only uniform patterns. The basic LBP approach considers the 8 neighbours of each pixel in a fixed rotational order and assigns `0` or `1` to a bit string if the central pixel intensity is larger or smaller than its neighbour. This produces an 8-bit unsigned integer for each pixel, and histograms of these values over different regions have proved effective in various computer vision tasks. Many extensions of the basic LBP approach have been considered, here we use uniform LBPs. Uniform patterns of texture units are those that have a maximum of 2 bit-wise transitions, i.e. from `0` to `1`. For instance, in an eight-circle neighbourhood texture unit, bit patterns `00000000` (0 transition), `00110000` (2 transitions) are uniform patterns, while non-uniform patterns such as `00010100` (4 transitions) and `00101010` (6 transitions) are not. In ULBP, there is a separate output label for each uniform pattern and one output label for all the non-uniform patterns. Thus, the number of output labels for the mapping of patterns P is $P(P-1)+3$. For instance, ULBP produces 59 output labels for an eight-neighbourhood texture unit and 243 for 16 circular neighbourhood sampling points (Pietikainen et al. , 2011; Tapia et al. , 2014).

However, because of the peculiar nature of active vision systems and the computational cost of evolutionary methods in training, we have implemented the ULBP method so that it will be suitable for the model. For instance,

all forms of pre-processing have to be done within the active window (retina region) per time step, instead of processing the entire image. We also have to use a considerably reduced number of cells (4 cells). We therefore prefer to term it Active-Uniform Local Binary Patterns (Active-ULBP), because of its adoption to the Active Vision System. The Active-ULBP algorithm is implemented as follows:

1. An image is presented to the active vision model in each trial of the evolutionary run.

2. In each time step of a trial:

   (a) a Gaussian blur function is used to reduce the noise within the active window (retina region); (b) the retina region is divided into 4 cells and a histogram of uniform patterns of size 59 is constructed for each cell; (c) the histogram of each cell is normalised with an $L2-norm$ scheme; (d) the normalised histograms of all cells are concatenated to form a feature vector of size 236; (e) the feature vector is combined with the copies of the movement and categorisation output units at the previous time step which formed the input vector for the neural network.

Fig. 8 shows the histograms and the concatenated histograms of the 4-cells of the active-window of a patch of the soft-toy image for the Active-ULBP method.

Active Histogram of Oriented Gradients method

The HOG descriptor was originally developed by Dalal and Triggs (2005) for describing edges and gradients over a local image region using a sliding window over an entire image. It computes histograms over dense grids of uniformly spaced cells and normalises contrast for improved performance. In their work Dalal and

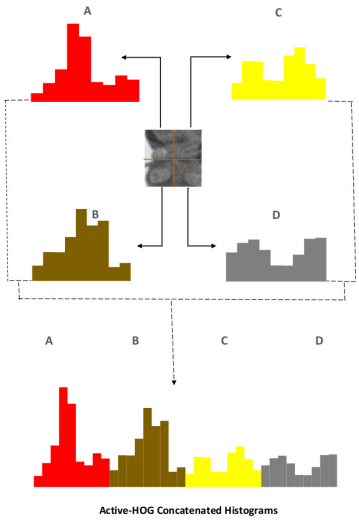


Figure 9: Active-HOG histograms of the active window image patch and the concatenated histograms.

Triggs (2005) used HOG as a feature descriptor for pedestrian recognition data and used a Linear Support Vector Machine as the classifier for the normalised histogram features.

The fundamental idea is that object appearance and shape over a local region can be characterised very well with intensity gradients distribution. The image window is divided into small spatial cells over dense grids. Histograms are computed for the cells and contrast normalised to form the feature sets.

However, in the adoption of HOG in our model we considered two major factors: (i) the computational complexity of the pre-processing, since evolving a neural network will only be practicable with lower dimensional feature vectors; and (ii) suitability for the active vision concept, which processes a part of the image scene at each time step. Consequently, the HOG used in our model is a very simple version of the original algorithm and we prefer to call it Active-Histogram of Oriented Gradients (Active-HOG) because of its adoption to

the Active Vision System. We list the complete steps of the Active-HOG algorithm applied each time step below:

1. compute the gradients for each pixel in the active window in x and y direction i.e dx and dy
2. divide the active window into 2×2 cells giving a total of 4 cells;
3. in each cell compute gradient magnitudes as $\sqrt{dy^2 + dx^2}$ and gradient directions as $\Theta = arctan(\frac{dy}{dx})$;
4. quantize gradient orientations into 9 bins with a bin size of 40 degrees of orientation space between $0 - 360$ degrees;
5. add magnitude into each bin;
6. concatenate all histograms into a feature descriptor of dimension 4 cells $\times$ 9 bins giving a feature vector of size 36;
7. normalise the feature vector with $L2 - norm$, i.e. $V = \frac{V}{||V||}$;
8. input a normalised feature vector into the neural network along with the copies of motor and categorisation outputs from the previous time step.

Fig. 9 shows the concatenated histograms of a patch of the soft toy image of the Active-HOG method.

3.4 Evolutionary Algorithm

The free parameters of the agent's neural controller are adapted through an evolutionary algorithm using roulette wheel selection scheme (see Goldberg , 1999). The initial population for each generation of the evolutionary process consists of 100 or 60 randomly-generated genotypes (for 2D and 3D experiments respectively); sampled from a uniform distribution in the range $[0, 1]$, each encoding the free parameters of the corresponding neural controller, which includes all the connection weights, gain factors, biases, and the time constants of the hidden neurons. In order to generate the phenotypes, weights and biases are linearly mapped in the range $[-10, 10]$ and $[-5, 5]$ respectively, while the time constants are exponentially mapped into $[10^{-1}, 10^{1.8}]$ for the 2D and into $[10^{-1}, 10^{2.2}]$, for the 3D experiments, with the lower bounds corresponding to the integration step-size used to update the controller. Generations following the first are produced by a combination of selection with elitism, recombination and mutation (Goldberg , 1999). For each new generation, the genotype with the highest fitness value ("the elite") from the previous generation is retained unchanged. The remaining 99 and 59 genotypes of the new generation for both 2D and 3D tests are formed by randomly selecting two genotypes from the older generation from the best 70 and 50 genotypes using roulette wheel selection, and a new genotype is created by combining the genetic material of these two old genotypes with a probability of 0.3 with crossover point selected during the recombination. Mutation which entails that a random Gaussian offset is applied to each real-valued component encoded in the genotype is done with the probability of 0.05 in the 2D and 0.04 in the 3D. The mean is 0 and its standard deviation is 0.1.

3.5 Fitness Function

In each trial of the evolutionary adaptation process, the artificial eye (active window) is left to freely explore the visual scene in the first part of the trial. The task of the active vision agent is to correctly classify an object when it has explored the image for a sufficient length of time, that is during the second half of a trial.

