# VISUAL NOVELTY DETECTION FOR AUTONOMOUS INSPECTION ROBOTS


**Hugo Vieira Neto**

A thesis submitted for the degree of Doctor of Philosophy

Department of Computer Science

University of Essex

June 2006

*This thesis is dedicated to my well-beloved wife, Michele Patrícia.*

# Abstract

Mobile robot applications that involve automated exploration and inspection of environments are often dependant on novelty detection, the ability to differentiate between common and uncommon perceptions. Because novelty can be *anything* that deviates from the normal context, we argue that in order to implement a novelty filter it is necessary to exploit the robot's sensory data from the ground up, building models of normality rather than abnormality.

In this work we use unrestricted colour visual data as perceptual input to on-line incremental learning algorithms. Unlike other sensor modalities, vision can provide a variety of useful information about the environment through massive amounts of data, which often need to be reduced for real-time operation. Here we use mechanisms of visual attention to select candidate image regions to be encoded and fed to higher levels of processing, enabling the localisation of novel features within the input image frame.

An extensive series of experiments using visual input, obtained by a real mobile robot interacting with laboratory and medium-scale real world environments, are used to discuss different visual novelty filter configurations. We compare performance and functionality of novelty detection mechanisms based on the Grow-When-Required neural network and incremental Principal Component Analysis. Results are assessed using both qualitative and quantitative methods, demonstrating advantages and disadvantages of each investigated approach.

# Acknowledgements

First of all, I would like to express my gratitude to my supervisor, Dr. Ulrich Nehmzow, for providing me with the encouragement and guidance needed to pursue a PhD and for being such an inspiring friend. Much appreciation is also given to the constructive criticism and suggestions received from my supervisory board members, Prof. Owen Holland and Dr. Francisco Sepulveda.

I appreciate the friendly scientific environment provided by Dr. Roberto Iglesias Rodriguez and Dr. Theocharis Kyriacou, and the competent and courteous technical support received from Mr. Robin Dowling and Mr. Adrian Simmonds.

I am very grateful to CAPES Foundation and the Federal Technological University of Paraná (UTFPR) for supporting my studies at the University of Essex. I am also indebted to my colleagues at the Electronics Department at UTFPR, who carried on with my duties during the period I was on leave.

The friendship of many fellow PhD students, staff and their families, who I had the privilege to meet at Essex made life much more pleasant. Thank you all. A special token of gratitude goes to Richard Newcombe, for the many inspiring discussions about Computer Vision.

I also thank my family and friends not named above for their constant support across the Atlantic. Finally, conducting this research and writing this thesis would not have been possible without the love, understanding and support of my dear wife, Michele Patrícia.

# Contents

# Chapter 1

# Introduction

The ability to identify perceptions that were never experienced before, referred to as *novelty detection*, is an essential component of intelligent agents aspiring to operate in dynamic environments. Animals, for example, are able to quickly detect and focus their attention in unusual situations by using different sources of sensory information. This sort of competence maximises chances of survival, not only because it helps to reduce threats and exploit opportunities, but also because it enables the animal to learn from experience.

Novelty detection mechanisms and, more generally, attention mechanisms are also extremely important to autonomous mobile robots with limited computational resources. From an operational point of view, the robot's resources can be used more efficiently by selecting those aspects of the surroundings which are relevant to the task in hand or uncommon aspects which deserve closer analysis. By using such mechanisms, previously unknown aspects of the environment can be incrementally learnt by the robot, while already known aspects can be used for the purposes of the desired task.

In fact, identification of new concepts is central to any learning process, especially if knowledge is to be acquired incrementally and without supervision. In order to learn concepts it is necessary to determine first if

they are not already part of the current knowledge base of the agent. Simultaneous learning and recognition (Artač et al., 2002) is a fundamental ability to robots aiming at true autonomy, adaptability to new situations and continuous operation.

From the point of view of applications, reliable novelty detection mechanisms would facilitate automated environment inspection and surveillance tasks using mobile robots. Novelty detection and incremental learning are also vital in applications that demand unsupervised environment exploration and mapping.

In this thesis we are particularly interested in environment inspection using mobile robots to assist in the detection of faults. A practical application is the automatic identification of cracks, tree roots or any other kind of fault in sewer pipes. Sewer inspection is currently performed manually by human operators watching video footage for hours on end, a very tiring and error-prone duty. Human operators would benefit immensely from the assistance of an inspection robot able to highlight just the unusual features in the sewer which are likely to correspond to potential faults.

This type of fault inspection task is different from usual pattern recognition problems in which the features of interest are usually determined beforehand. The aim in fault inspection is to detect unusual entities whose features are likely to be very difficult to be fully determined *a priori*. Therefore, we argue that the most feasible approach to be followed is to learn a model of *normality* of the environment and use it later on to filter out *any abnormal* sensory perceptions — abnormal perceptions are thus defined as anything that does not fit the model of normality. Previous work has demonstrated that the approach of learning models of normality and then using them to highlight abnormalities is very effective for mobile robots that use sonar readings as perceptual input (Marsland et al., 2002a).

Obviously, the sensor modality used as perceptual input plays an im-

portant role in the agent's performance for a given task or behaviour. If relevant features in the environment can not be properly sensed and discriminated, it will be impossible for the agent to respond appropriately. Mobile robots are usually equipped with tactile (bumpers), distance (infrared, sonar and laser range finders) and vision sensors (cameras). From the range of sensors available, vision allows measurement and estimation of several environmental features, some of them exclusive to the visual domain (texture and colour) and others not (shape, size and distance to a physical object, for instance). Therefore, vision is clearly the most versatile sensor modality available. Moreover, vision also has the advantage to provide high resolution readings in two dimensions, making the detection of small details of the environment more feasible.

We argue that it is necessary to use vision as primary source of information for a mobile robot in real world applications such as the sewer fault inspection mentioned earlier. The main reason is that the environment needs to be sensed with high resolution and a two-dimensional field of view, so that the chances of missing important details are minimised. Furthermore, vision is a sense shared with humans and therefore provides common ground for collaboration between robots and human operators while performing the inspection task.

Much of the previous research done in novelty detection applied to environment inspection using real mobile robots was made using exclusively sonar sensing (Crook et al., 2002; Marsland et al., 2002a) and little work was done using monochrome visual input in very restricted ways (Crook and Hayes, 2001; Marsland et al., 2001). There is also work related to novelty detection in simulated robots using sonar readings (Linåker and Niklasson, 2000, 2001). These approaches have the strength of using online unsupervised learning mechanisms to acquire models of normality for the environment. More details about related work and relevant background

for this thesis will be given in Chapter 2.

Marsland, Nehmzow, and Shapiro (2002a) developed the Grow-When-Required (GWR) neural network and used it to highlight novel patterns in sonar scans, while Crook and Hayes (2001) used a novelty detector based on the Hopfield neural network (Hopfield, 1982). Their approaches were qualitatively compared in (Crook et al., 2002) during a novelty detection task using sonar readings in a corridor. Linåker and Niklasson (2000) developed the Adaptive Resource Allocating Vector Quantisation (ARAVQ) network and used it in simulations. All of these approaches work very well with low resolution sonar data according to qualitative assessment criteria. However, none of them was employed using high-resolution visual data in real world application scenarios. Also, there was a lack of quantitative criteria to assess and compare the performance of novelty filters objectively.

Here we are interested in investigating novelty detection using colour visual input in real robots — the particular application in mind is automated visual inspection of sewer-like environments. For that, we use the work of Marsland, Shapiro, and Nehmzow (2002b) on the GWR network as foundation. The GWR network uses a model of habituation, a reduction in response to stimuli that are repeatedly presented, to identify new perceptions and quantify their degree of novelty (Marsland et al., 2000). Perceptions with a high degree of novelty can therefore be incrementally incorporated to the structure of the GWR network.

As mentioned earlier, our interest in using vision for novelty detection purposes stems from the much wider range of information about the environment, in high resolution, that this sensor modality can provide. However, a major difficulty that comes with vision is how to select which aspects of the data are important to be encoded and processed. In mobile robots, it is undesirable to process raw high-dimensional visual data directly due to restrictions in computational resources. Therefore, a natural solution to cope

with massive amounts of visual input (tens of thousands of pixels per image frame) is the use of a mechanism of attention to select aspects of interest and concentrate the available resources on those (Itti and Koch, 2001).

A mechanism of visual attention basically selects interest points within the input image according to some criteria (for instance, edges or corners). Interest points selected by such attention mechanisms are usually locations containing very descriptive information — they are visually salient in the scope of the image frame. A small region in the vicinity of an interest point can then be encoded to represent the local visual features. This process not only localises salient features within the image, but also concentrates computational resources where it is necessary. Local encoding of a small image region also has the advantage of reducing data dimensionality while preserving details.

In the context of visual novelty detection, a particularly interesting attention mechanism is the saliency map model (Itti et al., 1998). This approach combines different visual features (such as intensity, colour and orientation in multiple scales) to obtain a general measurement of saliency for each image location. Saliency can be thought as the property to stand out from the background. This approach is very convenient for novelty detection and, more specifically, inspection tasks in which the identification of uncommon features is precisely what is desired. Also, the use of a model of visual attention is essential to localise *where* the unusual features are in the image. Details about interest point detectors relevant to this work are given in Chapter 2.

In order to be processed efficiently by the novelty filter, the input image needs to be encoded. Here we refer to the general term "novelty filter" to describe any learning mechanism which acquired a model of normality from the environment and is able to use it to filter out abnormal inputs. The purpose of image encoding is to reduce dimensionality of the data to

be processed by the novelty filter while trying to preserve the ability to discriminate between different classes of features as much as possible. It is therefore necessary to devise some sort of internal representation through the encoding of visual aspects in the form of feature vectors. These feature vectors, which are simplified abstractions of the original visual aspects, are expected to describe relevant characteristics and eliminate unnecessary details.

However, designing the image encoding stage is not a trivial job because it is not always clear to the designer which elements of the data are relevant and which are unnecessary. It is therefore desirable to select the important parts of the data using information from the data itself, following a bottom-up approach. An example of such an approach is the Principal Component Analysis (PCA) algorithm, which consists of projecting the data onto principal axes (the axes in which variance is maximised) and selecting the components with larger variances.

The image encoding is also desired to be robust to geometric transformations that result from the fact that images are acquired from a moving platform. As the robot navigates around the environment, visual features are subject to changes in scale, translations and other affine transformations. The visual novelty filter should not classify known visual features as novel just because they were sampled from a different perspective. Also, in the context of the target application of fault inspection, the location of visual features does not determine novelty, although it is important to identify the location of novel features. For example, faults on the walls of a sewer should be always detected in spite of where they appear, but to determine their precise location is also important.

## 1.1 Research Objectives

The main objectives of the research presented in this thesis were:

1. To develop an experimental framework for the investigation of visual novelty detection;

2. To conduct experiments using real mobile robots operating in laboratory and medium-scale real world environments;

3. To devise qualitative and quantitative assessment tools that allow performance comparisons between different visual novelty detection algorithms;

4. To devise image encoding procedures that enable real-time processing* and localisation of novel visual features in the operating environment;

5. To compare and discuss different strategies for the image encoding, visual attention mechanism and novelty filter.

Item 1 in the research objectives is dealt with in Chapter 3, where we introduce a general framework for detection and localisation of novelty in visual data. The approach we follow is to use a mechanism of attention to select a number of salient regions in the input frame, which are encoded into feature vectors and then fed to an unsupervised learning mechanism. The learning mechanism is used to build a model of normality for the perceptions acquired in the environment and, after the learning process, is used as a novelty filter to highlight *arbitrary* novel features that may be encountered in the environment. Figure 1.1 shows a block diagram which summarises the entire process.

*Real-time processing in this context means that the time to process an image frame should not interfere with the efficiency of the robot's navigation behaviour in the environment, *i.e.* we desire the robot to navigate continuously while processing visual data at a frame rate that is high enough to avoid losing any environmental details because of motion.

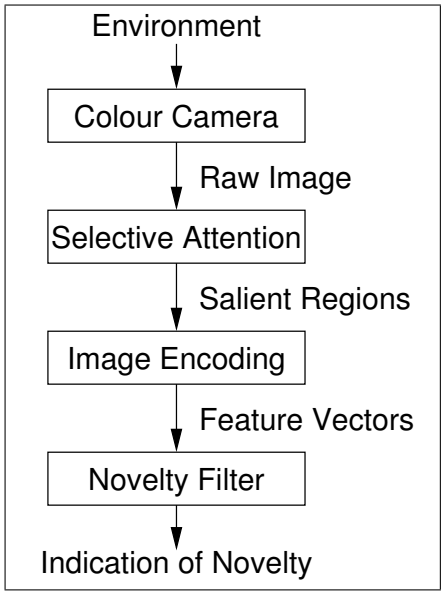



Figure 1.1: The framework for investigation of visual novelty detection: an attention mechanism selects patches from the input image frame, which are then encoded and classified by the novelty filter as novel or non-novel.

Because a model of normality needs to be acquired prior to the use of the novelty filter in inspection tasks, our experimental procedure is divided into two phases. First, exploration of the environment takes place with learning enabled so that the model of normality can be acquired. During the exploration phase, performance of the learning mechanism can be evaluated. After the model of normality is acquired for a particular environment, the trained system can be used in inspection tasks to filter out any abnormal perceptions in that context. By carrying out the inspection in controlled environments, where novel features are deliberately introduced and whose locations are known beforehand, the performance of the system can be objectively assessed.

Still in Chapter 3, we present the mobile robot platform and the environments used in our experimental setup (item 2 in the research objectives). Laboratory environments were built to resemble a corridor or tube-like environment that allowed continuous navigation of the robot while using a simple obstacle avoidance behaviour, also described in detail. Finally, qualitative and quantitative assessment tools, based on contingency table anal-

ysis and statistical tests are presented and discussed (item 3 in the research objectives).

Chapter 4 describes experiments using image encoding techniques based on colour statistics. Performances using both global (entire image frame) encoding and local (attention-based) encoding are compared and discussed (items 4 and 5 in the research objectives). The relevance of the attention mechanism to determine the location of visual features within the image frame is made evident with these experiments, as well as its contribution towards general robustness to translations in the image encoding stage. Contrasting with previous work in novelty detection using visual input, images from the environment were acquired in colour and with no restrictions in field of view.

More experiments are presented in Chapter 5, this time using raw image patches to encode not only colour, but also structural information (texture and shape). Besides producing more specific representation of image features, this approach provides extra functionality to the framework by allowing the reconstruction of the image patches acquired in the model of normality. Hence, visual assessment of which aspects of the environment were actually learnt becomes available to the operator of the robot.

Also in Chapter 5, performances of the GWR network and an alternative novelty filter based on incremental PCA (Artač et al., 2002) are compared and discussed (item 5 in the research objectives). Experiments concerning the influence of the robot's navigation trajectory in the performance of the system are also discussed.

Experiments with visual attention mechanisms are presented in Chapter 6, where different implementations of two interest point detection techniques are evaluated (item 5 in the research objectives). Stability and accuracy in the location of interest points are discussed, as well as the possibility of performing automatic scale selection for regions of interest in the image.

## 1.2 Contributions

The main contributions of this thesis are:

- Implementation and extensive experimentation of visual novelty detection mechanisms with applications in automated inspection using mobile robots. In contrast with previous work done in novelty detection using low resolution sonar readings (Crook et al., 2002; Marsland et al., 2002a) or very restricted monochrome visual input (Crook and Hayes, 2001; Marsland et al., 2001), the work presented here uses colour visual stimuli with unrestricted field of view.

- Development of quantitative assessment methods based on contingency table analysis and statistical tests to support performance comparisons between different visual novelty filters objectively. Such assessment tools were previously lacking in the literature, but were used in our work to contrast the efficiency of the GWR network and the incremental PCA algorithm as novelty filters against manually generated ground truth.

- Demonstration that on-line unsupervised learning mechanisms are able to readily acquire a visual model of normality and later use it to detect novelties introduced in various operating environments correctly, without the installation of any *a priori* models or human intervention, contrasting with other proposed novelty detection algorithms for visual input (Diehl and Hampshire II, 2002; Singh and Markou, 2004).

- Demonstration that the use of attention mechanisms extends the functionality of visual novelty filters to localise *where* the novel regions are in the input frame, while simultaneously improving robustness to geometric transformations due to robot motion, particularly translations and changes in scale. Explicit segmentation of the input image is

avoided, contrasting with the approach followed by Singh and Markou (2004).

- Demonstration that the use of raw image patches adds the functionality of image patch reconstruction from the acquired model of normality, allowing the user to perform a visual evaluation of which aspects of the environment were actually learnt.

The next chapter presents related work in novelty detection and visual attention models, also providing technical and implementation details of the algorithms used throughout this thesis. Emphasis is given on unsupervised clustering methods capable of on-line learning to be used as novelty filters and interest point detectors to be used as visual attention mechanisms.

# Chapter 2

# Related Work and Background

**Learning models of normality.** Novelty detection has been used in numerous problems, from medical diagnosis of masses in mammograms (Tarassenko et al., 1995) to fault monitoring and detection (Taylor and McIntyre, 1998). These problems share a common characteristic, which is the existence of large amounts of data in which the result of the test is negative (no disease diagnosed or no fault detected), and relatively few examples of the important features that have to be detected. It is therefore usually not possible to install or learn models of *abnormality*, because too little training data is available, if any (in some cases, one often does not know even what to look for). Instead, a model of *normality* is acquired and used to filter out any input stimulus that does not fit the learnt model (Marsland, 2003).

The implementation of novelty detection systems is usually based on statistical approaches (Markou and Singh, 2003a) or artificial neural networks (Markou and Singh, 2003b). In either case, a model of normality is built and used to filter out any previously unobserved situation. From an application point of view, an important aspect that distinguishes different novelty detection mechanisms is their ability to perform on-line learning. In mobile robotics, on-line learning is of particular interest for applications that demand simultaneous learning and recognition (Artač et al., 2002).

Identification of previously not perceived or uncommon information is central to unsupervised self-organising learning mechanisms and, in fact, is fundamental to any agent aiming at true autonomy, continuous operation and adaptability to new situations.

A typical task for mobile robots is to navigate and self-localise in their operating environment as a basis to other, more complex, tasks. In order to do this, one can resort to identify distinguished landmarks and use them to build an internal model of the robot's surroundings. The most useful landmarks are the perceptions that differ from the common observations from the environment and therefore can be used to determine the robot's location. When dynamic environments are concerned, which is the case of the most interesting applications, unusual perceptions also need to be identified so that the robot can react accordingly. These unusual observations in the environment, be it static or dynamic, can be conveniently highlighted by a novelty filter and then used by higher levels of processing.

**Low resolution sensor input.** On-line learning of landmarks in a mobile robot's environment through novelty detection was successfully done in the past by using mostly sonar sensor readings as perceptual input (Crook et al., 2002; Marsland et al., 2000). In these approaches, real robots identified and learned landmarks from the environment through unsupervised learning mechanisms. Their goal was to use the learnt model to identify changes (novel landmarks) in the environment, such as open doors in a corridor. Using the same idea and training several distinct novelty filters for different environments also made possible to identify in which environment the robot was operating. In this case, the novelty filter that produced less novelty indications was the most likely to correspond to the particular environment (corridor) in which the robot was operating at the moment (Marsland et al., 2002a).

A different approach used to identify landmarks is to perform change

detection in the flow of a robot's sonar readings while it navigates in the operating environment. Changes in the flow of sensory patterns according to a moving average window enable the robot to identify and learn local environmental configurations such as corners, walls and corridors. This approach was tested in simulated robots by Linåker and Niklasson (2000) and was also used to identify the robot's operating environment (Linåker and Niklasson, 2001).

The strength of the approaches described above is that they are based on unsupervised on-line learning, allowing the robot to acquire a model of normality for the operating environment without human intervention. Nevertheless, a major disadvantage is the use of unreliable low resolution sonar sensor readings (Crook et al., 2002; Marsland et al., 2000). An additional drawback concerning the work of Linåker and Niklasson (2000, 2001) is that only results using simulated robots are presented and discussed.

