Object Localisation and Tracking Through Subsymbolic Classification

Hugo Vieira Neto and Ulrich Nehmzow
Department of Computer Science
University of Essex
Wivenhoe Park
Colchester, Essex CO4 3SQ, UK
{hvieir, udfn}@essex.ac.uk

Abstract

This paper presents results of experiments in subsymbolic processing of visual data, to achieve identification and tracking of *arbitrary* objects, which are intended to be used in autonomous robots for novelty detection and navigation purposes.

Artificial neural networks with unsupervised training are used as the classification stage for the vision system, in order to provide the robot the ability to develop its own representations from perceptual data without the need of any external human-provided information. We present an evaluation of the behaviour of the system when using very simple feature extraction techniques, such as horizontal and vertical average histograms, as well as average coarse coding.

1 Introduction

Animals can rapidly detect novelties in their normal environment using different sources of perceptual information. As it is relevant for survival, this particular ability to be aware of environmental changes is very important, for example, to find food or avoid predators.

The sense of vision plays an essential role in animal behaviour either for tracking movements or identifying objects of interest. It follows that novelty detection using artificial vision can be of equal importance to autonomous mobile robots. If such an ability is feasible to be implemented artificially, possible applications are in automated inspection and surveillance tasks.

Previous work done by Marsland et al. (Marsland et al., 2000, 2002) has given successful results in novelty detection, using habituation and sonar data in real robots. Habituation is a reduction in responses to stimuli that are repeatedly presented to the system and therefore can be used to highlight new perceptions.

The ultimate aim of the present research is to achieve a novelty detection system to be used in autonomous mobile robots using the known model of habituation and artificial vision as the main perceptual input.

Dealing with visual information is often a difficult task, because of the large amounts of data involved and the processing power required. As a consequence, it is necessary to arrange some sort of internal representation through models of the aspects of interest in the environment. These models, which are simplified abstractions of the original aspects, are expected to describe relevant features and eliminate unnecessary details.

However, "relevant features" are not always clear to the designer of such a system, nor what constitutes "unnecessary details". Nehmzow (Nehmzow, 1999) argues that in this case a method of model acquisition through robot learning is a more feasible solution, rather than explicitly supplying *a priori* models.

In this first stage of our research we are looking for suitable ways to extract relevant features from raw image data in order to manage processing in reasonable time and storage size, which are especially important for autonomously operating robots that have to respond to stimuli in real-time.

We are particularly interested in implementing an attention mechanism to be used in an active vision approach to localise and physically track objects of interest, as part of the whole novelty detection system to be developed.

2 Experiments

2.1 Objective

Our system was designed to determine where an arbitrary target object was located within an image frame. The target was provided by the user and its location was determined by the system as one of 25 non-overlapped subimages of the frame (see Figure 3).

2.2 Experimental Setup

Our experiments were conducted using a Magellan Pro mobile robot, which is shown in Figure 1. The robot is equiped with a colour CCD camera, pan-tilt unit and framegrabber board, which are able to acquire RGB images with 160×120 pixels in size. Images were acquired while the robot was not in motion.

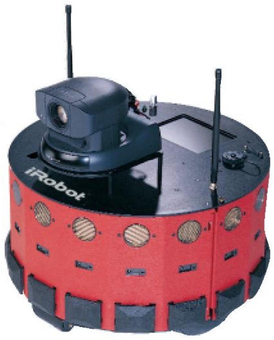


Figure 1: The Magellan Pro mobile robot

2.2.1 System Architecture

The functional blocks of our approach towards a system that can be trained to localise and track a small set of objects is given in Figure 2. Two very simple and computationally inexpensive image coding techniques were selected to be evaluated as the pre-processing stage.

As the classification stage, two different connectionist approaches with unsupervised training were chosen to be compared.

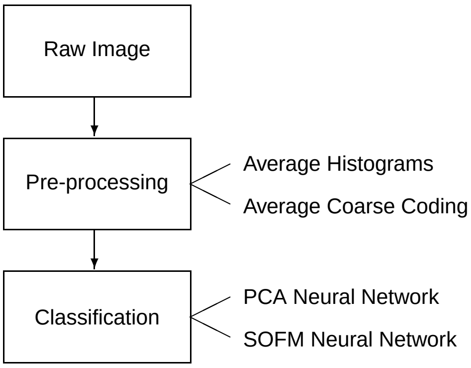
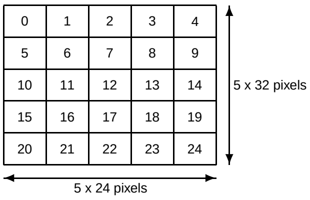


Figure 2: Functional blocks of the vision system

The pre-processing scheme consisted of partitioning the original image into 25 sub-images with 32×24 pixels in size, as shown in Figure 3. After partitioning, pre-processing continues with each raw sub-image serving as input for a feature extraction stage, in order to generate more compact input vectors to the classification stage.