The agent is evaluated by the by the fitness function F as used in (Mirolli et al. , 2010), and is comprised of two components: the first, $F_1(t,c)$ rewards the agent's ability to rank the correct category higher than the other categories; the second, $F_2(t,c)$ rewards the ability to maximise the activation of the correct unit while minimising the activations of the wrong units, with both terms given equal weighting:

$$F = \frac{\sum_{t=1}^{T}\sum_{c=S}^{C}\left(0.5 * F_1(t,c) + 0.5 * F_2(t,c)\right)}{T * (C - S)} \tag{10}$$

$$F_1(t,c) = 2^{-rank(t,c)} \tag{11}$$

$$F_2(t,c) = 0.5 * y_r^{t,c} + \sum_{w \in W}\left(1 - y_w^{t,c}\right) * \frac{0.5}{N-1} \tag{12}$$

where $F_1(t,c)$ and $F_2(t,c)$ are the values of the two fitness components at time step c of trial t, $rank(t,c)$ is the ranking of the activation of the categorisation corresponding to the correct category (that is, from 0, meaning the most activated and l, meaning the least activated: where l is 1 less than number of categories), $y_r^{t,c}$ is the activation of the output corresponding to the current (correct) category, $y_w^{t,c}$ is the activation output of the wrong category w at trial t and time step c (where W is the set of wrong categories). N is the number of categories, T is the number of trials, C is the number of time steps in a trial and S is the time step in which we start to compute fitness.

3.6 Implementation

Due to the high cost of evolving the main parameters of the neural network and evaluating each phenotype for multiple trials and timesteps, we employ a parallel computing cluster. We use High-Performance Computing Wales infrastructure (Super Computing Wales, see SCW , 2019)). The Message Passing Interface (MPI) (Gropp et al. , 1999) is used to parallelise the implementation using a root and individual sub-processes. Each individual runs its evaluation as a separate process and the respective fitness is communicated to the root process, which in turn carries out the evolution and subsequent generation of a new set of controllers.

4 Results

In this section, we present the results and analysis for all the experiments in the 2D and the 3D environments (i.e. object and indoor-outdoor categorisation). First, we present the fitness graphs for all runs and the best run for each visual extraction method. Second, we show the results of the re-evaluation tests. In all experiments, categorisation performance is based on the percentage of times in which the categorisation unit corresponding to the correct category is the most activated in all trials of the re-evaluation.

For each experiment we statistically compare the 3 extraction methods.

To compare the three techniques, we apply ANOVA tests, with a p-value$<$0.05 and a more stringent p-value$<$0.01. Where a significant difference is detected, further pairwise evaluation of the three possible pairs is performed using t-tests. Bonferroni correction is used in the t-tests to account for the greater chance of a significant result occurring by chance among the three methods.

To further evaluate the dynamic behaviour of the systems, we use the Modified Geometric Seperability Index (MGSI), for a quantitative behavioural analysis. The Geometric Separability Index (GSI) was originally proposed by Thornton (1998), while the MGSI is a mod-

ified version of the GSI and was proposed by Mirolli et al. (2010). The GSI computes the percentage rate at which the nearest pattern of each experienced pattern belonged to the same category; however the MGSI is more demanding in that it takes into account not only the nearest neighbour but all the stimuli belonging to the same category. We chose to use this more demanding measure because the nature of our problem is very similar to that of (Mirolli et al. , 2010). The MGSI is defined by the equation below:

$$MGSI(P) = \frac{\sum_{s \in P} \frac{\sum_{n \in N_s} I_{C_s}(n)}{|C_s|}}{|P|}$$

Which is defined as the average proportion of patterns belonging to the same category, that are in the $|C_s|$ nearest patterns (computed from Euclidean distance), where $|C_s|$ represents the total number of patterns in the same category as pattern s. Where P is the set comprising all the patterns, $|P|$ is the cardinality of the set P, C_s is the set of all patterns belonging to the same category as pattern s (s does not belong to C_s), N_s is the set of the $|C_s|$ patterns nearest to pattern s, and $I_{C_s}(n)$ is the indicator function of set C_s, that returns 1 if n is in set C_s and 0 otherwise. If the system is intelligently guiding the visual system we would hope to see an increase in the MGSI over time during the evaluation, indicating that it is moving to locations that improve the discrimination ability (Ferrauto et al. , 2009; Tuci et al. , 2010).

4.1 2D iCub Images

This section presents the results of the three methods of visual representation for active vision on 2D images from the iCub camera. In the 2D experiment, we performed 20 evolutionary runs, with 10 runs for each fold of the 2-fold cross validation and each evolutionary run lasted 3000 generations. We then re-evaluated the best genotypes of the last 1000 generations of the evolutionary runs for the categorisation task for the three methods of visual extraction. This is simply because the last 1000 generations should have a relatively higher fitness pattern than the other generations and as a result yield better performance in the categorisation tasks.

The best performing genotype in each run for all the evolutionary runs are presented and used in our statistical analysis.

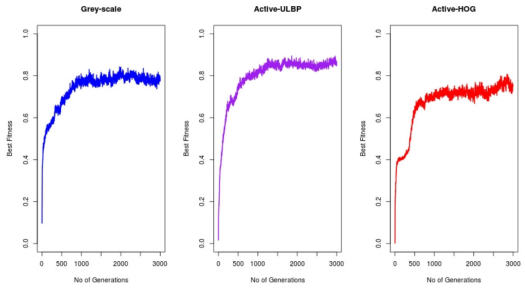


Figure 10: **2D iCub images:** The fitness graphs of the best evolutionary runs of the three visual extraction methods in the 2-fold cross-validation.

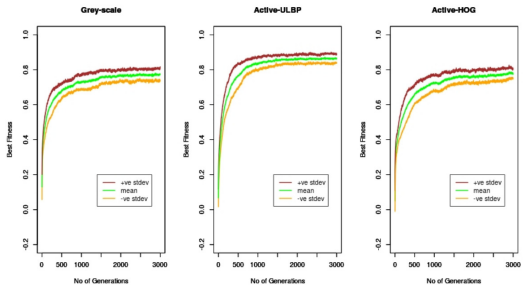


Figure 11: **2D iCub images:** Shows the graph of the mean (average) of all fitness in each generation of the 3000 generations for 20 evolutionary runs and their positive (+ve stdev) and negative (-ve stdev) standard deviation in each generation for the three methods of visual extraction.

Figure 10 (for the best runs) shows that fitness patterns for the three methods improve over all generations, while the grey-scale and Active-ULBP show greater improvement than the Active-HOG.

Also, observing standard deviation from the mean for the three methods in (Figure 11), one can deduced that all three methods produce a best fitness that is very close to the mean in the first few generations; however larger deviations are observed in the remaining generations. Moreover, Active-HOG exhibits a larger deviation from the mean at an earlier stage than the other two methods. Overall, the fitness patterns of all runs seems to be closer to the mean for the Active-ULBP than for the other two methods, especially from approximately 700 generations onwards. By contrast, the fitness patterns for the grey-scale and Active-HOG methods are very similar. This suggests that the fitness patterns for all runs of the Active-ULBP in general seem to progressively improve in all generations as compared to the other two visual extraction methods.

In the re-evaluation, a total of 700 trials were done, with each image of each fold (175 images) presented 4 times to the agent with a random initial eye position in each trial.

The results of revaluation are presented in Table 2 for the best performing genotypes (20 genotypes) in all evolutionary runs (20 runs)for the three visual extraction methods.

The ANOVA shows a significant difference between the three methods (p=0.0106), Table 3. Further evaluation with Bonferroni corrected t-tests shows a significant difference between the Active-ULBP and the Active-HOG, but not between the grey-scale and either of the other two methods (Table 4). The good performance of Active-ULBP, though not significantly better than the grey-scale, may lend support to ULBP as an effective feature descriptor for texture information. The performance of Active-HOG also shows that it can be an effective feature representation for images characterised by some level of structural information.