**High resolution sensor input.** The noisy low resolution sensory perceptions provided by sonars pose serious limitations for real world scenarios. In real world applications, more reliable sensors are necessary to provide higher resolution and multi-modal information. The use of artificial vision gives the extra information needed to operate in situations where distance measurements do not suffice. Cracks, stains or graffiti on walls, for example, cannot be detected by distance measuring devices and yet constitute important features to be found in automated inspection or surveillance applications. Because we are particularly interested in inspection applications, vision is going to be our sensor modality of choice to perform novelty detection.

Some previous work in novelty detection using visual input and mobile robots was conducted by Marsland, Nehmzow, and Shapiro (2001) and used to inspect corridors. They have used a wall-following behaviour and a monochrome camera positioned to acquire close-up images of the wall being followed. The experimental setup used in this approach implicitly

constrained the visual input to a very limited field of view and therefore restricted visual features mostly to the texture of the wall. Because the camera was positioned to take close-up images, only a small portion of the wall could be inspected.

A similar approach was employed by Crook and Hayes (2001), who also used a wall-following robot equipped with a camera that acquired close-up images of the wall being followed. Although the camera was able to sense colour, the pre-processing stage selected a particular colour (orange), effectively resulting in monochrome image sensing. The input space of the experiments was restricted to a "gallery" of pictures of simple geometric arrangements. Neither Marsland, Nehmzow, and Shapiro (2001) nor Crook and Hayes (2001) investigated the performance of different image encoding procedures thoroughly, using only oversimplified techniques to represent perceptual stimuli.

Diehl and Hampshire II (2002) have implemented a real-time novelty detection mechanism for video surveillance using a real mobile robot. Their method was based on the extraction of monochrome spatial features in image sequences to represent moving objects. A classifier based on support vector learning was then trained off-line with examples of moving people and moving cars, being used later to bring different moving items — such as bicycles, vans and trucks — to the attention of a human operator. The human operator was then responsible for labelling the new classes generated by the novelty filter. Their main objective was to allow the human operator to provide additional classes of moving items to the model of normality in order to adapt to contextual changes, while avoiding the process of reviewing and relabelling large portions of the available data. The update of the model of normality, however, could only be made off-line.

More recently, a new framework for novelty detection applied to region-segmented outdoor scenes in video sequences was proposed by Singh and

Markou (2004). Their approach makes use of a feature-selection mechanism for a large set of visual features in order to encode segmented image regions and a feedforward neural network as a classifier. Training of the classifier was performed using still images and an off-line supervised learning strategy. The trained classifier was then used to rejected any encoded input that was not present in the training dataset so that the network could be retrained off-line. Like in (Diehl and Hampshire II, 2002), their main objective was to allow the user to append additional classes to the model of normality without having to relearn all the available data from scratch.

The last two novelty detection methods that were described use unrestricted visual input from outdoor environments, opposed to the former approaches, which used just close-ups of walls. The image encoding procedures employed in (Diehl and Hampshire II, 2002; Singh and Markou, 2004) are much more elaborate in order to cope with the various geometrical transformations that unrestricted visual stimuli are subject to when acquired from a moving camera. The classifier proposed in (Singh and Markou, 2004) yielded excellent results. However, off-line supervised training makes the proposed classifier unattractive to mobile robotics applications in which *continuous* on-line learning and adaptation are often desired (*e.g.* for map building and navigation in dynamic environments).

**This thesis.** In this thesis we are interested in performing on-line unsupervised novelty detection using unrestricted visual input acquired from real robots. Our approach contrasts with the ones mentioned before because it combines unsupervised continuous learning and colour vision without restrictions in the camera's field of view. Instead of explicit image segmentation as in (Singh and Markou, 2004), we are interested in using mechanisms of selective visual attention to generate candidate regions of interest to the novelty filter. The use of visual attention, which follows a recent trend in the Computer Vision community, is meant to help tackling the problem of

geometrical transformations in the input images because of robot motion. Furthermore, we are interested in qualitatively and quantitatively assessing the performance of different configurations of image encoding mechanisms and novelty filters for the inspection of sewer-like environments.

The following sections describe the relevant techniques used throughout the thesis in two main areas: unsupervised clustering for novelty detection and mechanisms of visual attention (interest point detectors). Details about the motivation and use of each of these algorithm in this work will be given in Chapter 3. Methods regarding image encoding using colour statistics (which are conceptually simpler than the algorithms about to be presented) will be described and discussed later in Chapter 4.

## 2.1 Unsupervised Clustering Mechanisms

Several novelty detection techniques exist in the literature (Marsland, 2003), which are mostly based on statistical (Markou and Singh, 2003a) or neural network classifiers (Markou and Singh, 2003b). The usual approach is to train a classifier with examples of normal data and then use it to filter out any abnormal perception. Depending on the algorithm being used, learning can be performed in a supervised or unsupervised manner, either in batch (off-line) or continuous (on-line) mode.

In this thesis we are interested in unsupervised on-line learning strategies, as we desire to build a model of the robot's environment from scratch, with minimal human intervention. We also desire to have the possibility of continuous adaptation of this model. A natural choice of algorithm for this task is the Self-Organising Map (SOM) originally proposed by Kohonen (1984). In fact, many novelty filters proposed in the literature make use of the SOM (Taylor and McIntyre, 1998; Ypma and Duin, 1997).

A new approach for novelty detection based on a model of habituation was proposed by Marsland, Nehmzow, and Shapiro (2000). The use of

habituation, a reversible response reduction to repeated stimuli, allows not only to detect new perceptions but also to quantify their degree of novelty.

Habituation for novelty detection purposes was initially tested with the Habituating Self-Organising Map (HSOM) (Marsland et al., 2000) and eventually resulted in the development of the Grow-When-Required (GWR) neural network (Marsland et al., 2002a), which has a constructive architecture that allows the addition of new concepts to the model as they are presented during training.

The GWR network was successfully used in mobile robotics with sonar readings. In contrast to this, here we investigate its capability to work with visual data. To make this possible, it is necessary to encode the input images appropriately in order to reduce data dimensionality while preserving the ability to discriminate between different visual features. The complex task of image encoding is further discussed in Chapter 3.

### 2.1.1 The Grow-When-Required Neural Network

The Grow-When-Required network (Marsland et al., 2002a,b), which constitutes the basis of our visual novelty detection framework, is a self-organising neural network based on the same principles as Kohonen's Self-Organising Map (Kohonen, 1984). It is composed of nodes that represent the centres of clusters (model weight vectors) in input space — every time that an input is presented, each network node will respond with higher or lower activity depending on how good its weight vector matches the input vector.

Figure 2.1 shows a schematic representation of the GWR network, which basically consists of a clustering layer of nodes and a single output node. The connecting synapses to the output layer are subject to a model of habituation, which is a reduction in behavioural response to inputs that are repeatedly presented. In other words, the more a node in the clustering layer fires, the less efficient its output synapse becomes.

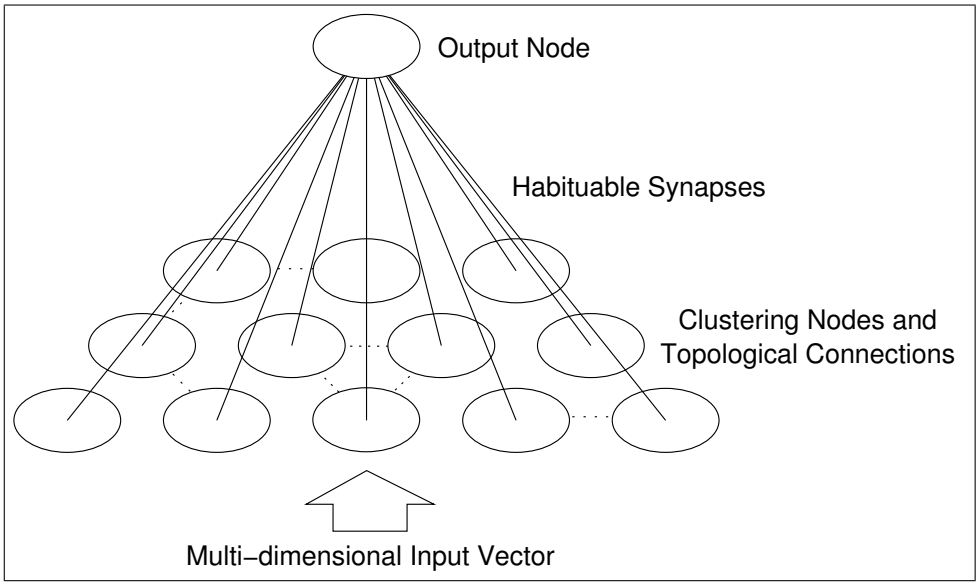



Figure 2.1: The GWR neural network architecture: A clustering layer responds to a multi-dimensional input vector using a winner-take-all strategy. The clustering nodes are connected to the output node through habituable synapses, whose efficacies decrease with repetitive stimulation. This characteristic enables the output node to report the degree of novelty for a given input stimulus. The dotted lines represent topological connections between neighbouring nodes in the clustering layer.

Learning is performed using a winner-take-all approach, which means that only the best matching node responds to a given input, inhibiting all other nodes. The match of the winner node to the given input vector is then reinforced and, in addition, topological information is used to adapt neighbouring nodes, although to a lesser extent than the winner. This learning mechanism is very similar to the one used in the SOM.

What makes the GWR network superior to the SOM is its ability to add nodes to its structure — hence the name Grow-When-Required — by identifying new input stimuli through the habituation model. Given an input vector, both the winner node's activity and habituation are used to a determine if a new node should be allocated in order to represent the input space better.

The habituation rule of a clustering node's output synaptic efficacy is given by the following first-order differential equation:

$$\tau \frac{dh(t)}{dt} = \alpha[h_0 - h(t)] - S(t), \tag{2.1}$$

where $h_0$ is the initial value of the efficacy $h(t)$, $S(t)$ is the external stimulus, $\tau$ and $\alpha$ are time constants that control the habituation rate and the recovery rate, respectively.

Figure 2.2 demonstrates the dynamics of equation 2.1 for a given stimulus $S(t)$. $S(t) = 1$ causes habituation (reduction in efficacy) and $S(t) = 0$ causes dishabituation (recovery of efficacy). It is important to mention that only habituation was modelled in our implementation. Dishabituation was disabled by setting $S(t) = 1$.

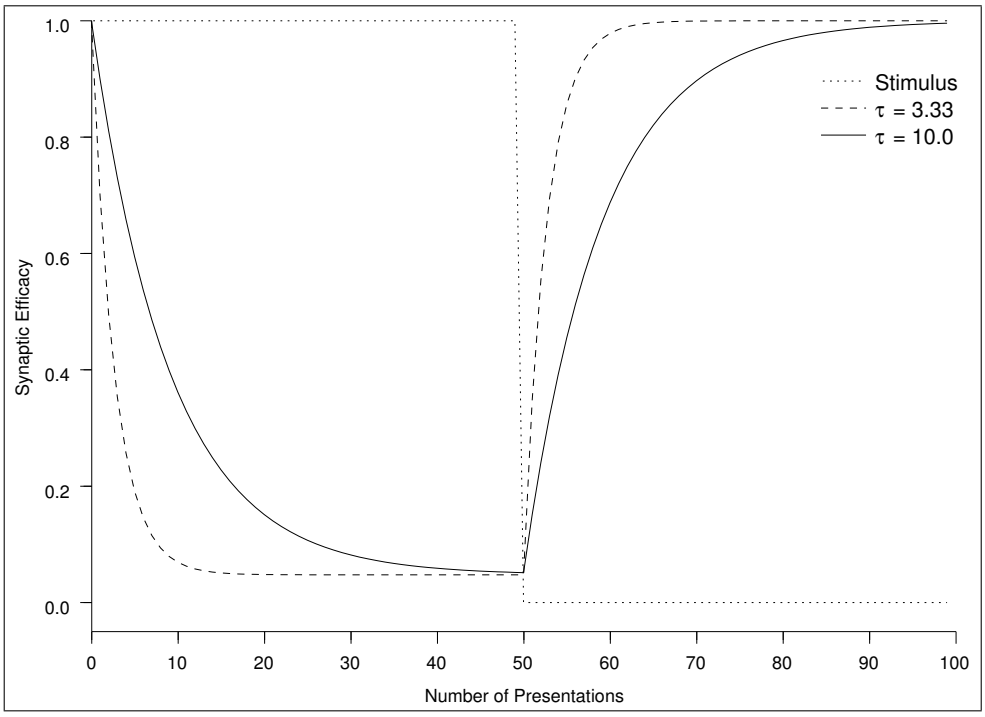



Figure 2.2: Habituation and dishabituation: synaptic efficacy decreases while the stimulus is present ($S = 1$) and increases when the stimulus is removed ($S = 0$). Larger values for $\tau$ result in slower rates of change.

The parameter $\tau$ influences in how fast habituation occurs, as shown in Figure 2.2. The curves were plotted using $\alpha = 1.05$ and $h_0 = 1$, resulting in efficacy values ranging from approximately 0.05 (meaning complete habituation) to 1 (meaning complete dishabituation). As synaptic efficacy has a bounded output, it can be used neatly as a measure of the degree of novelty

for any particular input: higher efficacy values correspond to higher degrees of novelty.

Because the network grows when required, it is first initialised with two completely dishabituated nodes ($c_1$ and $c_2$) in its clustering map $M$:

$$M = \{c_1, c_2\}. \tag{2.2}$$

The weight vectors for the two initial nodes are suggested to be taken randomly from the input distribution (Marsland et al., 2002a). Weight vectors correspond to model input vectors which are stored in the network nodes to represent acquired concepts.

An alternative, used throughout this thesis, is to initialise them with the first two input vectors presented to the network. This approach allows repeatability and fair comparison of results obtained in different experiments. At first there are no topological connections between the nodes and therefore the connection set $C$ is initialised to the empty set:

$$C = \emptyset. \tag{2.3}$$

Training is done with an unsupervised winner-take-all approach, as mentioned before, where the node that best matches the input and its topological neighbours have their output synapses habituated using equation 2.1 and their weights adapted as follows.

The match of the input vector $\mathbf{x}$ to the weight vector of each node in the clustering layer is computed using the Euclidean distance:

$$d_i = \|\mathbf{x} - \mathbf{w}_i\|, \tag{2.4}$$

where $\mathbf{w}_i$ is the weight vector of node $i$, with $i$ covering all the existing nodes in the current map $M$.

The best matching (winner) node is selected as follows:

$$s = \arg\min_{i \in M} \|\mathbf{x} - \mathbf{w}_i\| \tag{2.5}$$

and the second best matching node as:

$$t = \arg \min_{i \in M - \{s\}} \|\mathbf{x} - \mathbf{w}_i\|. \tag{2.6}$$

If there is already a connection between nodes $s$ and $t$, its age is set to zero (the age of a connection corresponds to how many iterations of the algorithm have elapsed since the connection was created), otherwise a new connection is created with age zero:

$$C = C \cup \{(s, t)\}. \tag{2.7}$$

Activation of the winner node is computed using the following radial function:

$$a_s = \exp(-\|\mathbf{x} - \mathbf{w}_s\|^2). \tag{2.8}$$

Both activation and habituation values of the winner node are used to decide whether a given input is considered novel or not. Every time that both activation and habituation values are below predefined thresholds $a_T$ and $h_T$, respectively, a new node $r$ is added to the clustering layer:

$$M = M \cup \{r\}. \tag{2.9}$$

According to the original GWR algorithm, the weight vector $\mathbf{w}_r$ of the new node is set to the average between the winner node's weight vector $\mathbf{w}_s$ and the input vector. However, this approach introduces the new node in a location in input space that does not correspond to a "real" data sample. An alternative strategy is to set the new weight vector $\mathbf{w}_r$ to the current input vector $\mathbf{x}$.

After inserting a new node, it is also necessary to update the network's topological connections by removing the link between nodes $s$ and $t$:

$$C = C - \{(s, t)\}, \tag{2.10}$$

and by connecting node $r$ to node $s$ and to node $t$:

$$C = C \cup \{(r, s), (r, t)\}. \tag{2.11}$$

Cluster centres of the winner node and all of its topological neighbours are adapted according to the following learning rule:

$$\Delta \mathbf{w}_i = \epsilon(\mathbf{x} - \mathbf{w}_i), \tag{2.12}$$

where $\epsilon$ is the learning rate for each of the nodes concerned (the winner and its connected neighbours only).

The implementation of the GWR network training algorithm used in this thesis is slightly different from the original presented in (Marsland et al., 2002a), because we have altered the way learning and habituation are applied to topological neighbours of the winner node. The original approach used fixed values for parameters $\epsilon$ (equation 2.12) and $\tau$ (equation 2.1) for the neighbour nodes, which were a constant fraction of the winner node's parameters. Therefore, $\epsilon$ and $\tau$ of the neighbour nodes were completely independent of the distance between neighbour and winner nodes in input space. In our approach, we made the learning and habituation rates of the neighbour nodes (denoted here by $\epsilon_n$ and $\tau_n$, respectively) proportional to the ratio between winner and neighbour nodes activations:

$$\epsilon_n = \frac{\eta a_n}{a_s}\epsilon, \tag{2.13}$$

$$\tau_n = \frac{a_s}{\eta a_n}\tau, \tag{2.14}$$

where $a_s$ and $a_n$ are the activation of the winner and neighbour nodes, respectively, and $\eta$ is a proportionality factor ($0 < \eta < 1$).

It can be noticed from equation 2.13 that the neighbour nodes will have their weights adapted to a lesser extent than the winner node, while equation 2.14 shows that neighbour nodes will habituate in a slower rate than the winner node. This happens because $a_n < a_s$.

The final step of the GWR training iteration consists of incrementing the age of every existing connection and checking for nodes and connections to be deleted. Nodes that no longer have any neighbours and connections whose age is greater than a predefined threshold $age_{max}$ are all removed.

A summary of the operation of the GWR network is given in algorithm 1.

**Algorithm 1**: GWR network novelty detection

**Input**: current set of nodes $M$, current set of connections $C$, new input vector $\mathbf{x}$.

**Output**: updated set of nodes $M$, updated set of connections $C$, novelty indication $N$.

Find the best and second best matching nodes $s$ and $t$: $s = \arg\min_{i \in M} \|\mathbf{x} - \mathbf{w}_i\|$, $t = \arg\min_{i \in M - \{s\}} \|\mathbf{x} - \mathbf{w}_i\|$, where $\mathbf{w}_i$ is the weight vector of the node $i$.

If there is a connection between $s$ and $t$, set its age to zero, otherwise create it: $C = C \cup \{(s, t)\}$.

Compute the activity of the best matching node: $a_s = \exp(-\|\mathbf{x} - \mathbf{w}_s\|^2)$.

Test if the activity and habituation values of the best matching node characterise novelty:

**if** $a_s < a_T$ *and* $h_s < h_T$ **then**

  Add a new node: $M = M \cup \{r\}$.

  Set the weight vector of the new node: $\mathbf{w}_r = (\mathbf{x} + \mathbf{w}_s)/2$.

  Remove the connection between the best and second best matching nodes: $C = C - \{(s, t)\}$.

  Create connections between the new node and the best and second best matching nodes: $C = C \cup \{(r, s), (r, t)\}$.

  Indicate novelty detected: $N = 1$.

**end**

**else** Indicate no novelty detected: $N = 0$.

Compute the activity of the best matching node's neighbours (nodes with connections to the best matching node): $a_n = \exp(-\|\mathbf{x} - \mathbf{w}_n\|^2)$.

Adapt the positions of the best matching node and its neighbours: $\mathbf{w}_s = \mathbf{w}_s + \epsilon(\mathbf{x} - \mathbf{w}_s)$, $\mathbf{w}_n = \mathbf{w}_n + \frac{\eta a_n}{a_s}\epsilon(\mathbf{x} - \mathbf{w}_n)$.