0	1	2	3	4
5	6	7	8	9
10	11	12	13	14
15	16	17	18	19
20	21	22	23	24

Figure 3: The 25 sub-images that result from partitioning the original image

2.2.2 Image Pre-processing

Two image coding techniques were used in different sets of experiments. The first consisted of computing the horizontal and vertical average histograms of the three RGB colour channels and stacking them into single feature vectors with $3 \times (32 + 24) = 168$ elements (Figure 4).

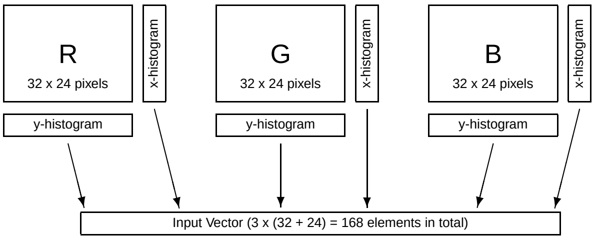


Figure 4: Image pre-processing using histograms

The second image pre-processing mechanism tested consisted of computing the average coarse coding of the three RGB colour channels within an 8×8 neighbourhood, resulting in feature vectors with $3 \times (\frac{32}{8} \times \frac{24}{8}) = 36$ elements (Figure 5).

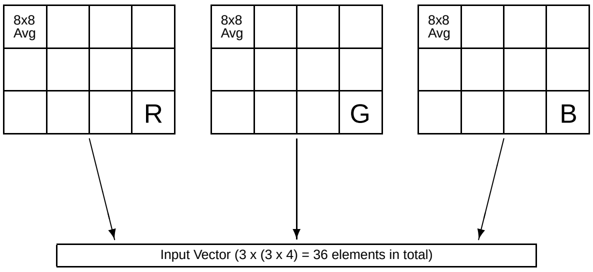


Figure 5: Image pre-processing using coarse coding

2.2.3 Image Classification

Two different neural network approaches were also tested in different sets of experiments. The first was a torus-shaped self-organising feature map (SOFM) (Kohonen, 1984) with 10×10 units, the second was a single-layer feedforward neural network that performs principal component analysis (PCA) (Sanger, 1989; Ballard, 1997) with 16 units.

The SOFM network was trained according to the usual winner-takes-all approach (Kohonen, 1984), using the similarity matching given in Equation 1, where $\vec{w_c}$ is the winner among all $\vec{w_i}$ units for a given input $\vec{x}$.

$$\| \vec{x}(t) - \vec{w_c}(t) \| = \min_i \{ \| \vec{x}(t) - \vec{w_i}(t) \| \} \qquad (1)$$

During the learning phase, the weight vectors of the winner and its neighbours were modified according to Equation 2, where α is the learning rate ($0 < \alpha < 1$) and N_c is the topological neighbourhood of the winner.

$$\vec{w}_i(t+1) = \vec{w}_i(t) + \alpha[\vec{x}(t) - \vec{w}_i(t)],\ i \in N_c \qquad (2)$$

In our experiments, both the learning rate and the topological neighbourhood size decreased with the training cycles, as shown in Equations 3 and 4.

$$\alpha(m) = \exp(-10m/M) \qquad (3)$$

$$N_c = \begin{cases} 3 & \text{if } 0 < m \leq 0.2M \\ 2 & \text{if } 0.2M < m \leq 0.5M \\ 1 & \text{if } 0.5M < m \leq M \end{cases} \qquad (4)$$

M is the total number of training cycles (100 in all experiments).

The PCA networks were trained with the Generalised Hebbian Algorithm (GHA) (Sanger, 1989). To compute the output vector $\vec{y}$ for a given input vector $\vec{x}$, Equation 5 was used. Equation 6 describes how the weights w_{ij} of the network were adapted.

$$y_i = \sum_{j=1}^{J} w_{ij} x_j \qquad (5)$$

$$\begin{array}{l} \Delta w_{ij} = \alpha y_i [x_j - \sum_{k=1}^{i} y_k . w_{kj}] \\ w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij} \end{array} \qquad (6)$$

For the PCA network it was necessary to compute the average vector of the training data and then subtract it from all vectors in the set, in order to obtain zero-mean data. The learning rate for the PCA network was also made to decrease exponentially (Equation 7).

$$\alpha(m) = 0.1 \exp(-m/M) \qquad (7)$$

Each neural network architecture was trained and tested for each image coding scheme, resulting overall in four different experiments. The data set for training consisted of 50 images, each of an orange football, a blue cylinder and a green box, acquired against the unstructured background of the Brooker Laboratory at the University of Essex. The objects were distributed in random positions and orientations, as shown in Figure 6.