Also, we computed the MGSI of the best performing re-evaluated evolved genotypes (3 genotypes) for all three visual extraction methods for 1750 trials during which the agent experiences the five different categories (i.e. soft toy, remote control set, microphone, board wiper and hammer) of the 35 different samples for each category, 10 times each with different initial eye positions. For each type of visual extraction method of the sensory patterns the MGSI has been calculated for each of the 100 time steps of a trial (Figure 12).

The results show that the MGSI increase for all visual extraction methods and for all objects, i.e. the system moves away from very ambiguous to less ambiguous stimuli. The MGSI never reached a value of 1, so the system never managed to discover completely unambiguous stimuli for any of the visual extraction methods. The Active-ULBP method generates less ambiguous stimuli (i.e. the highest peak in the MGSI graph) than the grey-

Table 2: **2D iCub images:** The statistics of the best performing re-evaluated genotypes (20 genotypes) in all runs for each visual extraction methods.

Visual extraction methods	Max	Average	Worst	Stdev
Grey-scale averaging	99.65	95.77	87.26	±4.13
Active-ULBP	99.77	96.82	91.75	±2.49
Active-HOG	98.16	92.87	77.81	±5.26

Table 3: **2D iCub images:** The results of the ANOVA test.

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	167.29	2	83.65	4.93	0.0106	3.16
Within Groups	967.33	57	16.97			
Total	1134.62	59				

scale and Active-HOG methods, however grey-scale is more consistent. For the Active-HOG, the system did not exhibit as great a tendency to move towards less ambiguous stimuli when compared with the other two methods. For some objects, the system managed to generate less ambiguous patterns than for other objects. This means the system produced more discriminative patterns for those objects than for others. On the whole in the 2D experiments, the MGSI results indicates that system used some form of eye movements in improving the categorisation tasks.

4.2 3D Object Categorisation

We performed 6 evolutionary runs for each of the visual extraction methods, and each run was for 5000 generations.

However, in this experiment, we assessed the performance of the system using the best evolved genotypes of 100 consecutive generations that had a relatively higher and more stable fitness pattern as compared to the other generations in all evolutionary runs. This differs from the 2D experiment, where we took a more systematic approach by re-evaluating the best genotypes of the last 1000 generations. The number of genotypes chosen for re-evaluation has been reduced in order to keep the re-evaluation time within reasonable limits, considering the high computational costs of the 3D experiments.

Figure 13 shows that the fitness pattern of the three methods of visual extraction generally goes up in all generations of the best evolutionary runs for the three methods, with that of Active-HOG showing more improvement than the other two.

Also, comparing the pattern of fitness of all runs of the three visual extraction methods (Figure 14), one can observe that the mean fitness pattern of the Active-HOG is generally higher than that of the other two methods in all generations of the evolutionary runs, while that of the grey-scale is a bit higher than that of the Active-ULBP. This suggest that Active-HOG fitness values over all generations in all evolutionary runs are generally higher than those of the other two methods.

In the re-evaluation, the system was tested on the four categories of object used in the training by randomly scaling and rotating each object presented in a trial. The objects were randomly scaled within the range [-15%, 15%] relative to their original size and rotated in the range $[-10°, 10°]$ on the y axis, with a uniform distribution. A total of 200 trials were performed, with each object presented 50 times to the agent in all trials and the eye was initialised in each quadrant of the iCub gaze space.

The re-evaluated best 100 genotypes, shows that the best performance was by the Active-HOG method, followed by the grey-scale method and then the Active-ULBP (Table 5). Statistical evaluation of these results using ANOVA (Table 6) shows a highly significant difference between the results ($p = 0.0004$). Further investigation using pairwise, Bonferroni corrected t-tests (Table 7) shows a signif-

Table 4: **2D iCub images:** The significant test results using a paired t-test with test conditions of (p-value<0.05) and (p-value<0.01).

Compared Groups	*Mean Difference*	*t-value*	*p-value*	*Signf. Level=0.05* *Bonf. Corr=0.0167*	*Signf. Level=0.01* *Bonf. Corr=0.003*
Active-ULBP and Greyscale	1.05	0.81	0.2862	Not Significant	Not Significant
Active-ULBP and Active-HOG	3.95	3.03	0.0052	Significant	Not Significant
Greyscale and Active-HOG	2.9	2.23	0.0354	Not Significant	Not Significant

Table 5: **3D object categorisation:** The statistics of the best performing re-evaluated genotypes (6 genotypes) in all runs for each visual extraction methods.

Visual extraction methods	Max	Average	Worst	Stdev
Grey-scale averaging	93.76	74.47	66.19	±12.01
Active-ULBP	88.03	68.53	49.36	±13.32
Active-HOG	99.48	98.07	95.08	±1.9

icantly better performance of the Active-HOG method than both of the other 2 methods.

The fact that Active-HOG performed better than grey-scale and Active-ULBP in the 3D object classification scenario may be due to the more structural nature of object categorisation problem. This boosts the credentials of HOG as an effective feature descriptor for applications that involve structures, e.g. object detection (Zaytseva et al. , 2012) and human recognition (Dalal and Triggs, 2005). The fact that Active-ULBP also demonstrated good performance, though not as good as the ggrey-scale, may also provides further evidence of ULBP as an effective feature descriptor in many applications (Pietikainen et al. , 2011; Tapia et al. , 2014).

The MGSI has been computed for the best performing re-evaluated evolved genotypes (3 genotypes) for the three visual extraction methods in all evolutionary runs. This was done for 200 trials during which the agent experienced the stimuli from the four categories (i.e. sphere, cube, cone, and torus), where each object was uniformly and randomly scaled between [10%, -10%] to the original size and rotated within the range $[-10°, 10°]$ relative to the original orientation with 50 different initial eye positions. For each type of visual extraction method using the sensory patterns, the MGSI was computed for each of the 100 time steps of a trial (Fig. 15).

The MGSI increased for grey-scale, showing that the system moved away from very ambiguous to more discriminative stimuli but showed only modest improvement for the Active-ULBP. The Active-HOG generated less ambiguous stimuli than grey-scale and Active-ULBP but generally did not show improvement over time and even deteriorated in the case of cone object, and also exhibited oscillatory behaviour in most time steps for all the objects. This might have been due to the reduced ambiguity provided by the Active-HOG stimuli from the start, and, as such, there was not much need in this case to use the eye movements to reduce ambiguity. On the whole in the 3D object classification experiments, the active vision system showed some form of movements in improving the discriminative tasks, but Active-HOG seems not to evolve movement strategies because it did very well in the early stages of the evolution.

4.3 3D Environment Categorisation

We performed 12 evolutionary runs, i.e. 6 runs for each of the 2-fold cross-validation, and each run lasted for 5000 generations. However, in the re-evaluation, we assessed the performance of the system using the best evolved genotypes of 100 consecutive generations that had a relatively higher and more stable fitness pattern as compared to the other generations in all evolutionary runs as was mentioned in the previous section for the 3D object categorisation.