Age connections to the best matching node: $age_{(s,n)} = age_{(s,n)} + 1$.

Habituate the best matching node and its neighbours: $\tau \frac{dh_s(t)}{dt} = \alpha[h_0 - h_s(t)] - S(t)$, $\frac{a_s}{\eta a_n}\tau \frac{dh_n(t)}{dt} = \alpha[h_0 - h_n(t)] - S(t)$.

Remove any nodes without any neighbours.

Remove any connections with age greater than $age_{max}$.

### 2.1.2 Incremental Principal Component Analysis

Principal Component Analysis (PCA) is a very useful tool for dimensionality reduction that allows optimal reconstruction of the original data, *i.e.* the squared reconstruction error is minimised. It consists of projecting the input data onto its principal axes — the axes along which variance is maximised — and is usually computed off-line because the standard algorithm requires that all data samples are available *a priori*, making it unsuitable for applications that demand on-line learning.

However, a method for the incremental computation of PCA recently introduced by Artač, Jogan, and Leonardis (2002) makes simultaneous learning and recognition possible. Their technique is an improvement to the one originally proposed by Hall, Marshall, and Martin (1998) and allows the original input data to be discarded immediately after the eigenspace is updated, storing only the reduced dimension projected data.

Standard PCA consists in solving an eigensystem for the covariance matrix $\mathbf{C}$ of normalised input vectors $\mathbf{x}_i \in \mathbb{R}^{m\times 1}$, $i = 1 \ldots n$:

$$\mathbf{C}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}, \tag{2.15}$$

$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\mathsf{T}}, \tag{2.16}$$

where $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ is the mean vector, the columns of $\mathbf{U}$ contain the eigenvectors $\mathbf{u}_i$ and the diagonal of $\mathbf{\Lambda}$ contains the eigenvalues $\lambda_i$.

A subspace of up to the original $m$ dimensions is spanned by the eigenvectors $\mathbf{u}_i$, $i = 1 \ldots n$, which correspond to non-zero eigenvalues. However, for dimensionality reduction purposes, one has the option to select only eigenvectors corresponding to the $k$ largest eigenvalues to be included in the eigenmodel. Hence, each input vector $\mathbf{x}_i$ can be projected to some vector $\mathbf{a}_i$ in this $k$-dimensional subspace spanned by the selected eigenvectors:

$$\mathbf{a}_i = \mathbf{U}^{\mathsf{T}}(\mathbf{x}_i - \boldsymbol{\mu}). \tag{2.17}$$

Obviously, the process can be reversed and the input vector can be reconstructed with minimal squared error:

$$\mathbf{x}_i = \mathbf{U}\mathbf{a}_i + \boldsymbol{\mu}. \tag{2.18}$$

For the incremental PCA algorithm proposed in (Hall et al., 1998), it is assumed that an initial eigenmodel is already available and was computed using $\mathbf{x}_i$, $i = 1 \ldots n$ input vectors. The eigenmodel is composed by the mean vector $\boldsymbol{\mu}^{(n)}$, eigenvectors $\mathbf{U}^{(n)} = [\mathbf{u}_j]$, $j = 1 \ldots k$ and eigenvalues $\boldsymbol{\Lambda}^{(n)}$, where the superscript $^{(n)}$ indicates the current iteration of the algorithm.

When a new input vector $\mathbf{x}_{n+1}$ is available, the set of eigenvectors is updated by appending a new orthogonal basis vector and then applying a rotational transformation (Hall et al., 1998):

$$\mathbf{U}^{(n+1)} = \mathbf{U}'\mathbf{R}, \tag{2.19}$$

where $\mathbf{U}'$ is the appended eigenvector set and $\mathbf{R}$ is the rotation matrix.

The new basis vector to be appended is obtained by projecting the new input vector in the current eigenspace using equation 2.17 and then computing the residual vector of its reconstruction (see equation 2.18):

$$\mathbf{r} = \mathbf{x}_{n+1} - \mathbf{U}^{(n)}\mathbf{a}_{n+1} + \boldsymbol{\mu}^{(n)}. \tag{2.20}$$

The normalised residual vector is orthogonal to the current eigenspace and therefore is a natural choice for the new basis vector (Hall et al., 1998):

$$\mathbf{U}' = \left[ \begin{array}{cc} \mathbf{U}^{(n)} & \frac{\mathbf{r}}{\|\mathbf{r}\|} \end{array} \right]. \tag{2.21}$$

The rotational transformation matrix $\mathbf{R}$ is obtained by computing the solution to the following eigenproblem:

$$\mathbf{D}\mathbf{R} = \mathbf{R}\boldsymbol{\Lambda}^{(n+1)}. \tag{2.22}$$

$\mathbf{D} \in \mathbb{R}^{(k+1)\times(k+1)}$ is composed as:

$$\mathbf{D} = \frac{n}{n+1}\begin{bmatrix} \mathbf{\Lambda}^{(n)} & \mathbf{0} \\ \mathbf{0}^\mathsf{T} & 0 \end{bmatrix} + \frac{n}{(n+1)^2}\begin{bmatrix} \mathbf{a}\mathbf{a}^\mathsf{T} & \gamma\mathbf{a} \\ \gamma\mathbf{a}^\mathsf{T} & \gamma^2 \end{bmatrix}, \tag{2.23}$$

where $\mathbf{a} = \mathbf{a}_{n+1}$ (computed using equation 2.17), $\gamma = \mathbf{r}^\mathsf{T}(\mathbf{x}_{n+1} - \boldsymbol{\mu})$ and $\mathbf{0}$ is a column vector $(k \times 1)$ of zeros.

$\mathbf{D}$ can be constructed in alternative ways (Chandrasekaran et al., 1997; Murakami and Vijaya Kumar, 1982), but the method described above is the only one that also takes into account the update of the mean vector (Hall et al., 1998):

$$\boldsymbol{\mu}^{(n+1)} = \frac{n}{n+1}(n\boldsymbol{\mu}^{(n)} + \mathbf{x}_{n+1}). \tag{2.24}$$

Updating the mean vector is important to keep track of the centre of the hyper-ellipsoidal cluster represented by the eigenmodel in the input space. Approaches that disregard the mean vector consider the centre of the cluster to be at the origin, which is obviously not always the case.

In order to enable simultaneous on-line learning and recognition, information about the original input vectors included in the model need also to be stored. The contribution made by Artač, Jogan, and Leonardis (2002) was to develop a method that allows the projected vectors $\mathbf{a}_i$, to be stored and updated, so that the original input vectors $\mathbf{x}_i$ can be discarded. Therefore, besides the mean vector $\boldsymbol{\mu}^{(n)}$ and eigenvectors $\mathbf{U}^{(n)}$, the model also needs to include an additional matrix $\mathbf{A}^{(n)} = [\mathbf{a}_i]$ of projected vectors.

Adding a new eigenvector to the eigenspace results in increasing the dimensionality of the stored projected vectors $\mathbf{a}_i$, which must then be updated as follows:

$$\mathbf{A}' = \begin{bmatrix} \mathbf{A}^{(n)} & \mathbf{a} \\ \mathbf{0} & \|\mathbf{r}\| \end{bmatrix}. \tag{2.25}$$

Instead of solving equations 2.22, 2.23 and 2.24, one can perform standard PCA on the projected vectors ($\mathbf{A}'$), obtaining a mean vector $\boldsymbol{\eta}$ and eigenvector matrix, which corresponds to the desired rotation matrix $\mathbf{R}$ (Skočaj and Leonardis, 2003) to be used for the eigenspace update in equation 2.19.

Finally, the mean vector and the projected vectors are updated by using:

$$\boldsymbol{\mu}^{(n+1)} = \boldsymbol{\mu}^{(n)} + \mathbf{U}'\boldsymbol{\eta}, \tag{2.26}$$

$$\mathbf{A}^{(n+1)} = \mathbf{U}'(\mathbf{A}' - \boldsymbol{\eta}\mathbf{1}), \tag{2.27}$$

where $\mathbf{1}$ is a row vector $(1 \times n + 1)$ of ones.

The algorithm is made completely incremental by initialising the eigenspace and projected vectors as follows: $\boldsymbol{\mu}^{(1)} = \mathbf{x}_1$, $\mathbf{U}^{(1)} = \mathbf{0}$ and $\mathbf{A}^{(1)} = 0$, where $\mathbf{x}_1$ is the first input vector and $\mathbf{0}$ denotes a column vector $(m \times 1)$ of zeros, $m$ being the dimensionality of the input.

In this thesis we use incremental PCA as an alternative method to the GWR network to perform on-line novelty detection. The magnitude of the residual vector — effectively the RMS error between original data and the reconstruction of its projection onto the current eigenspace — is used to decide if a given input is novel and therefore should be added to the model. If the residual vector is above some threshold $r_T$, the corresponding input vector is not well represented by the current model and therefore must be a novel input.

A summary of our implementation of the incremental PCA algorithm as a novelty filter is given in algorithm 2.

**Algorithm 2**: Incremental PCA novelty detection

**Input**: current mean vector $\boldsymbol{\mu}^{(n)}$, current eigenvectors $\mathbf{U}^{(n)}$, current projected vectors $\mathbf{A}^{(n)}$, new input vector $\mathbf{x}$.

**Output**: updated mean vector $\boldsymbol{\mu}^{(n+1)}$, updated eigenvectors $\mathbf{U}^{(n+1)}$, updated projected vectors $\mathbf{A}^{(n+1)}$, novelty indication $N$.

Compute the projection of the new input vector using the current basis: $\mathbf{a} = \mathbf{U}^{(n)\mathsf{T}}(\mathbf{x} - \boldsymbol{\mu}^{(n)})$.

Compute the reconstruction of the new input vector: $\mathbf{y} = \mathbf{U}^{(n)}\mathbf{a} + \boldsymbol{\mu}^{(n)}$.

Compute the residual vector (orthogonal to $\mathbf{U}^{(n)}$): $\mathbf{r} = \mathbf{x} - \mathbf{y}$.

Test if the magnitude of the residual vector is large enough to characterise novelty:

**if** $\|\mathbf{r}\| > r_T$ **then**

  Append residual vector as a new basis vector: $\mathbf{U}' = \begin{bmatrix} \mathbf{U}^{(n)} & \frac{\mathbf{r}}{\|\mathbf{r}\|} \end{bmatrix}$.

  Append projected vector: $\mathbf{A}' = \begin{bmatrix} \mathbf{A}^{(n)} & \mathbf{a} \\ \mathbf{0} & \|\mathbf{r}\| \end{bmatrix}$.

  Perform batch PCA on $\mathbf{A}'$, obtaining its mean vector $\boldsymbol{\eta}$ and eigenvectors $\mathbf{R}$.

  Update projected vectors using the new basis: $\mathbf{A}^{(n+1)} = \mathbf{U}'(\mathbf{A}' - \boldsymbol{\eta}\mathbf{1}_{1\times n+1})$.

  Update eigenvectors: $\mathbf{U}^{(n+1)} = \mathbf{U}'\mathbf{R}$.

  Update mean vector: $\boldsymbol{\mu}^{(n+1)} = \boldsymbol{\mu}^{(n)} + \mathbf{U}'\boldsymbol{\eta}$.

  Indicate novelty detected: $N = 1$.

**end**

**else** Indicate no novelty detected: $N = 0$.

## 2.2 Mechanisms of Visual Attention

Applications that demand real-time processing of large amounts of sensory data, such as vision, using the limited computational resources available to a mobile robot are often challenging. Generally it is not desirable to process the entire image frame delivered by the camera, even when using relatively low resolution, due to processing time constraints. A natural solution to cope with massive amounts of input stimuli is to use a mechanism of attention to select aspects of interest and concentrate the available computational resources on those. Selective attention is widely used in this manner in biological vision systems (Itti and Koch, 2001).

Furthermore, images acquired from a moving robot are subject to several transformations, such as translations, rotations, affine transformations, changes in scale and changes in illumination. Therefore, it is necessary to use encoding mechanisms that are able to generalise and cope with such transformations.

In the past few years many approaches for the selection of interest points were proposed, most of them being bottom-up approaches based on multi-scale pyramidal structures (Burt and Adelson, 1983; Crowley et al., 2002; Greenspan et al., 1994; Lindeberg, 1998; Simoncelli and Werman, 1995). These approaches demonstrate invariance or at least some degree of robustness to many of the aforementioned transformations. One of the most interesting approaches to select points of attention in an image builds a saliency map from different visual features.

The saliency map model for visual attention was proposed by Itti, Koch, and Niebur (1998) and is inspired by the neural architecture of the early primate visual system. Their model combines different modalities of visual features (basically intensity, colour and orientation) at different scales and is in agreement with the Feature-integration Theory of Attention (Treisman and Gelade, 1980). Interest points detected using the saliency map approach were successfully used in a number of applications, including multi-foveated MPEG compression (Dhavale and Itti, 2003) and object recognition (Navalpakkam and Itti, 2003).

In the context of visual novelty detection, the saliency map model is of particular interest. The algorithm that computes the saliency map uses a specially designed normalisation operator for each individual feature modality (*i.e.* intensity, colour and orientation), which renders unusual features more salient than common features in the scope of the image frame. This ability to localise the most unusual visual features in the input image is very useful for the pre-selection of candidate novel regions to be classified later

by a novelty filter, as we shall demonstrate later in our experiments (Chapters 4, 5 and 6). More details about the saliency map and its normalisation operator will be discussed in Subsection 2.2.1.

The use of a salient point detector to select local regions of the input image to be processed by the novelty filter also has the advantage to determine *where* the novel features are in the image frame, as opposed to processing the entire image frame in a global fashion, which only allows to determine *which* image frame contains novelty. Moreover, the use of multi-scale salient points to characterise regions of interest helps the system to cope with geometrical transformations that happen in images acquired from a moving platform. Perhaps the most obvious advantage is that salient points are stable in scale-space and therefore are invariant to image translation and scaling due to robot motion. Hence, the candidate regions selected by the saliency map are not affected by translations and can be made robust to changes in scale.

There are also many other interest point detector algorithms in the literature that are useful for the purposes of selecting candidate regions in the input image (Ferreira and Borges, 2004; Harris and Stephens, 1988; Kadir and Brady, 2003; Kadir et al., 2004; Mikolajczyk and Schmid, 2001, 2002, 2004; Shi and Tomasi, 1994). Among the best-known is the Harris detector (Harris and Stephens, 1988), which was extended by Mikolajczyk and Schmid (2001) to a multi-scale version. The multi-scale Harris detector offers invariance to translation and scaling, and eventually led to the development of the the Harris-affine detector (Mikolajczyk and Schmid, 2002, 2004), which also offers additional invariance to affine transformations.

The multi-scale Harris detector is based on the search for extrema in scale-space, which is a set of images at different resolutions created by convolution with the Gaussian kernel (a Gaussian pyramid). The main advantage of this approach is the fact that it allows automatic scale selection

for the region of interest. Furthermore, it can be implemented in a very efficient way. This interest point detector has been used as part of the Scale Invariant Feature Transform (SIFT) (Lowe, 1998, 2004), a very successful algorithm for object recognition.

In this thesis we concentrate on the use of the saliency map and the multi-scale Harris detector as mechanisms of visual attention, mainly because they both consist of multi-scale approaches that can be efficiently implemented for real-time operation. Here we exploit their ability to provide interest points that are invariant to translation and scaling and we also investigate their suitability for the automatic scale selection of regions of interest (see experiments in Chapter 6).

A detailed description of the scale-space models for visual attention that were implemented and used in this thesis is given next.

### 2.2.1 The Saliency Map

The simplified architecture for the computation of the saliency map model (Itti et al., 1998), which consists in the construction and combination of multi-scale feature maps that allow the detection of local variations in intensity, colour and orientation, is presented in Figure 2.3

The feature maps (intensity, colour and orientation) are computed from Gaussian and Gabor pyramids (Greenspan et al., 1994), which are obtained by successive filtering and subsampling of the input image. The number of levels of these pyramids is limited by the dimensions (width and height in pixels) of the input image, because of the successive subsampling procedure. Basically, the maximum number of subsampling steps, which corresponds to the maximum number of pyramid levels, depends on the dimensions of the image.

Clearly, resolution of the input image defines how well the saliency map is able to detect fine details. However, the higher is the resolution of the

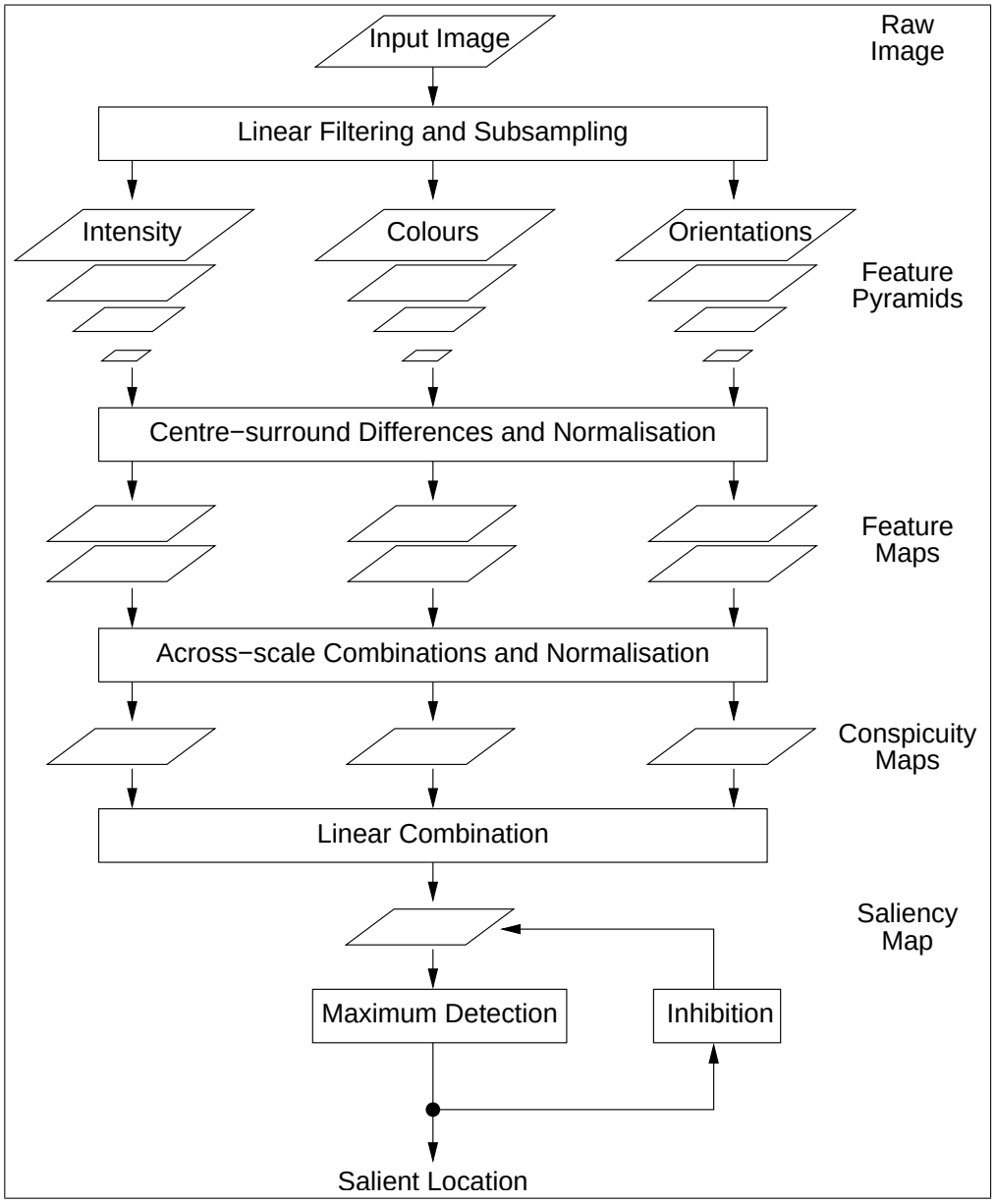



Figure 2.3: The saliency map model architecture: Multi-scale image pyramids are constructed from intensity, colour and orientation features of the input image. Centre-surround differences are computed from the image pyramids to yield feature maps, which are then combined and normalised into a final saliency map.

image, the longer it takes to build the pyramids, compute feature maps and the final saliency map. Therefore, the final number of levels to be used is a trade-off between the desired sensitivity to fine details and the available time to compute the algorithm. In the experiments reported in this thesis (Chapters 4, 5 and 6), five pyramid levels were used with reduction factors ranging from 1:1 (scale 0) to 1:16 (scale 4).