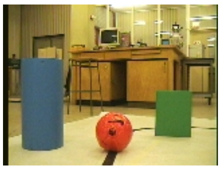


Figure 6: A sample image from the training set

2.3 Experimental Results

2.3.1 Test Images

A set of ten different images of each object in different locations within the image frame was used to test the neural networks. These test images were acquired in the same non-structured environment as the training images, altough against a different background. Additional target images were also taken from each object positioned exactly to lie in the central sub-image. Figure 7 depicts a target image for the orange ball.

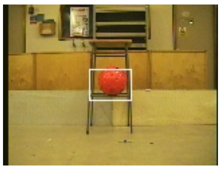


Figure 7: Target image of the orange football

2.3.2 Identification of Target Sub-images

For the SOFM architecture, the activation elicited by the target image was subsequently used to identify that sub-image in which the target object lay in each of the ten test images. The measure of similarity between the activations of the target sub-image and the other sub-images was a three-dimensional euclidean distance metric, which made use of the two-dimensional coordinates of the winner unit in the map and its activation value.

For proper results, the activation value of the winner must be scaled according to some criteria to make its contribution to the distance measure proportional to the contribution of its coordinates in the SOFM. The scaling criteria employed in our experiments is given by Equation 8, where S is the side dimension of the square SOFM (10 in our experiments).

$$a_c = S \times \parallel \vec{x} - \vec{w}_c \parallel / \max_i \{ \parallel \vec{x} - \vec{w}_i \parallel \} \qquad (8)$$

The two-dimensional coordinates of the winner units must also take into account the effect of the torus boundaries of the map to compute the distance between them. Ambiguous coordinates of the winner units were selected to minimise the euclidean distance between them.

For the the PCA network, a similarity measure of the same nature was also used. The euclidean distance was used again to evaluate the output vector of the network for the target sub-image in comparison to the output vectors of the sub-images from the test images. However, in the case of the PCA architecture, the simple n-dimensional euclidean distance was used without the need of any scaling procedure.

Figure 8 shows an example of the results for the test images of the orange football using the histogram-based image coding and the PCA network. It can be seen from these results that the system is reasonably robust to small changes in scale and translations of the target object within the detected sub-images.

To evaluate results, we manually determined the correct system response as that sub-image within which at least 25% of the target surface lay. We then composed

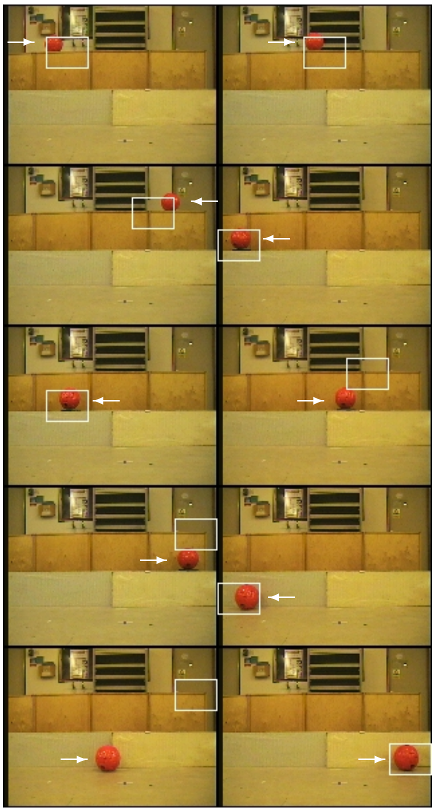


Figure 8: Results obtained using the histogram-based image coding and the PCA network (the arrows indicate the target object and the rectangles indicate the sub-image in which the target was detected)

this correct target identification against the one the system produced. A summary of the results obtained for all the four experiments is shown in Table 1.

Table 1: Success rates of the experiments

Average Histograms			
	Blue Cylinder	Green Box	Orange Football
SOFM Network	80%	30%	20%
PCA Network	90%	30%	70%

Average Coarse Coding			
	Blue Cylinder	Green Box	Orange Football
SOFM Network	70%	20%	20%
PCA Network	90%	10%	70%

A set of χ^2 tests were also performed on the results of the experiments. As only a restricted amount of test images was available, it was necessary to group neighbouring sub-images to satisfy practical constraints of the χ^2 test. This roughly resulted in the division of the image frame into four regions, as illustrated in Figure 9.

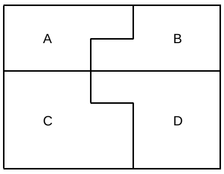


Figure 9: Sub-image groups used for the χ^2 test

The χ^2 tests for all experiments have shown statistically significant correlation between actual and identified target positions using all of the 30 test images ($p = 0.05$).