The fitness graphs of the best evolutionary runs (Figure 16) and all runs (Figure 17) for all three visual extraction methods shows that they exhibit a common fitness pattern in which fitness growth reached close to the op-

Table 6: **3D object categorisation:**The results of the anova test

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2930.97	2	1465.49	13.51	0.0004	3.68
Within Groups	1626.59	15	108.44			
Total	4557.56	17				

Table 7: **3D object categorisation:** The significant test results using a paired t-test with test condition of (p-value<0.05) and (p-value<0.01)

Compared Groups	*Mean Difference*	*t-value*	*p-value*	*Signf. Level=0.05* *Bonf. Corr=0.0167*	*Signf. Level=0.01* *Bonf. Corr=0.003*
Active-HOG and Greyscale	23.6	3.93	0.0014	Significant	Significant
Active-HOG and Active-ULBP	29.54	4.91	0.0002	Significant	Significant
Greyscale and Active-ULBP	5.94	0.99	0.2371	Not Significant	Not Significant

tima value of 1.0 in the early stage of the evolutionary run from about generation 1000.

The early convergence to optimal solutions of the three visual extraction methods as reflected in the training may be due to the system formulating easy solutions to the problem because of the small number of images used and trials that were done in order to reduce the time complexity of the evolutionary method. Therefore, the importance of the re-evaluation is to test the robustness of the model by introducing more variability into the system. For examples, changing the initial position of the eye in each trial, rotations of the environment/stimuli and increasing the number of trials. This is not possible in the evolutionary runs because of computational cost. For this reason, the complexity of the problem was in the generalisation of the skills learned by the evolved genotypes to unseen images coupled with the additional variability introduced in the re-evaluation.

We re-evaluated the 100 best evolved genotypes of the three visual extraction methods of all evolutionary runs for 200 trials during which the agent experienced 10 different indoor and outdoor environments (2 classes) in 2-fold cross-validation (20 images), with each environment uniformly and randomly rotated within the range $[-40°, 40°]$ to the original orientation with 20 different initial eye positions. We present the result of the best performing genotypes (12 genotypes) from the 12 evolutionary runs (Table 8). As shown in the table, even though reasonable results were obtained in the re-evaluation stage, the lowered performance did not reflect the early optima fitness reached by the system (grey-scale, Active-ULBP and Active-HOG) in the evolutionary (training) stage. Active-HOG shows the best performance, followed by Active-ULBP and then the grey-scale method. The ANOVA shows a highly significant difference between these means (Table 9) and further investigation using Bonferroni corrected t-tests shows a significant difference between the Active-HOG and grey-scale, but not between the other method combinations (Table 10).

Table 8: **3D indoor-outdoor classification:** The statistics of the best performing re-evaluated genotypes (12 genotypes) in all runs for each visual extraction methods.

Visual extraction methods	Max	Average	Worst	Stdev
Grey-scale averaging	88.31	69.82	58.55	±9.74
Active-ULBP	91.48	75.17	54.78	±11.23
Active-HOG	99.15	85.39	70.34	±9.74

The improvement shown by Active-ULBP in the environment categorisation problem may be due to the fact that ULBP is a good feature descriptor for detecting local binary texture patterns in texture images (Ojala et al. , 2002). HOG may also work well for texture images, especially if there are a lot of structures in the images (Dalal and Triggs, 2005).

Finally for the indoor-outdoor environment categorisation, the MGSI of the best performing re-evaluated evolved genotypes (3 genotypes) of the three visual extraction methods of all evolutionary runs was computed for 200 trials during which the agent experienced 10 different indoor and outdoor environments, with each environment uniformly and randomly rotated within the range $[-40°, 40°]$ to the original orientation with 20 different initial eye positions. For each type of visual extraction method of the sensory patterns, the MGSI was computed for each of the 100 time steps (Fig. 18).

Table 9: **3D indoor-outdoor categorisation:** The results of the anova test.

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1502.35	2	751.17	7.13	0.0027	3.29
Within Groups	3475.31	33	105.31			
Total	4977.66	35				

The fact that the MGSI did not show much improvement either for all conditions (visual extraction methods) or the two environments (indoor and outdoor) shows that the system did not make much use of coordinated sensory-motor control in order to disambiguate the ambiguous visual information. This actually was not a problem given the performance of the three visual extraction techniques. The system must have relied heavily on the internal states of the controller for the integration of sequences of experienced sensory states over time.

5 Discussion

In this paper we have investigated three visual extraction methods, that is, the grey-scale averaging (Mirolli et al. , 2010), Uniform Local Binary Patterns (Ojala et al. , 2002), and Histogram of Oriented Gradients (HOG, Dalal and Triggs, 2005), in the context of an active vision system.

Our first objective is to show that evolutionary based active vision systems can work in more complex scenes. This was demonstrated by using the grey-scale averaging method for more complex natural images taken from the camera of the iCub robot, as compared to synthetically generated hand-written italic-letter images with simple white background that were used in Mirolli et al. (2010). We also used this grey-scale method in the 3D iCub robot simulator platform for object categorisation and indoor-outdoor categorisation tasks.

Table 10: **3D indoor-outdoor categorisation:** The significance test result using paired t-test with test conditions of (p-value<0.05) and (p-value<0.01).

Compared Groups	*Mean Difference*	*t-value*	*p-value*	*Signf. Level=0.05 Bonf. Corr=0.0167*	*Signf. Level=0.01 Bonf. Corr=0.003*
Active-HOG and Greyscale	15.57	3.72	0.0010	Significant	Significant
Active-HOG and Active-ULBP	10.22	2.44	0.0237	Not Significant	Not Significant
Active-ULBP and Greyscale	5.35	1.28	0.1742	Not Significant	Not Significant

This was a more complex scenario, when compared to the (Mirolli et al. , 2010) letter experiment and our 2D iCub images experiment. In our 3D environment experiments, the agent (i,e. iCub) was confined within the environment, with the virtual camera located in the right eye position, which was extended to encourage more exploration of the scene. For instance, in the 3D indoor-outdoor environment categorisation experiment, the agent could only see a small fraction of the environment at the same time, encouraging the development of the ability to integrate sensory information over time to complete the task.

In addition to categorisation performance, the active vision system using the grey-scale method was able to use sensory-motor coordination for learning in the 2D images and 3D object categorisations. However, the system could not use intelligent control for learning the categorisation of the 3D outdoor-indoor environments and might have relied heavily on the internal states of the network, given the good performance of the system.

The second objective is to show that this active vision system can be further enhanced with pre-processing techniques for categorisation tasks.

In the 2D-image categorisation experiment, our proposed pre-processing technique, Active-ULBP had better average performance than the grey-scale, but was not significantly better. However, the grey-scale had better average performance than the Active-HOG, but also was not significantly better.

On the other hand, in the 3D object categorisation, Active-HOG showed better average performance than grey-scale, and was significantly better. While, the grey-scale outperformed the Active-ULBP in average performance but was not significantly better.

Also, in the 3D indoor-outdoor environment categorisation, Active-HOG showed better performance than the grey-scale and was

significantly better. While, the Active-ULBP was also better on the average than the grey-scale, but was not significantly better.

However, the active vision system, using pre-processing techniques has not demonstrated much use of intelligent control in the categorisation tasks, as evidenced by the MGSI. The Active-ULBP showed some evidence of intelligent control in the 2D and 3D object categorisation, and no evidence in the indoor-outdoor classification. While, Active-HOG showed some evidence of learning in the 2D but no significance evidence in the 3D experiments.

We will further discuss in the next sections: (i) the visual representation and active vision categorisation tasks; (ii) dynamics of the categorisation process.