The first step in the extraction of early visual features is to obtain an intensity channel ($I$) from the original red ($r$), green ($g$) and blue ($b$) channels of the input image:

$$I = \frac{r + g + b}{3}. \tag{2.28}$$

After that, intensity normalised channels $\hat{r}$, $\hat{g}$ and $\hat{b}$ are computed in order to decouple hue from intensity, but only at those locations where $I$ is larger than 1/10 of the maximum intensity ($I_{max}$):

$$\hat{r} = \begin{cases} r/I & \text{if } I > I_{max}/10, \\ 0 & \text{otherwise} \end{cases} \tag{2.29}$$

$$\hat{g} = \begin{cases} g/I & \text{if } I > I_{max}/10, \\ 0 & \text{otherwise} \end{cases} \tag{2.30}$$

$$\hat{b} = \begin{cases} b/I & \text{if } I > I_{max}/10. \\ 0 & \text{otherwise} \end{cases} \tag{2.31}$$

Four broadly tuned colour channels for red ($R$), green ($G$), blue ($B$) and yellow ($Y$) are then computed using the following equations:

$$R = \max\{0, \hat{r} - (\hat{g} + \hat{b})/2\}, \tag{2.32}$$

$$G = \max\{0, \hat{g} - (\hat{r} + \hat{b})/2\}, \tag{2.33}$$

$$B = \max\{0, \hat{b} - (\hat{r} + \hat{g})/2\}, \tag{2.34}$$

$$Y = \max\{0, -2(B + |\hat{r} - \hat{g}|)\}. \tag{2.35}$$

The intensity channel $I$ and the broadly tuned colour channels $R$, $G$, $B$ and $Y$ are used to construct the Gaussian pyramids $I(\sigma)$, $R(\sigma)$, $G(\sigma)$, $B(\sigma)$ and $Y(\sigma)$, respectively. $I$ is also used to construct oriented Gabor pyramids $O(\sigma, \theta)$ in the same way as described in (Greenspan et al., 1994). An alternative method to perform fast oriented Gabor filtering is to use

recursive filtering (Young et al., 2000). In our experiments, we used five scales ($\sigma \in \{0 \dots 4\}$), as already mentioned, and four orientations ($\theta \in \{0°, 45°, 90°, 135°\}$). The four oriented Gabor filters at steps of 45° span the 360° of orientation space with more than 99% accuracy, as shown by Greenspan et al. (1994).

Centre-surround linear operations similar to receptive fields found in neurons along the visual pathway of mammals are used to obtain the feature maps and are implemented as the difference between a fine centre scale and a coarser surround scale. Across-scale difference (denoted by $\ominus$) is obtained by bilinear interpolation from the coarse scale to the fine scale and subsequent pixelwise subtraction. In our implementation, the centre consists of pixels at scale $c \in \{1, 2\}$ and the surround of the corresponding pixels at scale $s = c + 2$, as shown in Figure 2.4.

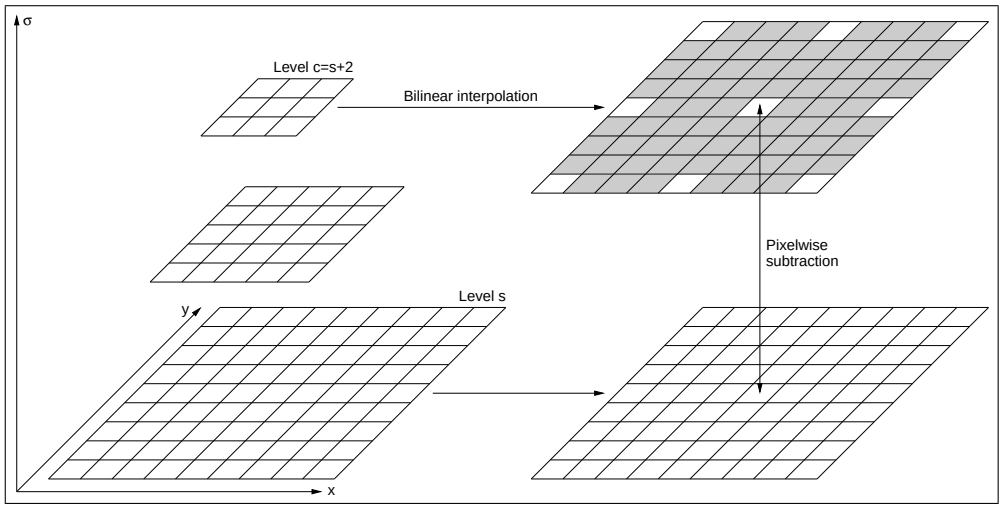



Figure 2.4: The pyramidal across-scale difference $\ominus$: the coarse scale $c$ is interpolated to the fine scale $s$, followed by pixelwise subtraction. Interpolated pixels in the coarse scale are shown in grey.

The first type of feature map is related to intensity contrast, detected by neurons sensitive to bright centres and dark surrounds or vice-versa. Both types of sensitivity are simultaneously obtained by the use of rectification:

$$\mathcal{I}(c, s) = |I(c) \ominus I(s)|. \tag{2.36}$$

The second type of feature map accounts for colour double-opponency, which is detected by neurons whose centres are excited by one colour and inhibited by another, while the opposite excitation relationship holds for the surrounds. Therefore colour feature maps are computed for red/green and blue/yellow double-opponent pairs as follows:

$$\mathcal{RG}(c,s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|, \tag{2.37}$$

$$\mathcal{BY}(c,s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|. \tag{2.38}$$

Finally, the third type of feature map is concerned with local orientation contrast between centre and surround scales. They are computed separately for every orientation, as shown below:

$$\mathcal{O}(c,s,\theta) = |O(c,\theta) \ominus O(s,\theta)|. \tag{2.39}$$

In order to combine feature maps with different dynamic ranges into a single saliency map it is necessary to use a normalisation operator $\mathcal{N}(.)$, otherwise salient features that are strongly present in a few maps may be masked by noise or less salient features that appear more frequently. This normalisation operator ultimately results in giving more weight to unusual features in the input image frame and therefore makes the saliency map an excellent choice for the task of selecting candidate regions to be processed by a novelty filter, as discussed before.

The original normalisation operator suggested in (Itti et al., 1998) involves searching feature maps for local maxima. However, in our implementation we have used a simpler way of normalising feature maps that yields comparable results: first we subtract the average value from the feature map $F$; then, we divide the resulting zero-mean feature map $F_0$ by its vector norm $\|F_0\|$; and finally we take the absolute value of the result. This form of normalisation derives from the colour angular encoding technique described in Subsection 4.3.

The feature maps are combined in three conspicuity maps at scale $\sigma = 2$, one for each feature: intensity $\bar{\mathcal{I}}$, colour $\bar{\mathcal{C}}$ and orientation $\bar{\mathcal{O}}$. The conspicuity maps are obtained by computing across-scale addition (denoted by $\oplus$), which consists of resampling each feature map to scale 2 and subsequent pixelwise addition:

$$\bar{\mathcal{I}} = \bigoplus_{c=1}^{2} \mathcal{N}(\mathcal{I}(c,s)), \tag{2.40}$$

$$\bar{\mathcal{C}} = \bigoplus_{c=1}^{2} [\mathcal{N}(\mathcal{RG}(c,s)) + \mathcal{N}(\mathcal{BY}(c,s))], \tag{2.41}$$

$$\bar{\mathcal{O}} = \sum_{\theta \in \{0°,45°,90°,135°\}} \mathcal{N}\left( \bigoplus_{c=1}^{2} \mathcal{N}(\mathcal{O}(c,s,\theta)) \right). \tag{2.42}$$

Finally, the three conspicuity maps are normalised and added to yield the final saliency map $\mathcal{S}$:

$$\mathcal{S} = \mathcal{N}(\bar{\mathcal{I}}) + \mathcal{N}(\bar{\mathcal{C}}) + \mathcal{N}(\bar{\mathcal{O}}). \tag{2.43}$$

**Biasing the saliency map.** It should be noted that the final saliency map can be biased by giving a higher weight for any particular feature in equation 2.43 (intensity, colour or orientation). By doing this, one can easily render colour features more salient than intensity or orientation features, for example. More weight can also be given to a particular scale and/or orientation in equations 2.40, 2.41 and 2.42. This is an important point because it makes possible to use top-down biasing (Itti and Koch, 1999) if some *a priori* information is available about the features of interest for a given application. For instance, if it is known beforehand that blue vertical lines are important to be detected in a certain inspection task, the saliency map can be easily biased and give more weight to the relevant feature maps (blue-yellow opponent colour and 90° orientation). The saliency map architecture also offers flexibility to be extended in order to include other visual features, such as flicker and motion (Dhavale and Itti, 2003), if the

application requires so.

The highest values in the saliency map correspond to the most salient locations within the input image, which can also be ranked according to their values. We use this information to select regions where further analysis should be performed (classification by the novelty filter, in our case). After that, we inhibit a circular region of the saliency map centred around the current location (by setting saliency values to zero) and search for the next most salient region to be analysed. This process is repeated until a desired number of salient locations is reached.

Location of interest points determined using the saliency map have shown to be robust to geometric transformations, which contributes to the desired general robustness of the image encoding mechanism. In our experiments we have used the nine highest saliency values to indicate which locations of the image are likely to be the most interesting so that further processing for feature encoding could be done in their vicinity.

### 2.2.2 The Multi-scale Harris Detector

An alternative attention mechanism to the saliency map is the multi-scale Harris detector (Lowe, 2004; Mikolajczyk and Schmid, 2001), which is based on the search for extrema in scale-space (Lindeberg, 1998). A fast and efficient algorithm to build a scale-space representation (a Laplacian image pyramid) was proposed by Crowley, Riff, and Piater (2002) and was used in our implementation. In this algorithm, half-octave pyramids are constructed by successive Gaussian filtering, subsampling and subtraction, as shown in Figure 2.5.

The half-octave pyramid algorithm builds simultaneously a Gaussian pyramid and a Difference-of-Gaussian (Laplacian) pyramid through the subtraction of adjacent Gaussian levels before subsampling. Filtering is performed by convolution with separable binomial Gaussian kernels, resulting

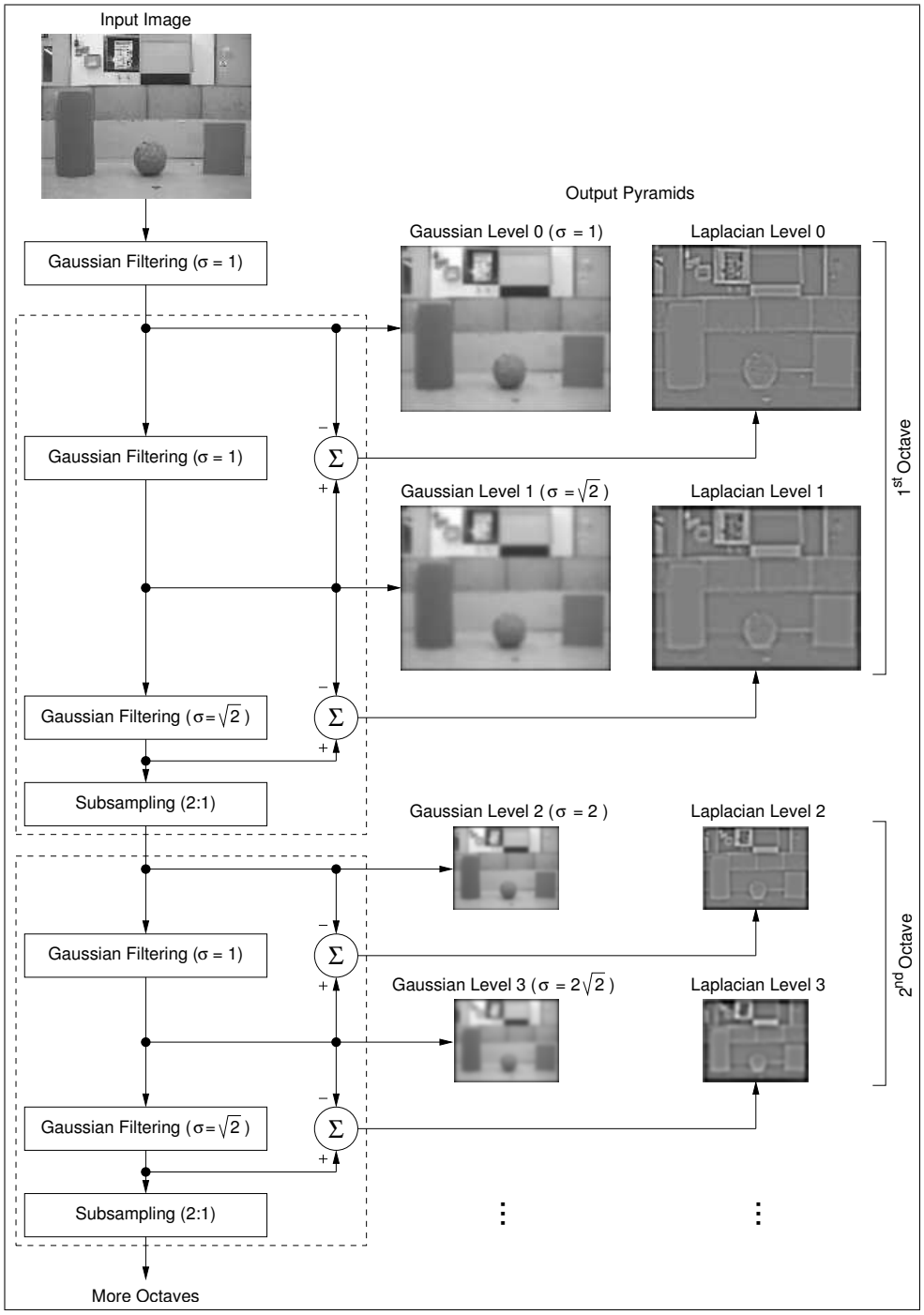



Figure 2.5: The half-octave pyramid construction scheme: The input image is successively Gaussian filtered and subsampled to yield a Gaussian pyramid. Subtraction of adjacent levels of the Gaussian pyramid also yields a Difference-of Gaussian (Laplacian) pyramid.

in a Gaussian pyramid with a scale factor of $\sqrt{2}$ (Crowley et al., 2002). As in the case of the saliency map model, the half-octave pyramid may be built with as many levels as the dimensions of the input image allow (because of the limit imposed by subsampling). More levels result in more resolution in

scale but also mean more processing time. In our implementation, pyramids of ten levels (five octaves) were used.

After the scale-space representation is built, search for extrema occurs in the Difference-of-Gaussian pyramid. In order to do that, each candidate pixel is compared to its eight neighbours in the same pyramid level and its eighteen neighbours in the levels above and below, as shown in Figure 2.6.

The computational cost of this operation is relatively low because most candidate pixels are eliminated within the first few comparisons. It should be noted that depending on which pyramid level is being searched, either the level above has to be subsampled (decimated as in Figure 2.6a) or the level below has to be upsampled (interpolated as in Figure 2.6b) to the same image size of the current level so that pixelwise comparisons can be properly made.

Once an interest point candidate has been found by comparing its pixel value to its neighbours it is necessary to determine its accurate location in scale-space by fitting a three-dimensional quadratic function (a second order Taylor expansion) to the neighbouring samples. According to Lowe (2004), the interpolated location provides a substantial improvement to stability and allows low contrast points to be rejected. The ratio of principal curvatures also makes possible to discard locations poorly localised along edges.

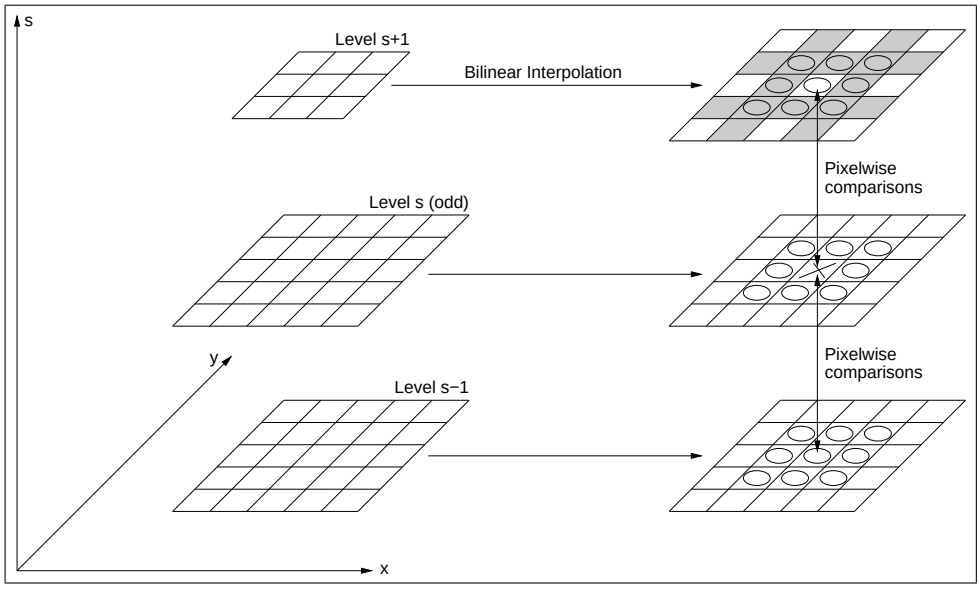
In order to interpolate the location of extrema, their offsets in scale ($\hat{s}$) and in space ($\hat{x}, \hat{y}$) are computed by equations 2.44, 2.45 and 2.46, respectively:

$$\hat{s} = -\frac{f_s}{f_{ss}} = \frac{f(x,y,s-1) - f(x,y,s+1)}{2[f(x,y,s+1) - 2f(x,y,s) + f(x,y,s-1)]}, \tag{2.44}$$

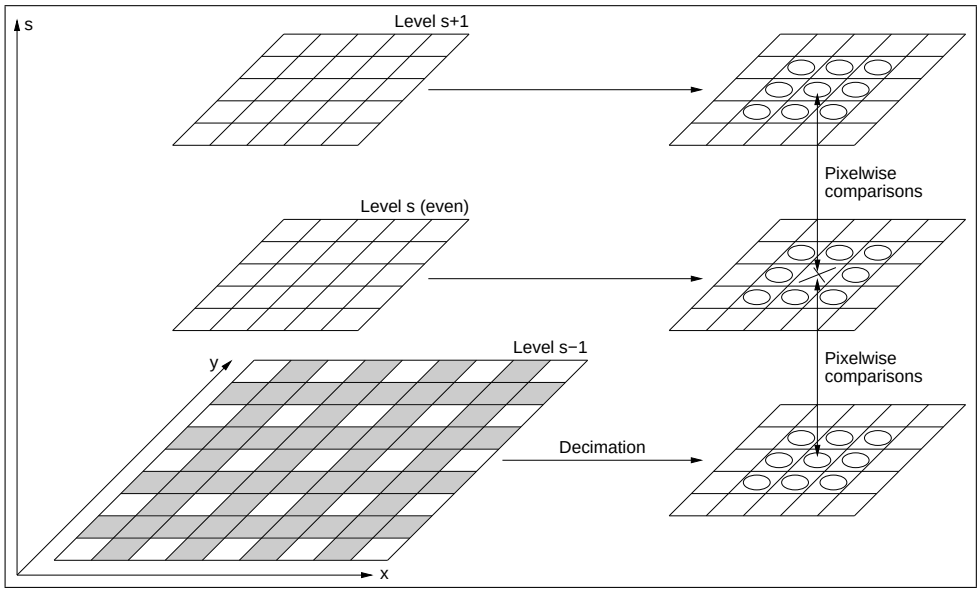
$$\hat{x} = -\frac{f_x}{f_{xx}} = \frac{f(x-1,y,s) - f(x+1,y,s)}{2[f(x+1,y,s) - 2f(x,y,s) + f(x-1,y,s)]}, \tag{2.45}$$

$$\hat{y} = -\frac{f_y}{f_{yy}} = \frac{f(x,y-1,s) - f(x,y+1,s)}{2[f(x,y+1,s) - 2f(x,y,s) + f(x,y-1,s)]}, \tag{2.46}$$

where $f_s$, $f_x$ and $f_y$ are the first partial derivatives of the scale-space function $f$ with respect to $s$, $x$ and $y$, respectively; $f_{ss}$, $f_{xx}$ and $f_{yy}$ are the second partial derivatives; $s$ and $(x,\ y)$ are the coordinates in scale and space, respectively. The final coordinates of the interpolated location are given by $s + \hat{s}$ and $(x + \hat{x},\ y + \hat{y})$.