From the previous analyses, it can be inferred that the PCA network has a better performance for both image coding techniques in this particular task. In a similar way, the average colour histograms image coding scheme provide better results for both neural network architectures.

More importantly, however, is that the results given by the χ^2 test indicate that our approach identifies the target object's location at least within four large areas of the image frame. Altough this location information might seem very coarse at first, it is enough to be used to close the loop between perception and action, driving the camera's pan-tilt unit in two-dimensions to follow the target object and position it in the centre of the frame.

Our argument is that the simple fact of moving the camera in the direction of the target object in an active vision approach will not only be feasible, but will also improve the target localisation performance.

3 Conclusions

The eventual application this research is aimed at is that of novelty detection and navigation in mobile robots using visual information. In this paper, we presented a sub-symbolic, unsupervised mechanism to identify and track objects within the image frame using passive vision.

The results of our experiments have shown good performances for the blue cylinder when compared to the performances for the orange football and specially the green box, which was very poor. This might be attributed to the fact that the blue cylinder was the biggest object in the set and thus occupied a greater area within the image frame, increasing its probability of being identified. Another plausible explanation resides in the colour of the cylinder, which offered the greatest contrast against a mostly yellow background.

Feature extraction using average coarse coding with a neighbourhood of 8×8 pixels gives as a result a very blurry version of the original image, where basically only colour features are preserved. In spite of reduced details, the presented unsupervised neural network approaches were able to identify the target objects reasonably, indicating their suitability to constitute a simple colour-based object detector. Future work includes testing approaches using hue, saturation and intensity parameters

(HSI colour model), which may result in improved performance in class separation with respect to colour features (Jain et al., 1995).

Average histograms seem to represent both shape and colour information and presented a good tolerance to scale and translation within the detected sub-image. This image coding scheme provided even better results than coarse coding. Further tests need to be conducted with objects that are not present in the training data set, in order to evaluate the robustness of this method to objects that have never seen before by the system.

The PCA network was faster to train and also had the advantage of having a simpler similarity measure procedure than the SOFM. However, the convergence of the algorithm used to train the PCA architecture is very sensitive to several details. This includes the need of zero-mean data and small learning rate values for the training phase.

A special conclusion that arises from our χ^2 analysis is that the combination of techniques employed in our experiments are able to estimate the position of the target image in four main regions of the image frame. Despite being very simple, this characteristic is useful enough to drive an attention mechanism for an active vision system, according to some preliminary results of our next step in future work.

Our expectations are that active vision will contribute to improve the ability of the system to locate the target object. Several case studies that support better performances of active vision approaches are presented by Pfeifer and Scheier (Pfeifer and Scheier, 1999). Future work also includes the development of a "self-motivated" attention mechanism, in order to eliminate the need of user-provided target objects. The target is desired to be automatically selected, and the attention of the system directed to it, by the amount of novelty in its features.

Our initial analysis suggests that the approach presented in this paper is useful as part of the entire visual novelty detection system we have in sight. The feature extraction techniques employed so far are fairly sensitive to objects with saturated colours, altough further experiments using monochromatic images need to be executed to confirm this hypothesis.

Acknowledgement

Hugo Vieira Neto is sponsored by a CEFET-PR and CAPES Foundation fellowship, whose support is gratefully acknowledged.

References

Dana H. Ballard. *An Introduction to Natural Computation*. MIT Press, Cambridge, MA, 1997.

Ramesh Jain, Rangachar Kasturi, and Brian G. Schunk. *Machine Vision*. McGraw-Hill, New York, NY, 1995.

Teuvo Kohonen. *Self-Organization and Associative Memory*. Springer-Verlag, Berlin; New York, NY, 1984.

Stephen Marsland, Ulrich Nehmzow, and Jonathan Shapiro. Detecting novel features of an environment using habituation. In *From Animals to Animats: Proceedings of the 6th International Conference on Simulation of Adaptive Behaviour (SAB'2000)*, pages 189–198, Paris, France, 2000. MIT Press.

Stephen Marsland, Ulrich Nehmzow, and Jonathan Shapiro. Environment-specific novelty detection. In *From Animals to Animats: Proceedings of the 7th International Conference on Simulation of Adaptive Behaviour (SAB'02)*, Edinburgh, UK, 2002. MIT Press.

Ulrich Nehmzow. Vision processing for robot learning. *Industrial Robot*, 26(2):121–130, 1999.

Rolf Pfeifer and Christian Scheier. *Understanding Intelligence*. MIT Press, Cambridge, MA, 1999.

Terence D. Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, 2:459–473, 1989.