5.1 The Visual Representation and Active Vision Categorisation Tasks

We investigated an active vision system based on Mirolli et al. (2010) for categorisation of more complex 2D images and 3D objects and indoor-outdoor environment categorisation.

All conditions in the categorisation tasks were the same, apart from the visual extraction methods (grey-scale averaging, ULBP, and HOG). Although, the size of visual inputs varied significantly between grey-scale (25) and ULBP (236), the number of inputs for HOG (36) was closer to that of the grey-scale; which works the best in some of the exeriments. This implies the number of inputs was not the only factor in the categorisation performance.

We used the iCub simulator platform as the basis for our 3D experiments. However, because of computational complexity, we only made use of 2 degrees of freedom for the right eye (pan and tilt) as we are not interested in calculating any depth information, and have excluded the vergence. Also for simplicity and because of the computational overhead, we excluded head, neck and other proprioceptive information from our experiments, where extending to the neck joints introduces redundancy into the degrees of freedom. For instance, in the 3D object categorisation, the objects were small enough that moving too much would lead to the objects being lost completely. Of course, if we introduce the torso then we could look at the object from different view points, but this will also introduce additional computational cost for the evolutionary method and goes well beyond the aims of this work. The focus is therefore on the eye movements alone exploring a scene.

In the experiment of object categorisation in 3D, the first challenge was the randomly varied size and orientations in each trial, and the second challenge was the high ambiguity of the stimuli of the objects that were investigated (i.e, sphere, cube, cone, and torus). Despite, the complexity of the problem, the three visual extraction methods that were investigated performed well.

On the other hand, the complexity of the indoor and the outdoor environment classification may be due to the following reasons: (i) In contrast to the object categorisation problem in which categorisation involves one category of object in each trial, environment categorisation can involve many objects within the same environment, which may or may not belong to shared category, and each of which may be in different spatial locations. Apart from this structural information, there is also textural information to be processed. (ii) The system therefore may have to use the totality of contextual information within each environment to complete the discrimination task, coupled with random rotation in each trial.

In spite of the complexity of the problem, the active vision system also performed well over the course of testing for all the visual extraction methods under investigation.

The improvement in performance of Active-HOG in the 3D object categorisation may be due to the more structural nature of the object categorisation problem. Equally the good performance of Active-HOG also in indoor-outdoor environment categorisation may have been due to more structural information in the datasets. Typically, in most indoor and outdoor environments, the objects and structures are more conspicuous. For instance, a typical indoor environment may have conspicuous objects, such as tables, chairs, beds, and so on, while outdoor environments may have structures, such as houses, cars, trees and the like. On the other hand, the fact that Active-ULBP performed well in categorisation tasks irrespective of the environmental context (2D images or 3D indoor-outdoor) is evidence that ULBP is a good feature descriptor for detecting patterns in texture images, and a good feature descriptor in many applications (Pietikainen et al. , 2011; Tapia et al. , 2014).

5.2 Dynamics of the Categorisation Process

The categorisation performance of an active vision system may not depend as much on the complexity of the system design as on the extent to which the agent may use the dynamic interaction of the sensory-motor components to exploit regularities that pertain to the different categories in the sensor input-space. We investigated the dynamics using the Modified Geometric Separability Index (MGSI) in order to analyse the extent to which the active vision system used its intelligent motor control to experience sensory stimuli that could be unambiguously associated with a particular category for each of the three visual extraction methods in the input space.

In the 2D environment in particular, the MGSI results showed that all three visual extraction methods generated sensory patterns that allowed the system to move from very ambiguous to less ambiguous stimuli. Active-ULBP also provided less ambiguous stimuli (i.e. the highest peak in the MGSI graph) than the other methods. However, grey-scale was a little bit more consistent over time than Active-ULBP.

In the 3D object categorisation, grey-scale was able to use sensory-motor coordination over time to experience more discriminative stimuli than the other two visual representation methods. Active-ULBP also showed some slight use of motor responses in moving to less ambiguous stimuli over time. However, even though Active-HOG generally had less ambiguous stimuli from the start, it was not to a great extent able to use eye movements to experience less ambiguous sensory stimuli. The low ambiguity of Active-HOG in most time steps may be due to the highly structural nature of the problem, and this may also have enhanced its recognition capability. That said, the inability to use sensory-motor coordination to experience less ambiguous stimuli over time, might have been due to the low ambiguity experienced by the system from the outset, and if the system can get good results from random eye movements it won't tend to evolve intelligent control. In this context, there was little need to make use of eye movements to reduce ambiguity over time.

Also, the occurrence of oscillatory behaviour by the Active-HOG stimuli, may in part be due to the best genotypes that were used for the computation of the MGSI. As the use of HOG only transforms the input pixels into a different representation, and of itself does not perform any classification. Hence the agent was learning to perform the classification but was not relying much on active vision to do so. This points to the need for further work to investigate the best combinations of representation and active learning.

On the whole, in both the 2D and 3D ob-

ject categorisation, grey-scale used more eye movements than the other two methods to influence the performance of the active vision system. Active-ULBP also showed more use of eye movements to reduce visual ambiguity in 2D than in 3D and outperformed Active-HOG in both environments.

On the other hand, in both indoor and outdoor environment categorisation experimental contexts, the active vision system seems to have relied heavily on the internal dynamics of the neural network controller. This was because there was only a slight improvement in the MGSI values for the three visual extraction methods over time. Since the performance of the three visual extraction methods was good, the system must have used the internal states to integrate the very ambiguous perceptual information over time. However, we are not committed to this view and this may be a subject of future research.

Also, the probable reason for the poor learning of the active vision system as compared to the object categorisation experiments may be due to the different context of categorisation. In the object categorisation experiments there was only one object to be categorised in an image/environment, whereas in the indoor and outdoor environment categorisations there was more variability. For example, there were many structures, each of varying sizes and spatial locations. There were also other variables such as texture, and some of the variables may not be peculiar to a particular environment, which is to say that some structures are common to both indoor and outdoor environments. It may therefore be difficult for the system to discover regularities that are particular to an environment (indoor or outdoor) through dynamic sensory-motor interaction alone.

6 Conclusion and Future Works

We have extended the architecture described in (Mirolli et al. , 2010) using grey-scale averaging feature extaction method to more complex scenes for 2D object categorisation and 3D object and outdoor-indoor environment categorisation.

We further sought to improve the categorisation performance of the active vision system with pre-processing techniques using Uniform Local Binary Patterns (ULBP, Ojala et al. , 2002), and Histogram of Oriented Gradients (HOG, Dalal and Triggs, 2005), in the 2D and the 3D environments.

In each experiment the performance of the 3 visual extraction methods were compared statistically, and the dynamics was evaluated using the MGSI measure of separability.

The results showed a mixed picture, dependent on the particular problem. For the 2D iCub images, the Active-ULBP showed the best average performance, and was significantly better than the other methods at the $p = 0.05$ level, but not at the $p = 0.01$ level, and all methods showed some evidence of intelligent control of the eye movement in the MGSI. For the 3D object experiment Active-HOG showed a significantly better performance than the other two methods at the stricter significance level of $p = 0.01$, but seemed to require less intelligent control due to high discrimination in the early stages of the re-evaluation trials. For the 3D environment experiments, again the Active-HOG showed the best performance, significantly better than the grey-scale method at the $p = 0.01$ level, but not significantly better than the Active-ULBP method, and the behaviour of the system exhibited virtually no evidence of intelligent selection of stimuli, according to the MGSI results.