(a)




(b)

Figure 2.6: Search for extrema in scale-space: (a) in odd-numbered levels, the level above needs to be interpolated; (b) in even-numbered levels, the level below needs to be decimated. Pixels that are interpolated in (a) or decimated in (b) are shown in grey. The current pixel in the search (denoted by $\times$) is compared to all of its neighbours in scale-space (denoted by $\circ$).

The scale-space function value at the interpolated location is calculated using the following equation:

$$f_i = f + f_s\hat{s} + f_x\hat{x} + f_y\hat{y}. \tag{2.47}$$

In our implementation, this information is then used to reject any extrema with $|f_i| < 0.02$ (assuming normalised values in the range [0,1]). For stability, however, it is not enough to discard points with low contrast because the Difference-of-Gaussian function has strong responses along edges, even if localisation is poor and unstable due to noise (Lowe, 2004). Poorly defined extrema have a large principal curvature across the edge but a small curvature in its perpendicular direction.

The spatial principal curvatures at a location in scale-space can be computed from a $2 \times 2$ Hessian matrix:

$$\mathbf{H} = \begin{bmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{bmatrix}. \tag{2.48}$$

The eigenvalues of $\mathbf{H}$ are proportional to the desired principal curvatures, but according to Harris and Stephens (1988) they do not need to be explicitly computed — obtaining their ratio is sufficient. If we let $\alpha$ and $\beta$ be the larger and smaller curvature eigenvalues, their sum and product can be computed from the trace and the determinant of $\mathbf{H}$, respectively:

$$Tr(\mathbf{H}) = f_{xx} + f_{yy} = \alpha + \beta, \tag{2.49}$$

$$Det(\mathbf{H}) = f_{xx}f_{yy} - (f_{xy})^2 = \alpha\beta. \tag{2.50}$$

Defining $r$ as the ratio between the largest and smaller eigenvalues ($\alpha = r\beta$) gives the following relation:

$$\frac{Tr(\mathbf{H})^2}{Det(\mathbf{H})} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(r\beta+\beta)^2}{r\beta^2} = \frac{(r+1)^2}{r}. \tag{2.51}$$

Therefore to verify if the ratio of principal curvatures is below some threshold $r$, all that needs to be done is verify the following condition (a threshold value of $r = 4$ was used in our implementation for the experiments in Chapter 6):

$$\frac{Tr(\mathbf{H})^2}{Det(\mathbf{H})} < \frac{(r+1)^2}{r}. \tag{2.52}$$

As it occurs with the saliency map model, the multi-scale Harris detector finds stable salient locations within the input image frame. However, the multi-scale Harris detector uses a mathematically well-defined mechanism that gives preference to edges with high curvature (corners). An additional functionality of the multi-scale Harris detector is that it can also provide information about the approximate scale (size) of the visual features in the vicinity of detected interest points by using a technique proposed by Lindeberg (1998). Information about size is valuable to the design of scale-invariant image encoding, as the experiments in Chapter 6 will show.

The next chapter discusses how the on-line clustering algorithms and visual attention mechanisms described here fit in our experimental framework for the investigation of visual novelty detection.

# Chapter 3

# An Experimental Framework for Visual Novelty Detection

The challenge in this project is to interface visual data to a known novelty detection technique which has been employed to classify robot perceptions using another sensor modality: distance readings obtained from sonars (Marsland et al., 2002b). Although novelty detection using sonar sensing proved to be useful to detect open doors in corridors (Marsland et al., 2000) and even to identify in which corridor the robot was operating (Marsland et al., 2002a), its very low resolution (a robot's sonar ring is typically composed of a small number of sensors) and unreliable noisy readings pose serious limitations to more demanding real world applications. For example, it would be impossible to detect small cracks in a wall by using sonar sensors alone.

Vision, on the other hand, provides detailed information about the environment in high resolution. Colour, texture and shape are the most obvious visual features, but more elaborate processing can provide information about size, pose, motion and even distance to visual objects. Of course, all of these come at the expense of large amounts of data to be processed, which constitutes a difficulty when one desires real-time operation.

Fortunately, the massive information provided by a vision sensor is highly redundant and therefore can be compressed prior to higher levels of processing. Selecting which aspects of the visual data are the most relevant, however, is not a straightforward procedure and usually is dependant on the application. Visual novelty depends on the multi-modal measures of the properties from the environment that the camera provides the robot with — some visual feature can be considered novel because of its colour, texture, shape, size, pose, motion or any combination of these and even other visual features — a much more complex case than the one of single mode sensors like sonars. Because multi-modal vision is very difficult to be accomplished in a mobile robot with limited resources, in this work we had to decide which visual features were the most important to define novelty in our application domain.

The primary application we had in mind was environment inspection. An example of such an application is sewer inspection, which is currently done by an operator watching video footage in search for cracks and tree roots inside the sewer pipes. This is obviously a very tiring and error-prone task that would benefit enormously from an autonomous mobile robot able to pre-select potential problems — perceptions that differ from perceptions experienced in fault-free sewer pipes — to the attention of the operator.

In the context of an environment such as the inside of a sewer pipe, the visual novelties we are interested in are static (they do not move). Nevertheless, the sewer is a dynamic environment in the sense that new visual features that correspond to faults — cracks and tree roots, for instance — may appear at any time, hence the need of regular inspections. For this type of application, higher level visual interpretations (such as the concepts of size, pose or motion) are not as important as low level features that characterise the essential appearance of visual objects. Therefore, we limited the visual features of interest in this work to colour, texture and shape.

Besides characterising novelty by visual appearance, spatial location of novel visual features in the environment is also important. Therefore, we are interested not only in detecting *which* features constitute potential faults but also *where* they are in the environment. Changes in location of visual features may also constitute novelty and are relevant in some application domains (automated surveillance, for instance). However, in the scope of this work we will not consider the location of a visual feature to be contextually important to determine novelty. In other words, it will not be possible to consider some visual feature as novel based solely on its location in the environment.

Another difficulty related to visual novelty detection using a mobile robot concerns invariance to image transformations. Because the images are acquired from a moving platform, visual features are subject to several geometric transformations and it is undesirable that known visual features happen to be labelled as novel just because there were changes in appearance due to robot movement (*e.g.* changes in perspective). Hence, the image encoding procedure should offer robustness to small geometrical transformations that result from robot motion.

All of these issues have an impact in the way to represent visual information so that it can be adequately processed by the novelty filters described in Section 2.1. The choice of image encoding also has great influence in the system's ability to generalise. This issue is important because if the learning mechanism generalises too much it will hardly ever detect any novelties. On the other hand, too little generalisation would result in frequent erroneous novelty detection.

A minimal configuration for a visual novelty detection system is given in the block diagram in Figure 3.1, in which the raw image acquired by the camera is globally encoded to generate a feature vector to be fed to the novelty filter, which is basically an unsupervised clustering mechanism.

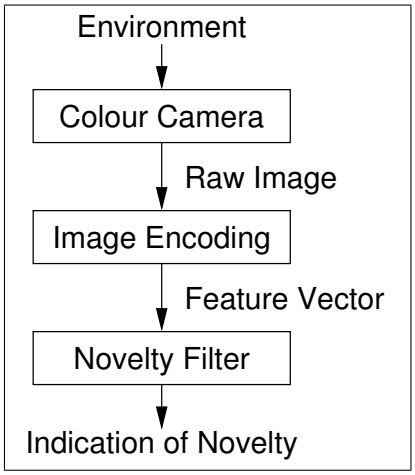



Figure 3.1: Global visual novelty detection framework: the entire image frame is encoded and the resulting feature vector is classified by the novelty filter.

The role of the image encoding stage is important for the correct functioning of the whole visual novelty detection system because the novelty filter itself is just a classifier whose performance depends on how different are the classes generated during image encoding. This is the classic problem of minimising intra-class distances while maximising extra-class distances for classification. Therefore, the main objective of the image encoding stage is to reduce data dimensionality while preserving the ability to discriminate between different classes of features.

In the case of the minimal configuration depicted in Figure 3.1, a global approach is used to encode image information. This global encoding method needs to be robust to the geometrical transformations due to robot motion. In Chapter 4 we report the results of experiments conducted with the minimal configuration using image statistics to encode features (among the choices of features available, we initially opted to use colour alone) to be fed to a novelty filter based on the GWR neural network. Because statistical measurements of image properties generally do not take into account the positions of pixels, they offer many interesting properties concerning robustness to geometrical transformations (translations and rotations), partial occlusions and even deformations in flexible objects (Mel, 1997).

The use of a global image encoding procedure enables the novelty filter

to identify *which* input image frames contain novel visual features. However, if one desires to determine *where* these novel features are localised within the input frame, further encoding and processing of local details in the frame are necessary.

In order to be able to localise novel features within the image frame, we decided to get inspiration from biological vision systems and use a mechanism of attention. This way, smaller image regions, selected by the visual attention mechanism from the input image, can be encoded as feature vectors. Figure 3.2 depicts the block diagram of such an approach, in which the image encoding stage is preceded by the attention mechanism. Instead of a single feature vector for the whole frame, several feature vectors are encoded per image frame using the vicinity of salient image points. Salient or interest points normally correspond to places with high information contents, *i.e.* strong peaks, edges or corners, depending on the criteria for their selection (see Section 2.2).

By selecting interest regions that are salient according to some criteria we reduce the dimensionality of the data to be processed by the novelty fil-

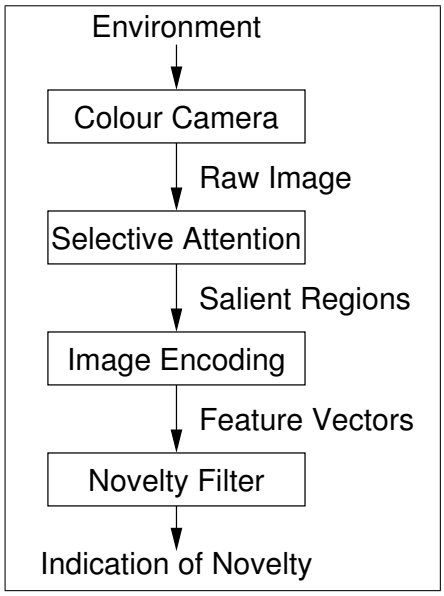



Figure 3.2: Local visual novelty detection framework: an attention mechanism selects patches from the image frame, which are then encoded and classified by the novelty filter.

ter and also gain robustness to some geometrical transformations, notably translations within the image frame due to the robot's movement. Furthermore, novel visual features can be immediately localised within the input image frame with the use of local encoding of salient regions.

If the attention mechanism is able to provide reasonably stable interest points, one can make use of more specific image encoding mechanisms that incorporate not only colour, but also texture and structural information. For example, the relative position of pixels plays an important role when using raw image data — hence the importance of stable and accurate interest points to minimise misalignment of image regions when comparing them.

Experiments with raw image patches extracted from salient locations within the input image frame are reported and discussed in Chapter 5, where we also compare performances of novelty filters based on the GWR neural network and incremental PCA. There is hardly any other visual representation more specific than raw image patches, therefore all the generalisation in these experiments is left to the learning mechanism used as novelty filter. A nice side-effect, though, is that the use of raw image patches allows visual feedback of the knowledge acquired during training.

The influence of the attention mechanism on the overall system performance is investigated in Chapter 6, where issues related to scaling are also discussed. The attention mechanisms reported in this work use multi-scale representations (Gaussian and Laplacian pyramids) to identify salient locations in space and scale. We compare different strategies to select interest points and discuss their advantages and disadvantages, also pointing out alternative methods.

## 3.1 Experimental Setup

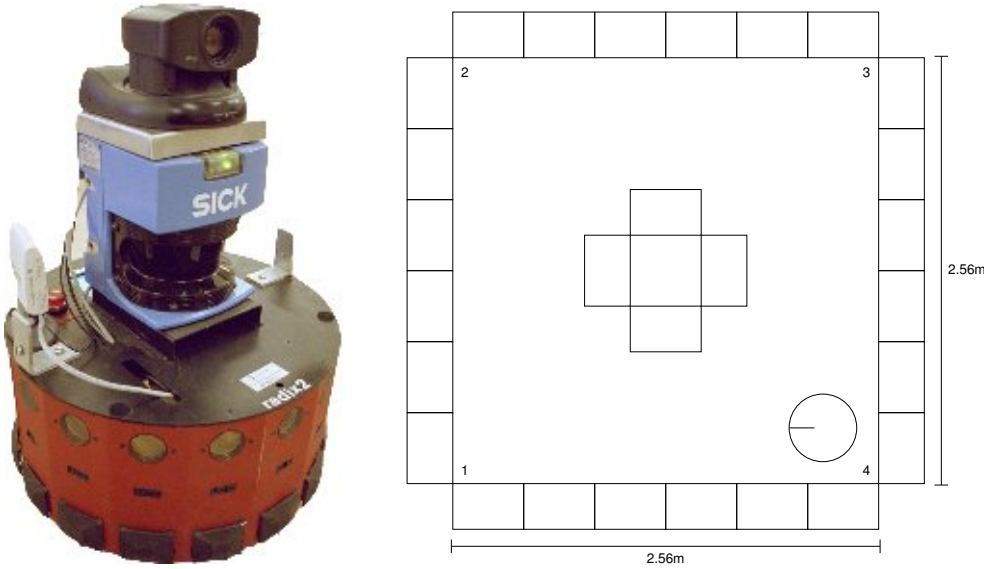
In order to evaluate the ability of the proposed framework to detect novel visual features that may appear in the robot's normal operating environment,

we devised and conducted experiments in controlled scenarios. Every experiment consisted of two stages: an exploration (learning) phase, in which the robot was used to acquire a model of normality of the environment, followed by an inspection (application) phase, in which the acquired model was then used to highlight any abnormal perceptions in the environment.

During the learning phase, images were acquired while the robot was navigating around a "baseline" environment (usually an empty arena or corridor). These images were processed to generate input feature vectors and train the novelty filter. After that, during the application phase, novel objects were placed in the environment so that a new sequence of images could be acquired and used to test the trained novelty filter.

The expected outcome of these experiments was that the amount of novelty detected would continuously decrease during exploration as a result of learning. During the inspection phase we expected that peaks in the novelty measure would appear in areas where a new object had been placed. This hypothesis was tested using a real robot navigating in engineered (laboratory) and medium-scale real world environments. Figure 3.3 shows the experimental setup used for the laboratory experiments.

The colour vision system of *Radix*, the Magellan Pro robot (iRobot Corporation, 2001) shown in Figure 3.3a was used to generate visual stimuli while navigating in the environment. The robot is equipped with standard sonar, infra-red and tactile sensors, and also with an additional laser range scanner whose readings were used for controlling the navigation behaviour. *Radix* operates autonomously, thanks to on-board batteries and computer (850MHz Pentium III processor, 128MB RAM) running Linux. The control software was implemented in C++ (Eckel, 2000; Horstmann, 1996) using iRobot's Mobility Robot Integration Software libraries (iRobot Corporation, 2002) and Robert Davies' Newmat library for matrix operations (Davies, 2002).




(a) (b)

Figure 3.3: Experimental setup: (a) Magellan Pro mobile robot; (b) top view of a typical operating environment, delimited by cardboard boxes shown as rectangles. The robot is represented by a circle with a line that indicates the front.

The robot's on-board computer was capable of processing on-line up to eight frames per second when running our control software, which was optimised for speed. Nevertheless, the images used in the experiments reported here were acquired at one frame per second (without stopping the robot) for off-line processing. This procedure was chosen in order to allow fair performance comparisons between different image encoding techniques and novelty detection mechanisms by using the same datasets.

Figure 3.3b also shows the top view of the engineered environment used in most of the laboratory experiments, a square arena delimited by cardboard boxes in whose corners (numbered from 1 to 4) novel objects were introduced. The cardboard boxes at the borders of the arena acted as walls (approximately 0.5m high) that limited the robot's trajectory and also its visual world. With the sole intention of obtaining a completely controlled visual world for the experiments, the images were acquired with the robot's camera tilted down to $-25°$, so that the field of view was constrained to the arena's walls and floor.

## 3.2 Navigation Behaviour

A simple obstacle-avoidance algorithm using the robot's laser range scanner measurements was used as the navigation behaviour for the robot. A force-field strategy was used, in which every distance measure covering 90° in front of the robot was weighted to act as a virtual spring that pushed the trajectory to the freest space in the environment. The robot's motion control was governed by translational and rotational velocities.

The translational velocity $v_t$ (in m/s) was given by:

$$v_t = \begin{cases} v_{t,nom} & \text{if } \min\limits_{\pi/4 \leq \theta \leq 3\pi/4} (d_\theta \sin\theta) > d_{min} \\ 0 & \text{otherwise,} \end{cases} \tag{3.1}$$

where $v_{t,nom} = 0.15$m/s is the nominal translational velocity, $d_{min} = 0.5$m is the minimum distance allowed to an obstacle and $d_\theta$ is the distance read by the laser range scanner in the direction of $\theta$ (ranging from 45° to 135°, the 90° corresponding to the front of the robot).

The rotational velocity $v_r$ (in rad/s) was given by:

$$v_r = \frac{n}{k_r} \sum_{\theta=\pi/4}^{3\pi/4} \frac{cos\theta}{d_\theta}, \text{limited to the interval } [-v_{r,max}, v_{r,max}], \tag{3.2}$$

where $n$ is the number of iterations since the robot last moved forward ($n = 1$ if the robot is moving forward; $n$ is incremented every time the condition in equation 3.1 repeatedly results in $v_t = 0$), $k_r = 100$ is an empirical gain constant that allows smooth turning and $v_{r,max}$ is the maximum rotational velocity allowed, set to $35°/s$. A negative value for $v_r$ results in clockwise rotation.

Basically, the robot slowly moved forward at 0.15m/s ($v_{t,nom}$) until it found an obstacle within a threshold distance of 0.5m ($d_{min}$), which caused it to stop and rotate at a maximum speed of $35°/s$ ($v_{r,max}$) towards free space again. In our experiments, this behaviour has shown to be very predictable and stable, as will be shown later in Section 5.5.

## 3.3 Assessment of Results

Qualitative and quantitative assessment tools were devised to analyse the performance of different arrangements of the proposed framework (attention mechanism, image encoding and novelty filter) as well as changes in other factors such as the robot's trajectory. These assessment tools are very important in order to establish a reference for comparisons and therefore determine which of the studied methods perform better according to the desired application.

**Qualitative assessment.** In the following chapters we use bar graphs in which novelty measurements provided by the novelty filter are plotted against time. They are used in order to obtain a qualitative indication of performance, in a similar fashion to (Marsland et al., 2002a). In these graphs, time essentially corresponds to a certain position and orientation of the robot in the environment because the navigation behaviour used in the experiments was highly repeatable (see Section 5.5 for details).