In summary, the investigation has demonstrated that evolutionary active vision system using greyscale averaging extraction method can work in complex scenes. However, the optimal choice of feature extaction technique is highly dependent on the specific problem. Also, the evidence for intelligent control is also dependent on the specific problem and choice of pre-processing, with some systems showing good performance, but little evidence of guiding the agent towards more easily discriminated stimuli. If the agent can solve the problem without intelligent movement the system will not evolve such behaviours. Thus, we demonstrate that simple pre-processing step can also increase categorisation performance in some scenarios, and that the reliance on active control is lower as the agents (non-active) categorisation performance increases. Such understanding could help focus research on developing the best combination of active and non-active components.

It should be noted that other form of behavioural analyses can be performed apart from the Modified Geometric Separability Index (MGSI) to understand more of the categorisation process. However, the focus of this research paper is in investigating the impact of representation on active learning and classification accuracy and not on underlying phenomena beyond the categorisation process.

In the future, it would be interesting to evaluate more of the behaviours of the currently used pre-processing techniques by using other behavioral analyses tools, so as to understand more of their categorisation process. Also, to better understand the internal dynamics of the system, neural analysis can be carried out to understand the patterns of the neuron activation over time using the best evolved genotypes of the three visual extraction methods used for the MGSI in the re-evaluation stage and under the same experimental conditions.

Similarly, one can fix the eye movement for the outdoor-indoor environment classification, so as know if the performance of the system (grey-scale, ULBP and HOG) was mainly as a result of internal states of the system. If the performance still remains at a level comparable to the system that uses adaptive eye movement, it will be a further indication of systemic reliance on the internal states of the controller to complement sensory-motor coordination.

Furthermore, one can investigate other pre-processing techniques for the problem at hand to know the best combinations of representation and active learning. For example, convolution neural networks have shown excellent results on object recognition and other problems in computer vision, but adapting them and their training to the active vision framework is challenging, particularly due to the high computational cost of the evolutionary training. Another area for future work is to implement the active vision system on a physical robotic hardware platform in order to see if the system could replicate the same level of performance in the real system. Although, we tried to simulate the conditions of the real world as much as possible, it is not automatic that the algorithms will perform as well in the real system.

References

Ojala T, Pietikainen M, and Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern,Analysis and Machine Intelligence, IEEE Transactions on, 24(7):971987.

Dalal N and Triggs B (2005). Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 886893. IEEE.

Belongie S, Malik J, and Puzicha J (2002). Shape matching and object recognition using shape contexts. IEEE transactions on pattern analysis and machine intelligence, 24(4):509522.

Avraham T and Lindenbaum M (2010). Esaliency (extended saliency): Meaningful attention using stochastic image modeling. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32(4):693-708.

Kagan I and Hafed Z (2013). Active vision: microsaccades direct the eye to where it matters most. Current Biology, 23(17):712-714.

Osuna E, Freund R, and Girosit F (1997). Training support vector ma- chines: an application to face detection. In Computer vision and pattern recognition. Proceedings., 1997 IEEE computer society conference on, pages 130 136. IEEE.

Tsotsos J (1992). On the relative complexity of active vs. passive visual search. International journal of computer vision, 7(2):127141.

Nolfi S (1998). Adaptation as a more powerful tool than decomposition and integration: experimental evidences from evolutionary robotics. In Fuzzy Systems Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on, volume 1, pages 141146. IEEE.

Mirolli M, Ferrauto T, and Nolfi S(2010). Categorization through evidence accumulation in an active vision system. Connection Science, 22(4):331 354.

Kato T and Floreano D (2001). An evolutionary active-vision system. In Evolutionary Computation, 2001. Proceedings of the 2001 Congress on, volume 1, pages 107114. IEEE.

Croon G (2008). Adaptive Active Vision. PhD thesis, Universiteit Maastricht, Gildeprint, The Netherlands, 3.

Denzler J and Brown C (2002). Information theoretic sensor data selection for active object recognition and state estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2):145157.

Borotschnig H, Paletta L, and Pinz A (1999). A comparison of probabilistic, possibilistic and evidence theoretic fusion schemes for active object recognition. Computing, 62(4):293319.

Tuci E (2014). Evolutionary swarm robotics: genetic diversity, task-allocation and task-switching. In International Conference on Swarm Intelligence, pages 98109. Springer.

Marocco D and Floreano D (2002). Active vision and feature selection in evolutionary behavioral systems. From animals to animats, 7:247255, 2002.

Nolfi S and Marocco D (2000). Evolving visually-guided robots able to discriminate between different landmarks. In In From Animals to Animats 6. Proceedings of the sixth International Conference on Simulation of Adaptive Behavior SAB-00. Citeseer, 2000.

Morimoto G and Ikegami T (2004). Evolution of plastic sensory-motor coupling and dynamic categorization. Proceedings of Artificial Life IX, pages 188-193, 2004.

Magnussen S (2000). Low-level memory processes in vision. Trends in neurosciences, 23(6):247251, 2000.

Le Meur O, Le Callet P, Barba D, Thoreau D, and Francois E (2004). From low-level perception to high-level perception: a coherent approach for visual attention modeling. In

Electronic Imaging 2004, pages 284295. International Society for Optics and Photonics, 2004.

Diamant E (2008). Unveiling the mystery of visual information processing in human brain. Brain research, 1225:171178, 2008.

Schembri M and Belardinelli M (2015). Evolved simulated agents exhibit size constancy abilities in solving an online size discrimination task. In EAPCogSci, 2015.

Ahonen T, Hadid A, and Pietikainen M (2006). Face description with local binary patterns: Application to face recognition. IEEE transactions on pattern analysis and machine intelligence, 28(12):20372041, 2006.

Stefanou S and Argyros A (2012). Efficient scale and rotation invariant object detection based on hogs and evolutionary optimization techniques. Advances in Visual Computing, pages 220229, 2012

kas Kass M, Witkin A, and Terzopoulos D (1988). Snakes: Active contour models. International journal of computer vision, 1(4):321331, 1988.

Terzopoulos D and Rabie T (1995). Animat vision: Active vision in artificial animals. In Computer Vision, 1995. Proceedings., Fifth International Conference on, pages 801808. IEEE, 1995.

Minut S and Mahadevan S (2001). A reinforcement learning model of selective visual attention. In Proceedings of the fifth international conference on Autonomous agents, pages 457464. ACM, 2001.

Vidal-Calleja T and Sanfeliu A and Andrade-Cetto J (2010). Action selection for single-camera slam. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 40(6):15671581, 2010.

Guerrero P, Ruiz-Del-Solar J, Romero M, and Angulo S (2010). Task-oriented probabilistic active vision. International Journal of Humanoid Robotics, 7(3):451476, 2010.

Dame A and Marchand E (2013). Using mutual information for appearance-based visual path following. Robotics and Autonomous Systems, 61(3):259270, 2013.

Davison A (2005). Active search for real-time vision. In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, volume 1, pages 6673. IEEE, 2005.

Nolfi S (2005). Categories formation in self-organizing embodied agents. Handbook of categorization in cognitive science, pages 869-889, 2005.