As discussed before, during the learning phase of an experiment, the measure of novelty was expected to decline over time as the robot repeatedly explored its environment and progressively acquired a model to represent it. The efficiency of learning during the exploration phase can be graphically assessed through inspection of the qualitative novelty graphs in multiple rounds. For instance, by inspecting the novelty graphs for the exploration phase, one can determine how fast learning occurred and also assess if the amount of learning was adequate for the acquisition of an environmental model of normality.

During the inspection phase, a new object was introduced in the normal environment in order to test the system's ability to highlight abnormal perceptions. The measure of novelty was expected to be high only in places where the new object could be sensed, an expectation that should be

reflected in the novelty graphs obtained. The inspection phase of experiments was also carried out in multiple rounds (usually five) with the learning mechanism disabled, so that unusual features in the environment were highlighted every time that they were perceived. Hence, the consistency of a novel feature being detected in a particular location of the environment but in different inspection rounds can also be evaluated using our qualitative assessment scheme.

**Quantitative assessment.** The off-line processing of image frames acquired with the robot in exploration and inspection phases also allows a quantitative assessment and fair performance comparison between different approaches through the use of the same datasets. For that, we manually generated ground truth in the form of a binary image for each input image where the novel object was present. In these binary images, the pixels corresponding to the novel object were highlighted (see an example in Figure 3.4, where the novel object present in the image is an orange football).

(a) (b)

Figure 3.4: Example of a typical input image containing a novel object (an orange football) and its corresponding ground truth novelty template.

Using the ground truth information, contingency tables were built relating system response to actual novelty status, as shown in Table 3.1. If a given region of the input image has a minimum percentage of highlighted pixels (10% was used in all experiments reported here) in the corresponding region of its respective ground truth template, then this region's novelty sta-

tus is considered as "novelty present". In this case, if the system response is "novelty detected" this configures true novelty and therefore entry $A$ in the contingency table shown in Table 3.1 is incremented; otherwise, if the system response is "novelty not detected", this would configure a missed novelty with entry $B$ being incremented. On the other hand, if the novelty status of a given image region is considered as "novelty not present" (less than 10% of highlighted pixels in the corresponding ground truth region) and nevertheless the system responds as "novelty detected", this configures detection of a false novelty and causes entry $C$ to be incremented. Finally, if the system response agrees with the novelty status by attributing "novelty not detected" to a region whose novelty status is "novelty not present", this represents true non-novelty and entry $D$ is incremented.

Table 3.1: Contingency table for the quantitative assessment of novelty filters.

| | Novelty Detected | Novelty Not Detected |
|---|---|---|
| Novelty Present | $A$ | $B$ |
| Novelty Not Present | $C$ | $D$ |

An ideal association between system response and actual novelty status would have a contingency table in which values $B$ and $C$ in Table 3.1 are zero, while values $A$ and $D$ have non-zero values (in practice, $A$ will be small in comparison to $D$ as usually there are few examples of novel features in the inspected environment). The statistical significance of the association between the actual novelty status (ground truth) and the novelty filter response can be tested using the $\chi^2$ analysis (Nehmzow, 2003; Sachs, 2004). For the $2 \times 2$ contingency table shown in Table 3.1, the $\chi^2$ statistic is computed using:

$$\chi^2 = \frac{N(AD - BC)^2}{(A + C)(C + D)(A + B)(B + D)}, \tag{3.3}$$

where $N = A + B + C + D$ is the total number of samples in the table.

If $\chi^2 > 3.84$ there is a significant correlation between novelty status and novelty filter response, with a probability $p \leq 0.05$ of this statement being wrong. If $\chi^2 > 6.64$ the significance level of the correlation is higher and the probability of being wrong decreases to $p \leq 0.01$. It is also important to mention that the $\chi^2$ test is valid only for well-conditioned contingency tables — this entails the computation of a table of expected values, which must have no entries with expected values below 5 (Nehmzow, 2003).

The strength of the association can be assessed by Cramer's $V$, which is directly based on the $\chi^2$ statistic (Nehmzow, 2003):

$$V = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{(AD - BC)^2}{(A + C)(C + D)(A + B)(B + D)}}. \tag{3.4}$$

The uncertainty coefficient $U$ — an entropy-based measure — can also be used to estimate the strength of the association. Computation of the uncertainty coefficient relies on the fact that each sample in the contingency table shown in Table 3.1 has two attributes, the actual novelty status $S$ and the novelty filter response $R$. The entropy of $S$, $H(S)$, the entropy of $R$, $H(R)$, and the mutual entropy of $S$ and $R$, $H(S, R)$, are given by the following equations (Nehmzow, 2003):

$$H(S) = -\frac{A + B}{N} \ln \left( \frac{A + B}{N} \right) - \frac{C + D}{N} \ln \left( \frac{C + D}{N} \right), \tag{3.5}$$

$$H(R) = -\frac{A + C}{N} \ln \left( \frac{A + C}{N} \right) - \frac{B + D}{N} \ln \left( \frac{B + D}{N} \right), \tag{3.6}$$

$$H(S, R) = -\frac{A}{N} \ln \left( \frac{A}{N} \right) - \frac{B}{N} \ln \left( \frac{B}{N} \right) - \frac{C}{N} \ln \left( \frac{C}{N} \right) - \frac{D}{N} \ln \left( \frac{D}{N} \right). \tag{3.7}$$

When applying equations 3.5, 3.6 and 3.7, one must remember that $\lim_{p \to 0} p \ln p = 0$.

The uncertainty coefficient $U$ of $S$ given $R$, $U(S \mid R)$, is finally computed using equation 3.8 (Nehmzow, 2003):

$$U(S \mid R) = \frac{H(S) - H(S, R) + H(R)}{H(S)}. \tag{3.8}$$

Both $V$ and $U$ provide normalised measures of strength ranging from zero to one. Good associations would result in $V$ and $U$ having values close to one, while poor associations would result in values close to zero. Therefore, the values of $V$ and $U$ can be used to determine which among two or more novelty systems perform better in a given situation.

A further statistic that can be used is the $\kappa$ index of agreement, which is computed for $2 \times 2$ contingency tables as follows (Sachs, 2004):

$$\kappa = \frac{2(AD - BC)}{(A + C)(C + D) + (A + B)(B + D)}. \tag{3.9}$$

This statistic is used to assess the agreement between ground truth data and novelty filter response, in a similar way to what is done with $V$ and $U$. However, it has the advantage of having an established semantic meaning associated with some intervals, as shown in Table 3.2 (Sachs, 2004).

Table 3.2: $\kappa$ intervals and corresponding levels of agreement between ground truth and novelty filter response.

| Interval | Level of Agreement |
|---|---|
| $\kappa \leq 0.10$ | No |
| $0.10 < \kappa \leq 0.40$ | Weak |
| $0.40 < \kappa \leq 0.60$ | Clear |
| $0.60 < \kappa \leq 0.80$ | Strong |
| $0.80 < \kappa \leq 1.00$ | Almost complete |

Unlike $V$ and $U$, the $\kappa$ statistic may yield negative values. If this happens, the level of *disagreement* between system response and manually generated ground truth can be assessed. Negative values occur when the entries $B$ and $C$ in the resulting contingency table are larger than the entries $A$ and $D$. In such a case, both $U$ and $V$ would still result in positive values because they are designed to measure the strength of the association (be it positive or negative) rather than the level of agreement (positive association) or disagreement (negative association).

**Evaluation of the statistical measures.** In order to determine if these statistics were in fact appropriate for the performance assessment of novelty filters, we conducted a simple experiment in which ground truth with probability $s$ of novelty being present was compared against a novelty filter randomly indicating "novelty" with probability $r$. The aim of this experiment was to establish which values of $V$, $U$ and $\kappa$ would result from randomly guessing novelty with a probability $r$ equal to the probability $s$ of novelty actually being present.

Figure 3.5 shows the maximum values (worst cases) obtained for $V$, $U$ and $\kappa$ over a hundred trials, each using a thousand samples. Results are shown for probabilities $r = s$ ranging from 0.05 to 0.95.

It can be noticed in Figure 3.5 that both $V$ and $\kappa$ had their values around 0.10, but always below 0.15 (corresponding to no agreement or weak

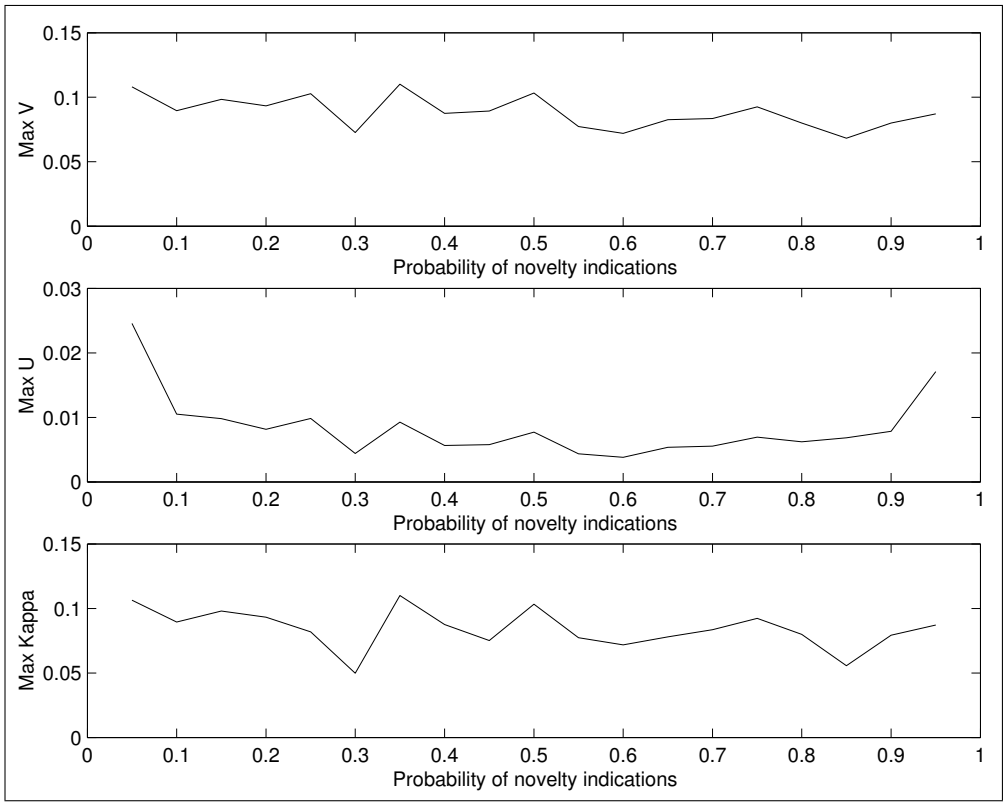



Figure 3.5: Maximum values of $V$, $U$ and $\kappa$ for random ground truth and random novelty filter responses with probabilities ranging from 0.05 to 0.95 ($r = s$). $V$ and $\kappa$ always resulted in values below 0.15, while $U$ always resulted in values below 0.03.

agreement between actual novelty status and novelty filter response). $U$ yielded a U-shaped graph with higher values at the extremes, but always below 0.03. The $\chi^2$ test correctly revealed no statistical significance ($p \leq 0.01$) in the association between novelty status and system response in 99% of the cases. With these results, we concluded that the statistics in question are valid for the assessment of novelty filter performance. Moreover, we can say that any novelty filter whose statistical analysis results in $V > 0.15$, $U > 0.03$ and $\kappa > 0.15$ performs better than random guessing.

**Further evaluation of the statistical measures.** Because the nature of the problem of novelty detection usually implies a much larger amount of non-novel than novel samples, the experiment was repeated with a fixed "novel status" ground truth probability $s = 0.05$ and a "novelty" response probability $r$ ranging from 0.05 to 0.95.

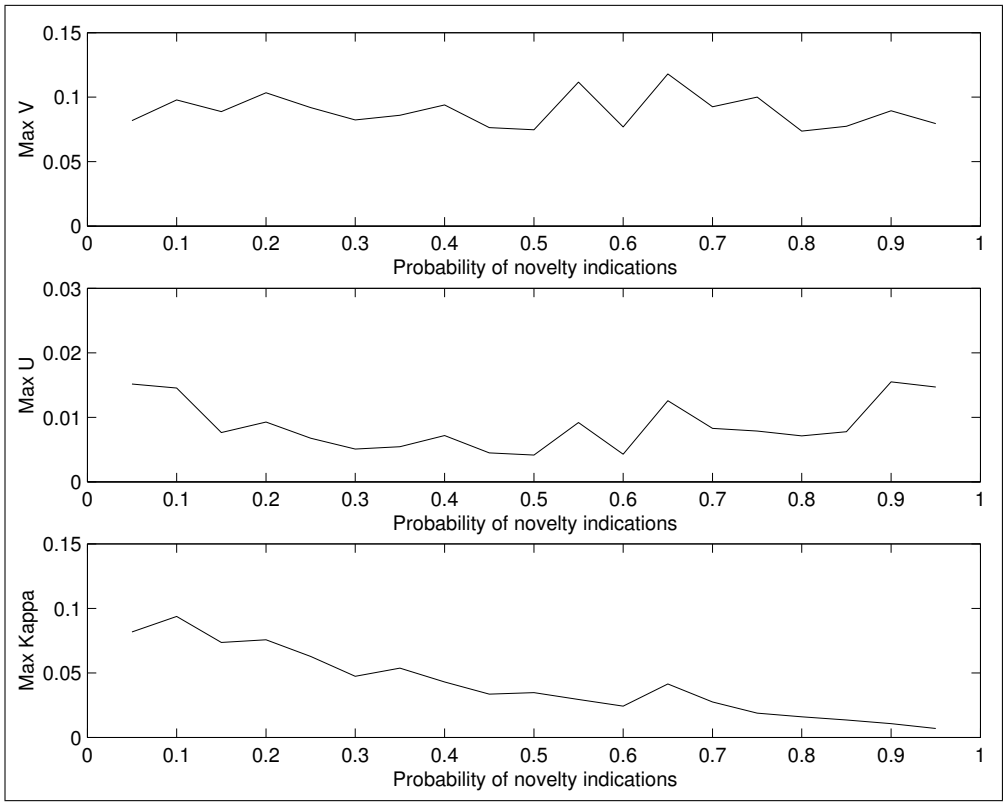



Figure 3.6: Maximum values of $V$, $U$ and $\kappa$ for random ground truth with fixed probability $s = 0.05$ and random novelty filter responses with probability $r$ ranging from 0.05 to 0.95. Only $\kappa$ showed to be sensitive to random guessing probability, but still maintained values below 0.10 (no agreement between novelty filter response and actual novelty status).

The results of this experiment are shown in Figure 3.6, where one can notice that guessing "novelty" randomly with different probabilities had little effect on $V$ and $U$, which maintained the same trend as in Figure 3.5. The $\kappa$ index, on the other hand, showed to be sensitive to the novelty guessing probability $s$, but nevertheless remained below 0.10. These results confirm the hypothesis that $U$, $V$ and $\kappa$ are indeed appropriate tools to assess the performance of novelty filters.

**Problems with the $\chi^2$ test.** One must be careful when using the $\chi^2$ analysis to test the statistical significance in the association between variables in $2 \times 2$ contingency tables. One of the existing problems concerns the number of samples in the contingency table. Consider for example the following contingency table:

$$\begin{array}{|cc|}\hline 10 & 30 \\ 20 & 100 \\ \hline\end{array} \tag{3.10}$$

The analysis for the table above results in $\chi^2 = 1.37$ (no statistical significance in the association between variables), $V = 0.09$, $U = 0.01$, $\kappa = 0.09$ (no agreement between variables). However, if the number of data samples is increased ten times, this results in the following table:

$$\begin{array}{|cc|}\hline 100 & 300 \\ 200 & 1000 \\ \hline\end{array} \tag{3.11}$$

The analysis of this new table yields $\chi^2 = 13.7$ (statistical significance in the association between variables, $p \leq 0.01$), $V = 0.09$, $U = 0.01$, $\kappa = 0.09$ (no agreement between variables).

This effect indicates that statistical significance in the association between variables in $2 \times 2$ contingency tables may result solely from the use of a large number of data samples. In spite of that, values of $V$, $U$ and $\kappa$ are not affected by the number of data samples and continue to be faithful

to the strength of the association between the variables, regardless of the number of data samples in the contingency table.

Another problem illustrated by Sachs (2004, p. 456) concerns the addition of two contingency tables with no statistically significant association between variables, which may yield statistically significant association between variables in the resulting table:

$$\begin{array}{ccccc} \begin{bmatrix} 1 & 10 \\ 10 & 100 \end{bmatrix} & + & \begin{bmatrix} 100 & 10 \\ 10 & 1 \end{bmatrix} & = & \begin{bmatrix} 101 & 20 \\ 20 & 101 \end{bmatrix} \\ \chi^2 = 0 & & \chi^2 = 0 & & \chi^2 = 108 \end{array} \tag{3.12}$$

The problem above refers to adding contingency tables with radically different $A$ and $D$ entries ($A << D$ in one of the terms shown above, while $A >> D$ in the other). Therefore, one should be aware of these limitations of the $\chi^2$ analysis before using it to test statistical significance in the association of variables in $2 \times 2$ tables resulting from addition. In this particular example, the resulting contingency table yields $V = 0.67$, $U = 0.35$ and $\kappa = 0.67$.

# Chapter 4

# Experiments using Colour Statistics

## 4.1 Experiments 1 and 2: Novelty Detection from Global Colour Histograms

For the first experiments using visual input to the GWR neural network, we used image encoding mechanisms based on image statistics. The idea was to use simple and fast image encoding techniques in order to reduce dimensionality of the input data (152 $\times$ 120 pixels), so that it could be processed efficiently by the novelty filter. Such statistical image encoding techniques, mainly in the form of histograms, have been successfully used in the past from content-based image retrieval (Boujemaa et al., 2001; Schmid and Mohr, 1997; Swain and Ballard, 1991) to pattern recognition (Chang and Krumm, 1999; Lowe, 1999, 2004; Mel, 1997; Schiele and Crowley, 1996, 2000) and robot localisation (Gonzales-Barbosa and Lacroix, 2002).

The main advantage of histograms is that when applied to image features, they show robustness against geometric transformations, changes in perspective and partial occlusion (Schiele and Crowley, 2000), all of which are of interest due to the fact that input images are acquired from a moving

platform. In fact, conventional image histograms disregard all the information about shape and structure — even if all the pixels in an image are randomly rearranged, its histogram will contain the same information. On the other hand, a clear disadvantage is that it is not possible to reconstruct the original image from its histogram.

In these initial experiments we analysed the performance of colour histograms, without encoding of any other image feature, such as shape or texture, for example. With the intention to separate intensity and colour information, we first converted the images to the HSI (Hue-Saturation-Intensity) colour space from the original RGB (Red-Green-Blue) colour space using equations (4.1), (4.2) and (4.3):

$$I = \frac{R+G+B}{3}, \tag{4.1}$$

$$S = 1 - \frac{\min(R,G,B)}{I}, \tag{4.2}$$

$$H = \arctan\left(\frac{\sqrt{3}(G-B)}{2R-G-B}\right). \tag{4.3}$$

We then divided the hue interval $[-\pi, \pi]$ equally into M regions, by defining the following membership functions $f_m$:

$$f_m = \begin{cases} 1 & \text{if} - \theta < H - (M-2m)\,\theta \leq \theta \\ 0 & \text{otherwise,} \end{cases} \tag{4.4}$$

where $\theta = \frac{\pi}{M}$ and $m = 0, 1, ..., M-1$.

A standard histogram is computed by evaluating the responses of the membership functions $f_m$ for each pixel in the image and adding them to the corresponding histogram bin ($b_m$), as shown in (4.5):

$$b_m = \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} f_m(H_{x,y}), \tag{4.5}$$

where $(x, y)$ are the pixel coordinates, $X$ and $Y$ are the image width and height, respectively, and $m = 0, 1, ..., M-1$.

However, for the colour histograms used in the experiments reported

here, we have also included information about colour saturation by weighting the response of the membership functions as given in equation (4.6):

$$b_m = \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} f_m(H_{x,y})\, S_{x,y}. \tag{4.6}$$

Finally, the weighted histogram was normalised to satisfy the constraint $\sum_{m=0}^{M-1} b_m = 1$.

### 4.1.1 Experiment 1: Global Colour Histograms with 32 Bins

**Experimental setup.** Our initial approach employed the above defined weighted histograms using $M = 32$ bins as input vectors for the GWR-based novelty filter. These histograms were computed in a global fashion, *i.e.* for the entire input image frame, and fed to the GWR network.

A first square arena was built in the Brooker Laboratory at the University of Essex, using wooden panels, cardboard boxes and plastic cylinders. The internal visual appearance of the arena was basically of light yellow hue (floor and lower part of the walls made of wooden panels) and dark blue (upper part of the walls made of cardboard boxes), with a large amount of visible marks on the floor and also on the cardboard boxes. Figure 4.1 shows a view of the arena from the top, where the robot can be seen on the top left corner and an orange football, used as novel object during inspection, can be seen on the bottom right corner.

The cylinders on the centre of the arena constitute an obstacle that forces the robot to navigate around the arena in a closed loop. The idea was to build a laboratory environment that would resemble a corridor or a tunnel, such as sewer pipes or air-conditioning ducts. Figure 4.2 shows an image acquired from the robot's start position.

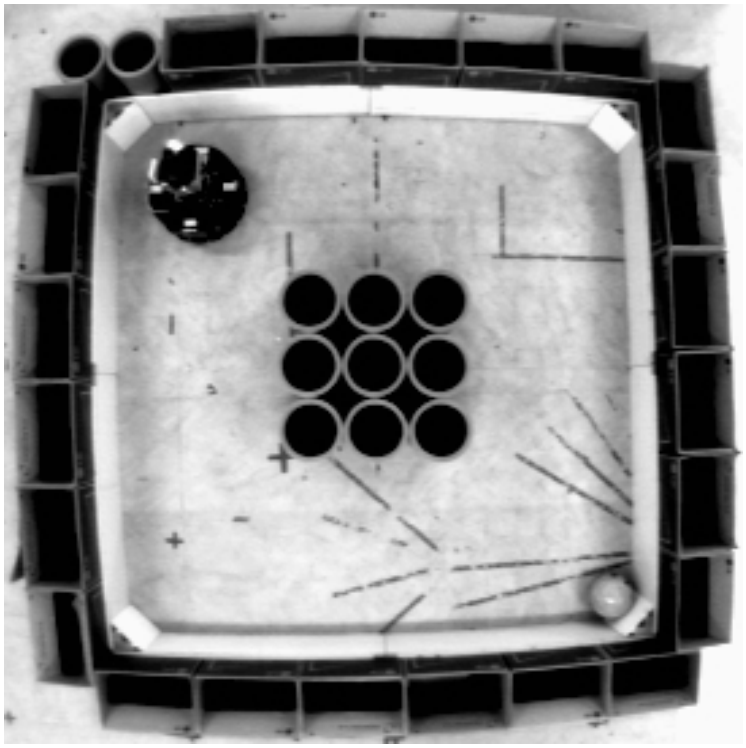


Figure 4.1: Top view from the first arena built in the Brooker Laboratory at the University of Essex. The robot is shown at the top left corner and an orange football, used as novel stimulus during inspection, at the bottom right corner.

Figure 4.2: View of the arena from the robot's start position. The camera was tilted down to -25°, so that the field of view included only the interior of the arena (floor and walls).

**Normality model acquisition.** The experiment started with an exploration phase, when the robot was used to acquire a model of normality of the empty arena. Exploration was conducted in five consecutive loops around the empty arena, with the robot being stopped and repositioned at the starting point in every loop. This procedure was used in order to ensure that the robot's trajectory would be as similar as possible for every loop,

resulting in consistent novelty graphs for qualitative assessment.

Images were acquired at the rate of one frame per second, resulting in a total of 45 images per loop around the arena. A global colour histogram was computed for every acquired image frame and fed as input vector to a GWR neural network with the following parameters: $a_T = 0.9$, $h_T = 0.5$, $\eta = 0.1$, $\epsilon = 0.1$, $\tau = 3.33$, $\alpha = 1.05$, $h_0 = 1$, $S(t) = 1$ and $age_{max} = 20$. These parameters were chosen in order to facilitate the insertion of nodes in the GWR network structure and allow fast habituation to the environment.

As previously explained in Chapter 2, the GWR network parameters $h_T$, $\tau$, $\alpha$, $h_0$ and $S(t)$ control node habituation (equation 2.1 on page 20), while $a_T$ controls cluster size (equation 2.8 on page 22) and $\eta$ controls node adaptation (equation 2.12 on page 23). The amount of habituation and adaptation within the topological neighbourhood of the firing node are governed by $\eta$ (equations 2.13 and 2.14 on page 23).

During the exploration phase, learning of the GWR network was obviously enabled to allow the acquisition of a model of normality. This can be observed in the novelty graphs depicted in Figure 4.3, corresponding to each of the five consecutive loops around the empty arena.

It can be noticed in Figure 4.3, as expected, that the amount of novelty measured — namely the efficacy of the habituable synapse of the winner node for a given input vector — decreases exponentially as the network habituates on repeated stimuli. The range of values for the novelty measure in the vertical axis of the graphs goes from approximately 0.05 to 1 as a result of the choices for the parameters $\alpha$, $h_0$ and $S(t)$. Because the images were acquired at the rate of one frame per second, the horizontal axis of the graphs can also be interpreted as time in seconds. Pictograms indicating the approximate position and orientation of the robot in the arena are also shown in Figure 4.3 (we use the notation "Corner 1+" to indicate position and orientation immediately after the robot has completely turned the first

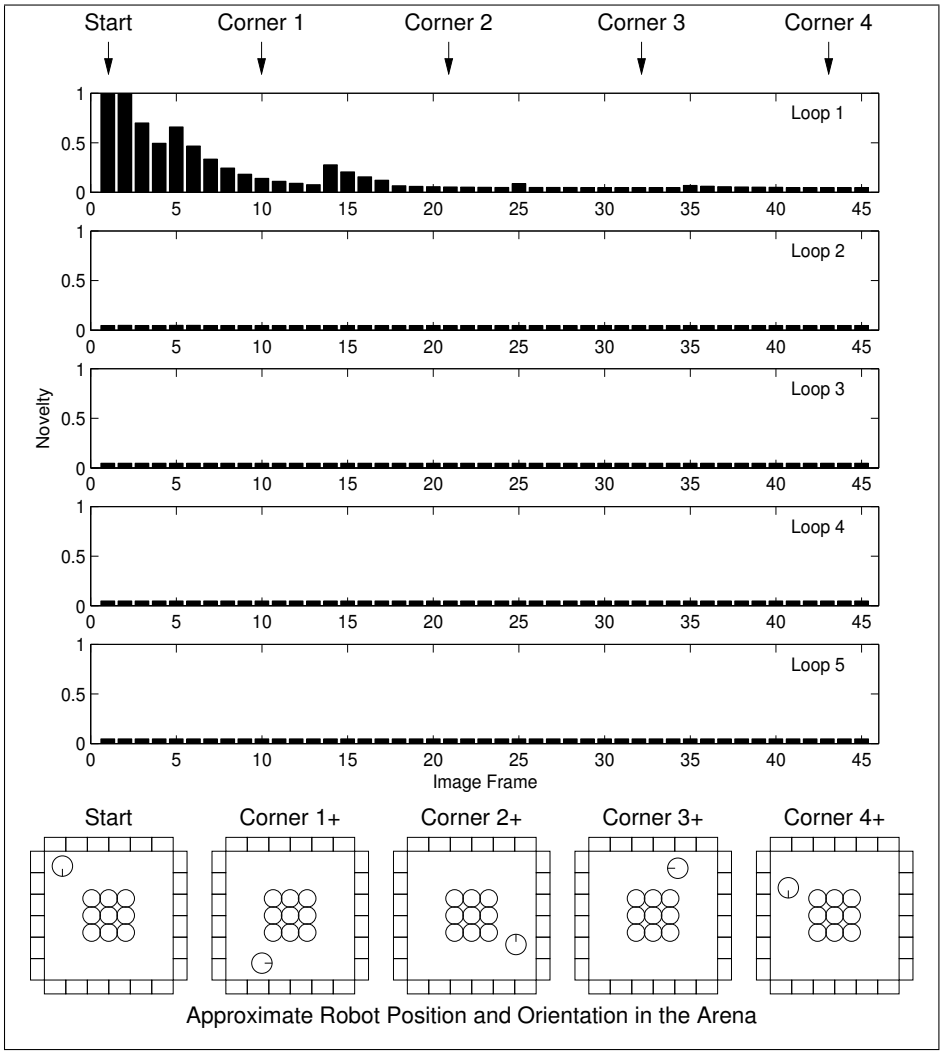



Figure 4.3: Exploration of the empty arena using global colour histograms (32 bins) as image encoding scheme. The graphs show that the amount of novelty decreases exponentially as the network habituates on repeated stimuli. Complete habituation is achieved by the end of the first loop. The pictograms below the graphs indicate the approximate position and orientation of the robot in the arena while performing the exploration loops.

corner). The network was completely habituated after the first loop around the arena. After training, the model of normality acquired by the GWR network had only two nodes, each containing a prototype colour histogram learnt from the explored environment.

**Novelty detection.** Having trained the GWR network during the exploration phase, we then used the acquired model to highlight any unusual visual features introduced in the empty arena: an orange football was placed

as novel object in one of the corners and the robot was used to inspect the arena. The ball was selected not only because it contrasted well with the arena's colour features, but also because it did not interfere with the robot's trajectory around the arena (it could not be sensed by the laser range scanner). Learning of the GWR network was disabled during inspection, so that consistency in novelty indications could be verified over different loops around the arena. The results obtained for the inspection phase of the arena containing the ball are shown in Figure 4.4.

The set of frames where the orange football appeared in the camera's field of view are indicated by dotted arrows on the top of Figure 4.4. These frames correspond to locations where high values for the novelty measure were expected to happen (the ball appeared always in the same frames in every loop because the navigation behaviour was very stable, as will be shown in Section 5.5), but it can be noticed from the presented graphs that this particular experiment has failed in this respect. Not only were the frames containing the orange ball not labelled as containing novel colour features, but also frames with supposedly already known colour features were misclassified by the system as being novel. It is interesting to notice that false novelties were consistently detected in each loop in the same part of the arena, when the robot was turning the first corner immediately before the ball was encountered. Later we discovered that the false novelties detected when turning corners were due to changes in illumination (to be discussed in Subsection 4.2.2).

This experiment yielded an ill-conditioned contingency table for the $\chi^2$ test (the corresponding table of expected values had entries with values below 5, see Section 3.3). Hence, it was not possible to assess the statistical significance in the association between actual novelty status and novelty filter response. The quantitative assessment resulted in $V = 0.08$, $U = 0.02$ and $\kappa = -0.07$ (no agreement between ground truth and system response).

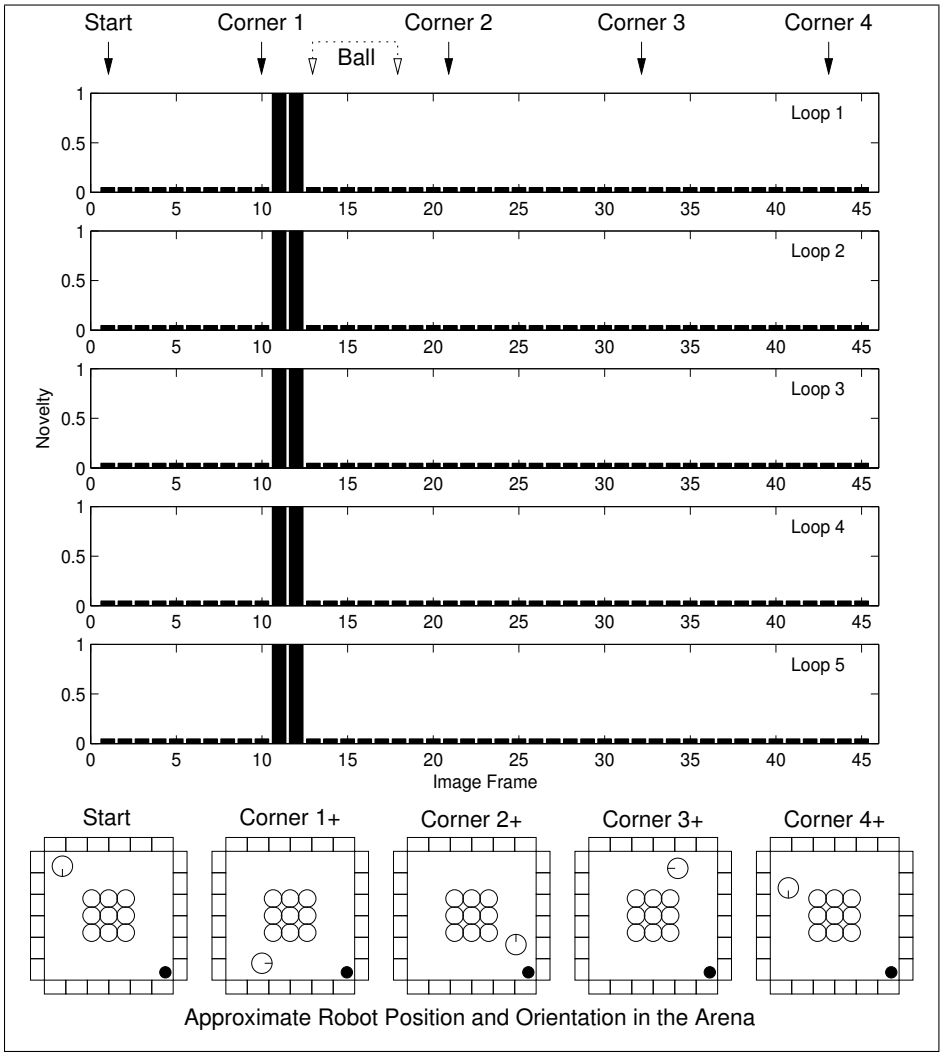



Figure 4.4: Inspection of the arena with an orange football (novel stimulus) using global colour histograms (32 bins) as image encoding scheme. Locations where the ball was in the camera's field of view are indicated by dotted arrows on the top of the figure. The graphs show that the system failed to detect the actual novel stimulus and erroneously labelled known features of the arena as being novel. Pictograms below the graphs indicate the approximate position and orientation of the robot in the arena while performing the inspection loops and also the location of the novel stimulus.

The explanation for the experiment's failure is that the resulting novelty filter is generalising too much, something that could be anticipated by the fact that only two nodes were learnt by the neural network during exploration of the environment — a simple environment perhaps, but not that simple. Possible causes of this over-generalisation could be either a bad choice of parameters for the GWR network or the use of an image encoding that is too poor to represent significant changes in visual features

experimented by the robot.

We rejected the hypothesis that a bad choice of parameters for the GWR network was the cause of over-generalised results because these parameters were actually chosen to facilitate the addition of nodes to the network structure (more nodes acquired means more specificity and less generality from the network's point of view). The combination of parameters $h_T = 0.5$, $\tau = 3.33$, $\alpha = 1.05$, $h_0 = 1$ and $S(t) = 1$ results in a particular node having to fire (*i.e.* be the winner) only three times, due to the dynamics of equation 2.1 (page 20), in order to be considered sufficiently habituated and well positioned in input space. The activation threshold $a_T = 0.9$ provides some room for generalisation in order to handle noise — $a_T$ basically sets the size of the clusters used by the GWR network in input space (see equation 2.8 on page 22). Finally, the learning rate $\epsilon = 0.1$ allows a bit of cluster centre adaptation (see equation 2.12 on page 23), but makes sure that centres will not move too far from their original location in input space.

### 4.1.2 Experiment 2: Global Colour Histograms with 64 Bins

Having made the considerations about the GWR network parameters, we decided to investigate the influence of the image encoding method. A simple way to reduce the generalisation of a histogram-based image encoding scheme is to enlarge its number of bins. Therefore, we repeated our previous experiments but this time using twice as many bins ($M = 64$) for the weighted colour histograms. Figure 4.5 shows the amount of novelty measured in five consecutive loops around the empty arena during the exploration (learning) phase.

Using global colour histograms with 64 bins as input vectors to the GWR network, the model acquired after five loops had four nodes, twice the amount of nodes acquired in the first experiment, indicating that gen-

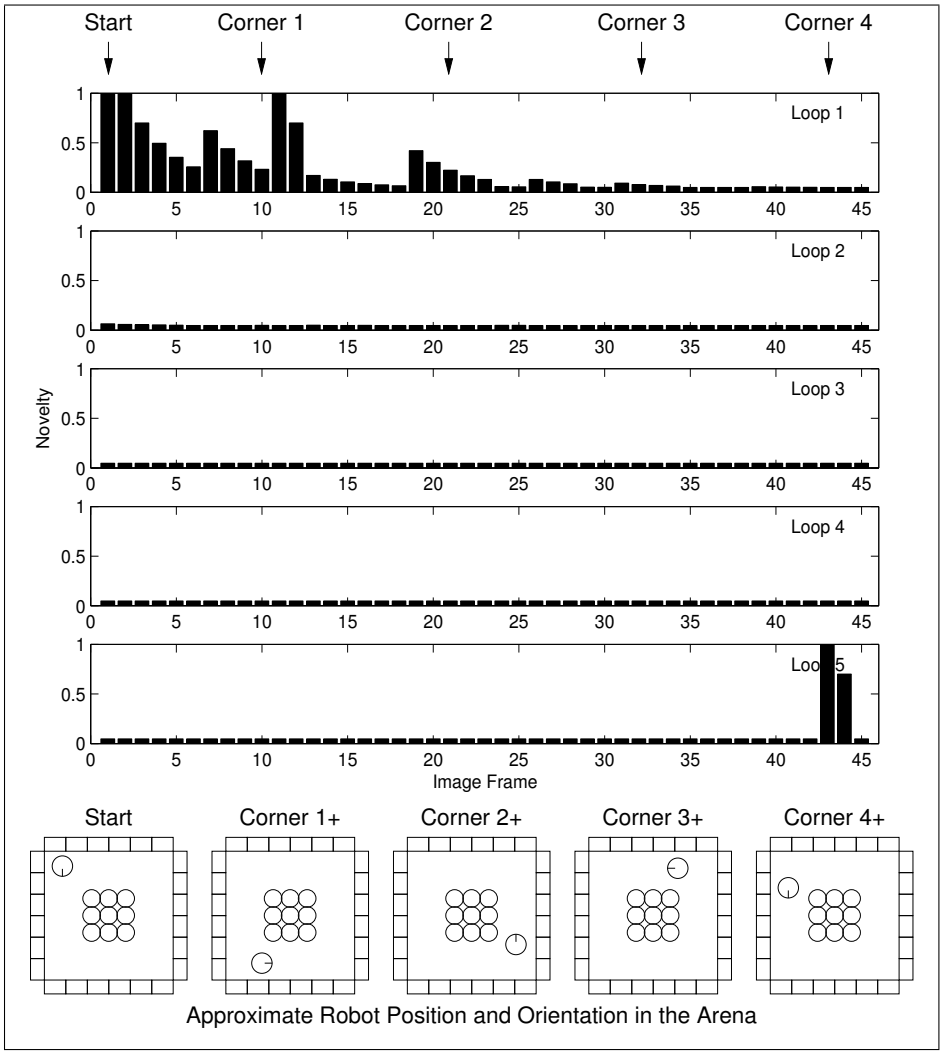



Figure 4.5: Exploration of the empty arena using global colour histograms (64 bins) as image encoding scheme. Learning and habituation are slower than when histograms with 32 bins are used, confirming an expected reduction in generalisation. Although the network was well habituated to the arena, by the end of the fifth loop a new node was acquired and did not have time to be completely habituated. The pictograms below the graphs indicate the approximate position and orientation of the robot in the arena while performing the exploration loops.

eralisation has been reduced. Also, comparing the first exploration loop in Figure 4.5 to the first exploration loop in Figure 4.3, one can notice that the network takes longer to habituate to the environment when using colour histograms with 64 bins.