Pugliese F (2014). Development of categorisation abilities in evolving embodied agents: A study of internal representations with external social inputs. In Evolution, Complexity and Artificial Life, pages 123134. Springer, 2014.

Nolfi S and Parisi D (1995). Evolving non-trivial behaviors on real robots: an autonomous robot that picks up objects. Topics in artificial intelligence, pages 243254, 1995.

Tuci E, Massera G, and Nolfi S (2010). Active categorical perception of object shapes in a simulated anthropomorphic robotic arm. IEEE transactions on evolutionary computation, 14(6):885899, 2010.

Lanihun O, Tiddeman B, Tuci E, and Shaw P (2015). Improving active vision system categorization capability through histogram of oriented gradients. In Conference Towards Autonomous Robotic Systems, pages 143148. Springer, 2015.

Tuci E, Nolfi S, Mirolli M, Ferrauto T and Massera G (2009). Two examples of active categorisation processes distributed over time. In Proceedings of the Ninth International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems, pages 4956, 2009.

Lanihun O, Tiddeman B, Tuci E, and Shaw P (2014). Enhancing active vision system categorization capability through uniform local binary patterns. In Artificial Life and Intelligent Agents Symposium, pages 3143. Springer, 2014.

Ojala T, Pietikainen M, and Harwood D (1996). A comparative study of texture measures with classification based on featured distributions. Pattern recognition, 29(1):5159.

Tsagarakis N et. al. (2007). icub: the design and realization of an open humanoid platform for cognitive and neuroscience research. Advanced Robotics, 21(10):11511175.

Leitner J et al (2017). Learning visual object detection and localisation using icvision. http://www.sciencedirect.com/ science/article/pii/S2212683X13000443, 2017. Accessed: 2017-06-28.

Tuci E (2016). The simple icub simulator used in this journal paper. Faculty of Computer Science University of Namur, rue Grandgagnage 21, 5000, Namur, Belgium.

Thornton C (1998). Separability is a learners best friend. In 4th Neural Computation and Psychology Workshop, London, 911 April 1997, pages 4046. Springer.

Google (2017). Google Images. https://www.google.co.uk/imghp?hl=entab=wi

Goldberg D (1999). Genetic algorithm in search, optimisation and machine learning. Reading, MA. Addison-Wesley.

Canny J. A (1986). Computational Approach To Edge Detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, 8(6):679698.

Pietikainen M, Hadid A, Zhao G, and Ahonen T (2011). Local binary patterns for still images. In Computer vision using local binary patterns, pages 13-47. Springer

Tapia J, Perez C, and Bowyer K (2014). Gender classification from iris images using fusion of uniform local binary patterns. In ECCV Workshops (2), pages 751-763, .

Zaytseva E, Segui S, and Vitria J (2012). Sketchable histograms of oriented gradients for object detection. In Iberoamerican Congress on Pattern Recognition, pages 374-381. Springer.

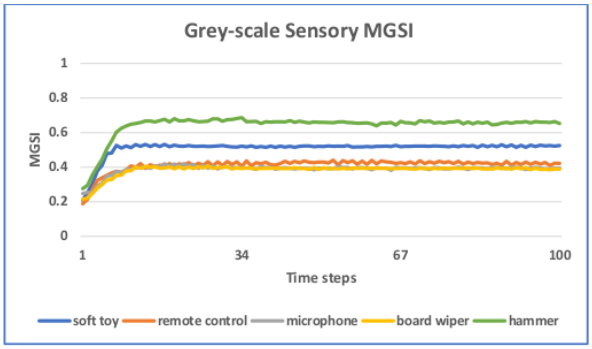
Kass M, Witkin A, and Terzopoulos D (2008). Snakes: Active contour models. International journal of computer vision, 1(4):321-331.

Super Computing Wales (2019). https://www.supercomputing.wales/ .

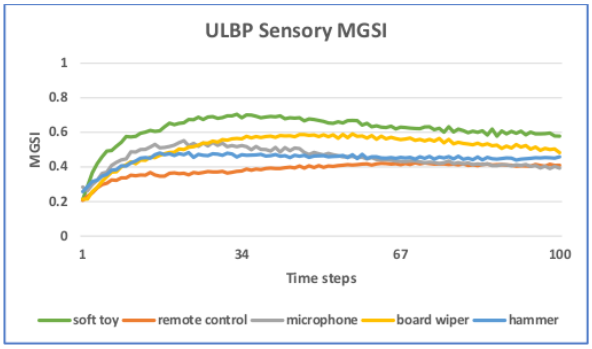
Gropp, W.D., Gropp, W., Lusk, E., Skjellum, A. and Lusk, A.D.F.E.E. (1999). Using MPI: portable parallel programming with the message-passing interface (Vol. 1). MIT press.

Ferrauto T, Tuci E, Mirolli M, Massera G,and Nolfi S (2009). Two examples of active categorisation processes distributed over time. In Proceedings of the Ninth International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems, pages 49-56.

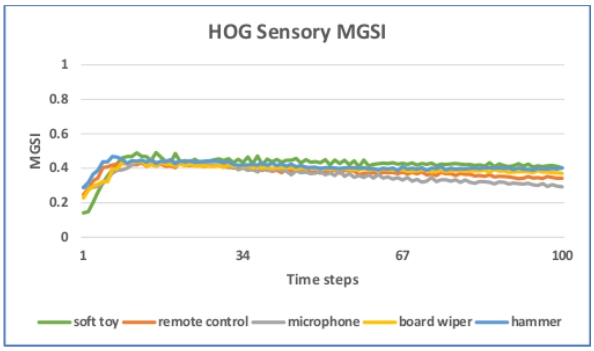
Tuci E, Massera G, and Nolfi S (2010). Active categorical perception of object shapes in a simulated anthropomorphic robotic arm. IEEE transactions on evolutionary computation, 14(6):885-899.



(a) Modified Geometric Separability (MGSI) of the stimuli provided by the greyscale averaging method



(b) Modified Geometric Separability (MGSI) of the stimuli provided by the Active-ULBP method.



(c) Modified Geometric Separability (MGSI) of the stimuli provided by the Active-HOG method.

Figure 12: **2D iCub images:** Modified Geometric Separability (MGSI) of the stimuli provided by the visual extraction methods for the 2D images: an increase over time indicates the agent is moving the system towards regions that increase discrimination ability.

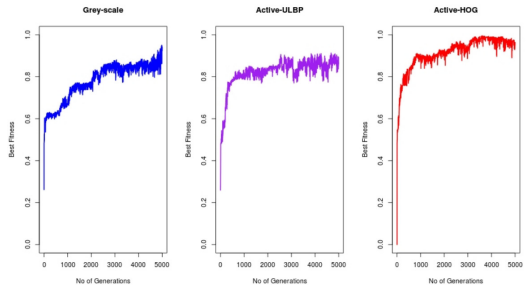


Figure 13: **3D object categorisation:** The fitness graphs of the best evolutionary runs of the three visual extraction methods.

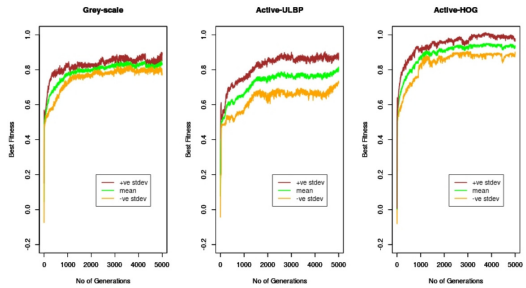
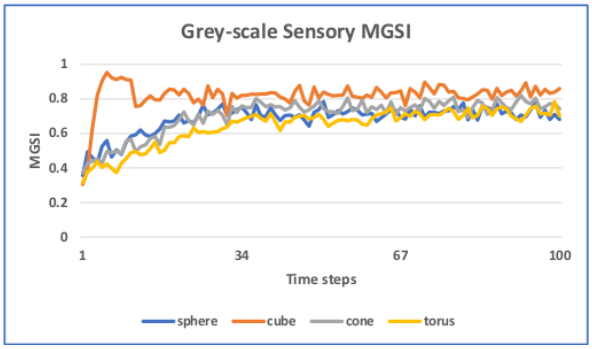
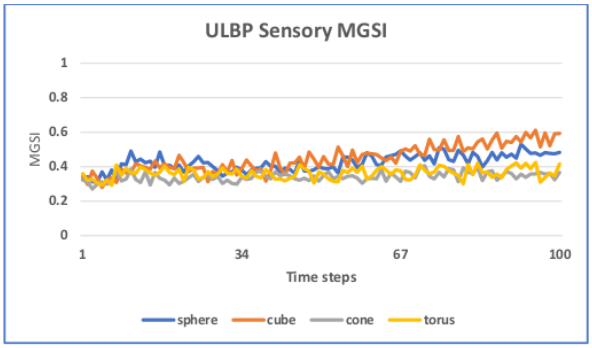


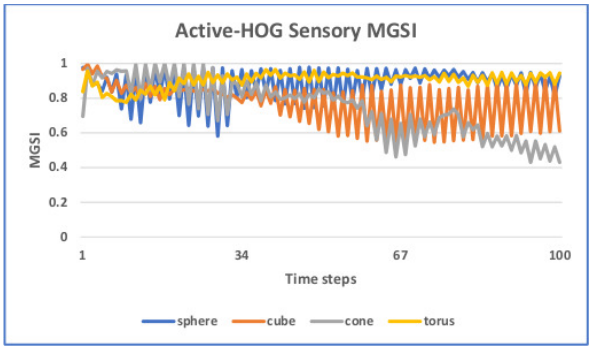
Figure 14: **3D object categorisation:** Shows the graph of the mean (average) of all fitness in each generation of the 5000 generations for 6 evolutionary runs and their positive (+ve stdev) and negative (-ve stdev) standard deviation in each generation for the three methods of visual extraction.



(a) Modified Geometric Separability (MGSI) of the stimuli provided by the greyscale averaging method



(b) Modified Geometric Separability (MGSI) of the stimuli provided by the Active-ULBP method.



(c) Modified Geometric Separability (MGSI) of the stimuli provided by the Active-HOG method.

Figure 15: **3D object categorisation:** Modified Geometric Separability (MGSI) of the stimuli provided by three visual extraction methods for the 3D objects: an increase over time indicates the agent is moving the system towards regions that increase discrimination ability.

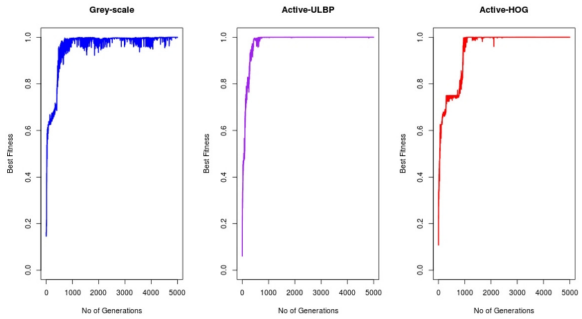


Figure 16: **3D indoor-outdoor classification:** The fitness graphs of the best evolutionary runs of the three visual extraction methods.

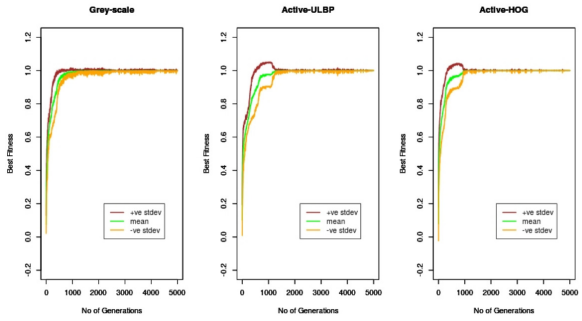
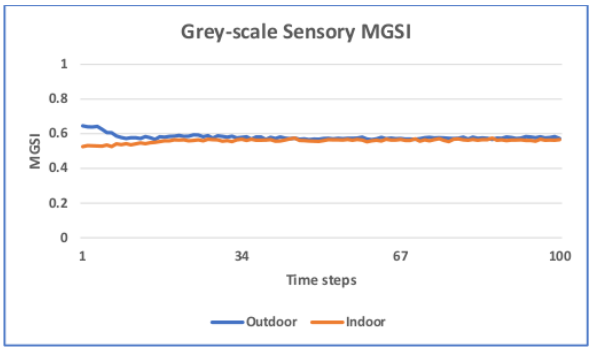
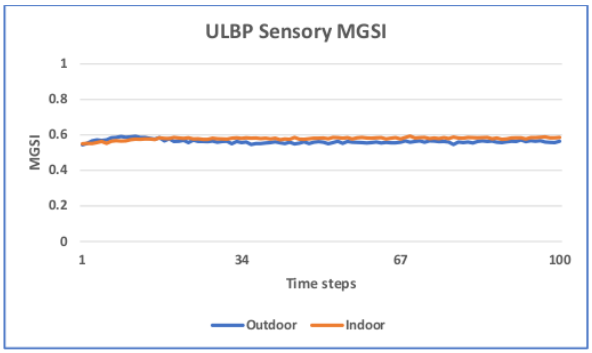


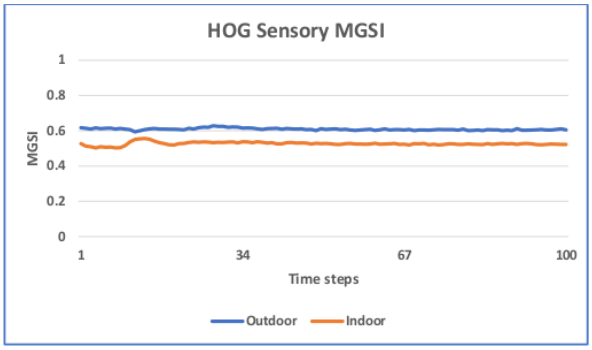
Figure 17: **3D indoor-outdoor classification:** Shows the graph of the mean (average) of all fitness in each generation of the 5000 generations for 12 evolutionary runs and their positive (+ve stdev) and negative (-ve stdev) standard deviation in each generation for the three methods of visual extraction.



(a) Modified Geometric Separability (MGSI) of the stimuli provided by the grey-scale averaging method



(b) Modified Geometric Separability (MGSI) of the stimuli provided by the Active-ULBP method.



(c) Modified Geometric Separability (MGSI) of the stimuli provided by the Active-HOG method.

Figure 18: **3D indoor-outdoor classification:** Modified Geometric Separability (MGSI) of the stimuli provided by the three visual extraction methods for the 3D indoor and outdoor environments: An increase over time indicates the agent is moving the system towards regions that increase discrimination ability.