It is interesting to notice that at the end of the fifth loop, when the robot was turning the last corner, the colour distribution of the environment was considered novel by the GWR network and resulted in the addition of a

new node, which could not be completely habituated. In fact, the efficacy of the habituable synapse of this newly added node was kept at 0.7 because it only had time to fire twice (in frames 43 and 44 of the last exploration loop). Nevertheless, the acquired model was used to inspect the arena with the orange football, yielding the results depicted in Figure 4.6.

Examining Figure 4.6, one can verify that some of the frames in which the orange ball appeared were correctly assigned as containing novel colour

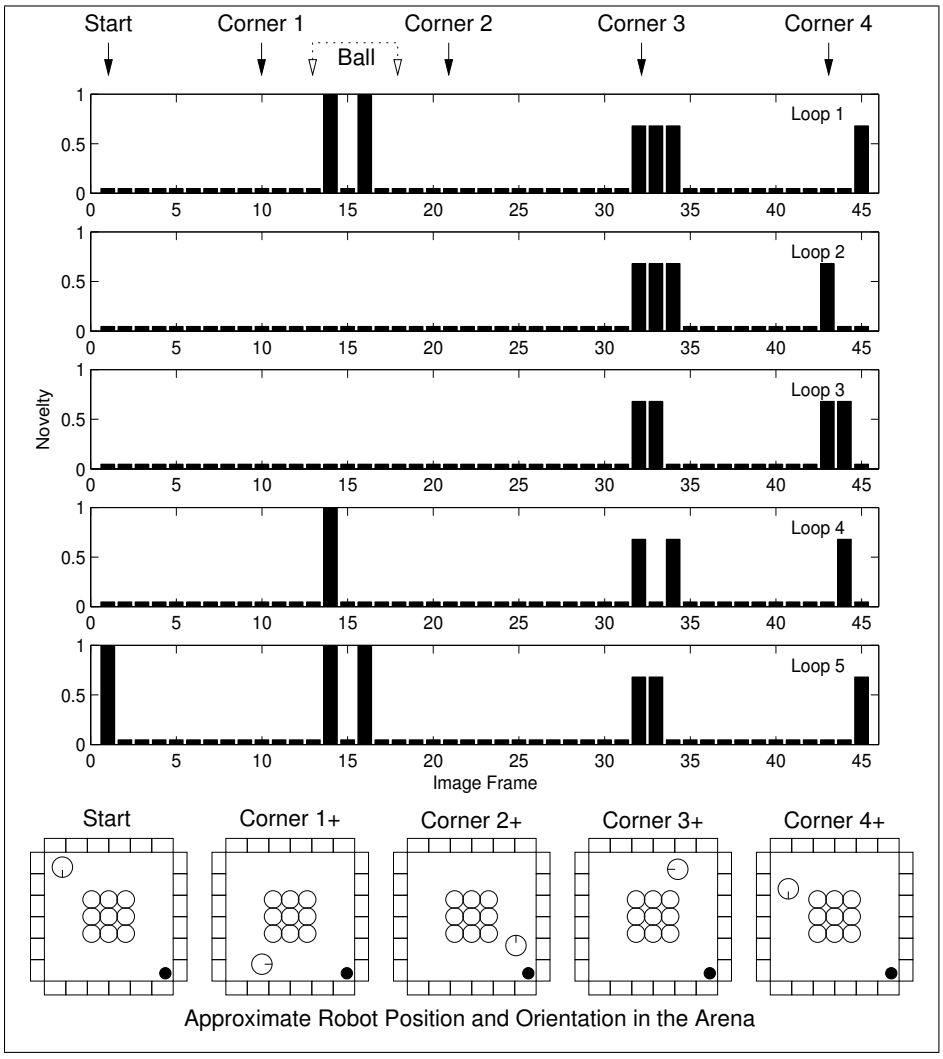



Figure 4.6: Inspection of the arena with the orange football (novel stimulus) using global colour histograms (64 bins) as image encoding scheme. The system was able to detected the novel stimulus in some of the loops around the arena. However, erroneous novelty indications occurred repeatedly when the robot was turning the last two corners of the arena, due to incomplete habituation of the last node acquired during exploration. Pictograms below the graphs indicate the approximate position and orientation of the robot in the arena while performing the inspection loops and also the location of the novel stimulus.

features (in the first, fourth and fifth loops). However, several frames towards the end of the loops were misclassified as if they contained novel features. Interestingly, all of these misclassifications involved firing of the incompletely habituated node discussed above (this can be easily identified by the novelty measure of 0.7 reported by the GWR network in Figure 4.6). Also, it can be noticed from the novelty graphs that the misclassification of these image frames seem to be somehow associated with the robot turning the last two corners of the arena.

The $\chi^2$ analysis of this experiment was inconclusive because the resulting contingency table was ill-conditioned (see Section 3.3). The quantitative analysis resulted in $V = 0.08$, $U = 0.01$ and $\kappa = 0.08$ (no agreement between system response and ground truth data).

**Repeating the experiment.** However, by repeating the second experiment and training the GWR network with data from the first four exploration loops only, we were able to eliminate most of the false novelties due to the weakly habituated node acquired during the end of the fifth exploration loop. The reason of the acquisition of this node and its firing during the inspection phase in the second experiment is going to be temporarily left aside, until the next set of experiments are presented.

The new results obtained, which avoided most of the false novelties that happened in the previous trial, are shown in Figure 4.7. In spite of this fact, the resulting contingency table was still ill-conditioned for the $\chi^2$ analysis. Nevertheless, the quantitative analysis yielded $V = 0.34$, $U = 0.09$ and $\kappa = 0.24$, revealing a weak agreement between system response and ground truth (better than random guessing, see Section 3.3).

The results from the experiments above are important for at least two reasons. First, they show the importance of having a well trained and habituated GWR network in order to minimise false novelties during its use as novelty filter; and second, the influence of the image encoding scheme

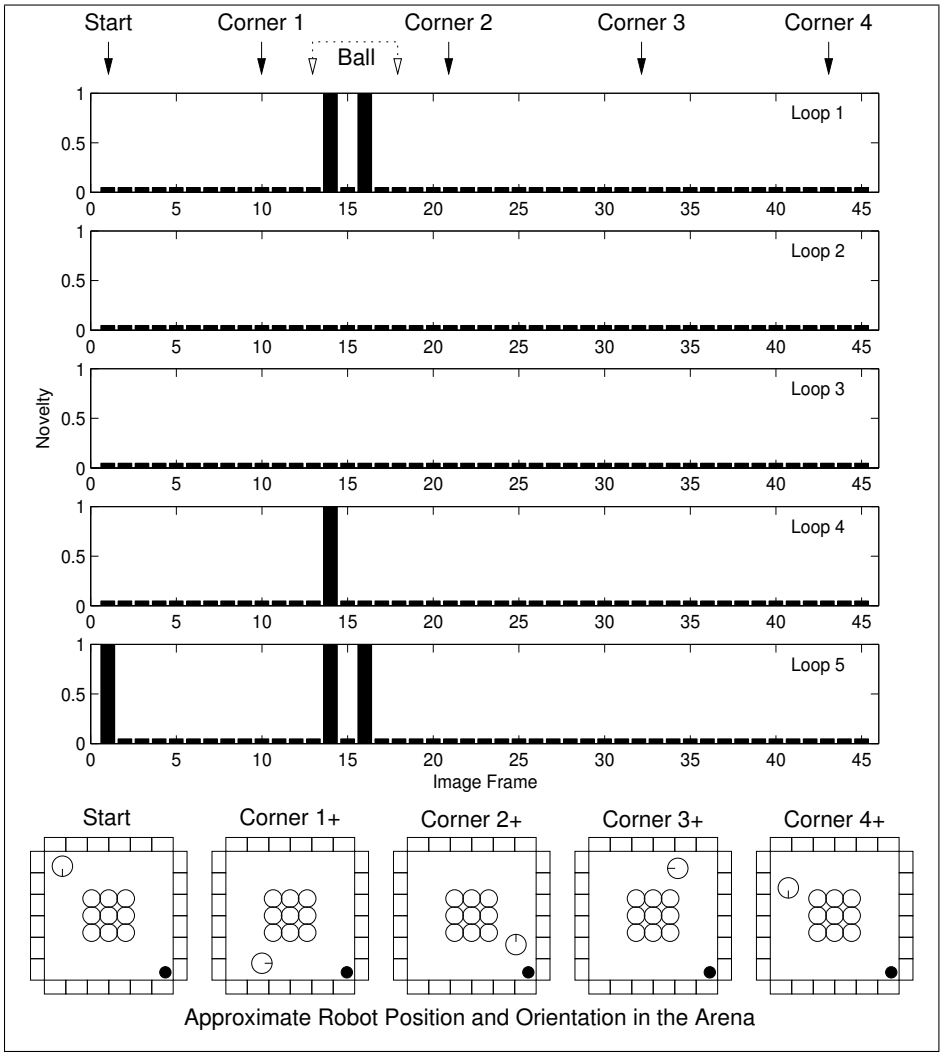



Figure 4.7: Re-inspection of the arena with the orange football (novel stimulus) using global colour histograms (64 bins) as image encoding scheme. The GWR network was retrained with data from the first four exploration loops only, in order to avoid acquiring a new node at the end of the fifth exploration loop. This resulted in the removal of false novelty detections when the robot was turning the last two corners of the arena.

in the overall system performance, particularly concerning the ability to generalise, was made evident.

Table 4.1 summarises the results obtained for the experiments using global histograms, all of which yielded ill-conditioned contingency tables for the $\chi^2$ test, making it impossible to assess the statistical significance in the association between actual novelty status and novelty filter response. $V$, $U$ and $\kappa$ were computed for comparison purposes.

Table 4.1: Performance comparison using global colour histograms (225 samples) while detecting the orange football as novel stimulus in the arena (Experiments 1 and 2). Resulting contingency tables were ill-conditioned for the $\chi^2$ test.

| | Experiment 1<br>32 Bins | Experiment 2a<br>64 Bins | Experiment 2b<br>64 Bins# |
|---|---|---|---|
| Orange ball | $V = 0.08^*$<br>$U = 0.02$<br>$\kappa$ = -0.07 | $V = 0.08^*$<br>$U = 0.01$<br>$\kappa = 0.08$ | $V = 0.34^*$<br>$U = 0.09$<br>$\kappa = 0.24$ |

*Ill-conditioned contingency tables for the $\chi^2$ test

#Exploration using only four loops — see text for details

## 4.2 Experiments 3 and 4: Novelty Detection from Local Colour Histograms

Although histograms of visual features can be powerful descriptors in many applications, their use in a global fashion weakens their ability to capture and represent small details present in the visual field. Strictly speaking, any statistical representation tends to dilute the contribution of features that appear less frequently in a sea of more common features. The fact is that an object or visual feature that occupies a small area relative to the size of the image will have a small contribution to a global histogram. Such a small feature would be probably disregarded by higher levels of processing as if it was noise.

As small details are often relevant for novelty detection tasks (for instance, a crack in a sewer pipe), global histograms — in reality, global representations in general — are not a good solution for the image encoding problem. The results shown in Table 4.1 demonstrate this. Furthermore, a novelty filter that highlights *which* image frames contain some novel visual features is not as useful as a filter that also locates *where* within these image frames the novel features are.

**Local histograms.** Having this in mind, we decided to conduct experiments using colour statistics at selected locations rather than global colour

statistics, which means computing weighted colour histograms in several distinct regions of the image. It was also decided to follow a recent trend in the Computer Vision community and use interest point or saliency detectors, instead of classical image segmentation methods. The main idea is to locate points that stand out within the input frame and then select their vicinities as candidate regions to be encoded and processed by the novelty filter.

Among many available saliency detectors (Ferreira and Borges, 2004; Itti et al., 1998; Kadir and Brady, 2003; Kadir et al., 2004; Lindeberg, 1998; Loupias et al., 2000; Lowe, 1999, 2004; Mikolajczyk and Schmid, 2001, 2002, 2004; Shi and Tomasi, 1994), the saliency map model developed by Itti, Koch, and Niebur (1998) was the one chosen for our experiments (a detailed description of the saliency map model and its advantages is given in Chapter 2).

The main reason for the choice of the saliency map is that this particular attention model identifies the most unusual features — the salient ones — within the image frame according to their intensity, colour and orientation in several scales. This effect happens due to the use of a normalisation operator $\mathcal{N}(.)$ that promotes less frequent features while suppressing the most frequent ones in any individual feature map (see Subsection 2.2.1). Features that are considered salient in this context are then more likely to represent novelty and therefore constitute good candidates to be encoded and classified by the novelty filter. In fact, recent research has shown that biological visual systems use local changes in image features to perform predictive coding at the retinal level, dynamically adapting to the visual statistics of the environment (Hosoya et al., 2005).

In the following experiments, we computed saliency maps for each input frame and used the points corresponding to the nine highest saliency values — the nine most salient points — to establish the centre of image

patches of 24 × 24 pixels in size (the image patch size was selected to be one fifth of the input frame height, 120 pixels). Each of the selected image patches had corresponding weighted colour histograms computed, which were individually fed to the GWR network (see Figure 3.2 on page 48).

### 4.2.1 Experiment 3: Local Colour Histograms with 32 Bins

**Normality model acquisition.** The first experiment with local statistics used once again colour histograms of 32 bins. The same set of images acquired in the arena built for Experiments 1 and 2 was used. Results obtained for the exploration of the empty arena (learning enabled) is shown in Figure 4.8. Because nine feature vectors were generated and classified for each input frame, the novelty graphs for qualitative assessment of results had to be adapted. Now, for experiments using the local approach, the novelty graphs depict the average measure of novelty given by the GWR network in each frame.

The clear exponential decay due to progressive habituation is not evident like in the experiments using global encoding because the measure of novelty shown is now the average of the nine salient regions in each frame. Nevertheless, it is still possible to notice that the novelty activity declines as the robot repeatedly explores the arena and that the network nodes are completely habituated by the end of the last loop. It is interesting to note that, once more, increases in novelty measure happen when the robot is turning corners. At the end of the exploration phase, the GWR network had acquired 8 nodes, four times as much as when using global colour histograms with the same 32 bins.

**Visual feedback.** Because a local image encoding procedure is now in use in conjunction with an attention mechanism, it is also possible to generate

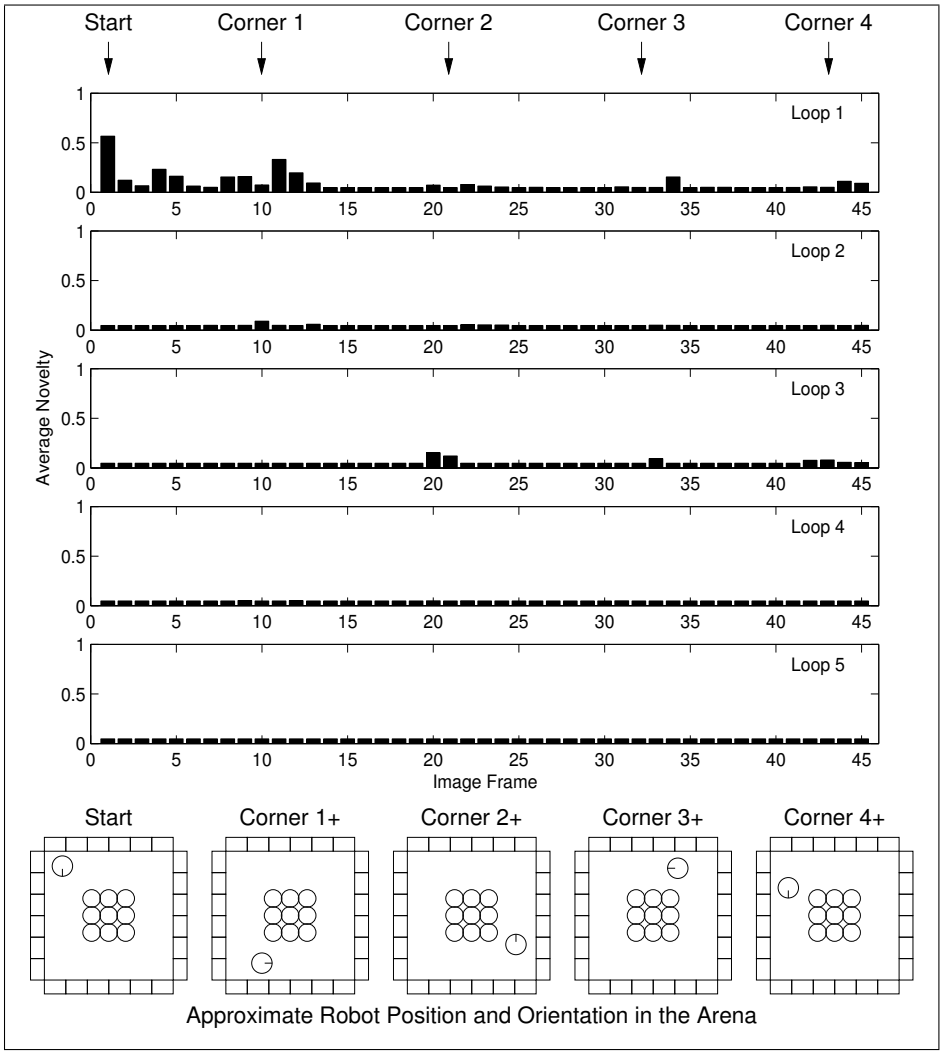



Figure 4.8: Exploration of the empty arena using local colour histograms (32 bins) as image encoding scheme. Nine salient regions were selected per image frame in order to assess local novelty and the graphs show the average novelty measure for each input image frame. Complete habituation is achieved after the third loop around the arena.

output images indicating which features of the environment were considered salient and, among those, which were considered novel. Figure 4.9 depicts the output image at the robot's start position in the first exploration loop (the input image corresponds to the one shown in Figure 4.2).

In Figure 4.9 salient points are marked with numbers (0 corresponding to the most salient location) and their corresponding regions, if classified as novel, are marked with white circles. We find our earlier hypothesis that global encoding schemes will suppress small features confirmed: regions 0 and 1 correspond to small marks on the arena's floor, constituting details

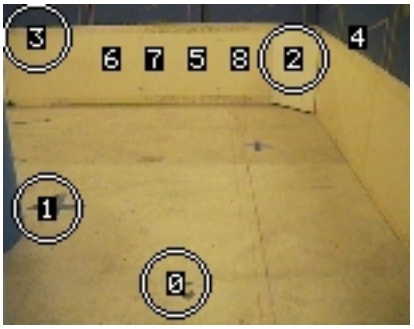



Figure 4.9: Output for the image at the robot's start position (first exploration loop). The numbers indicate the location of salient points in order of importance (0 corresponds to the most salient) and the white circles indicate that the region corresponding to a particular salient point was considered novel. Because this was the very first image to be presented to the robot, there are several regions that were marked as novel. As the robot explored the arena and the GWR network habituated on it, fewer and fewer regions were labelled as novel.

of the environment that would be ignored by the global image encoding approach. Of course, the detection of small details may be seen as an advantage or disadvantage depending on the application.

**Novelty detection.** The trained network (learning disabled) was used to inspect the arena containing the orange football and highlight any novelties, in the same fashion as in previous experiments. Performance of the system during the inspection phase of the arena can be evaluated by means of Figure 4.10.

**Qualitative assessment.** The impact of using a local approach for image encoding was clearly positive from a qualitative perspective. The ball was correctly detected as the novel feature in the environment (see also Figure 4.11) and there were almost no false novelties detected, except some cases that occurred close to the corner immediately before the orange ball appeared — coincidentally, the same area of the arena that generated false positives in the experiments using global image encoding (see Figure 4.4 on page 69).

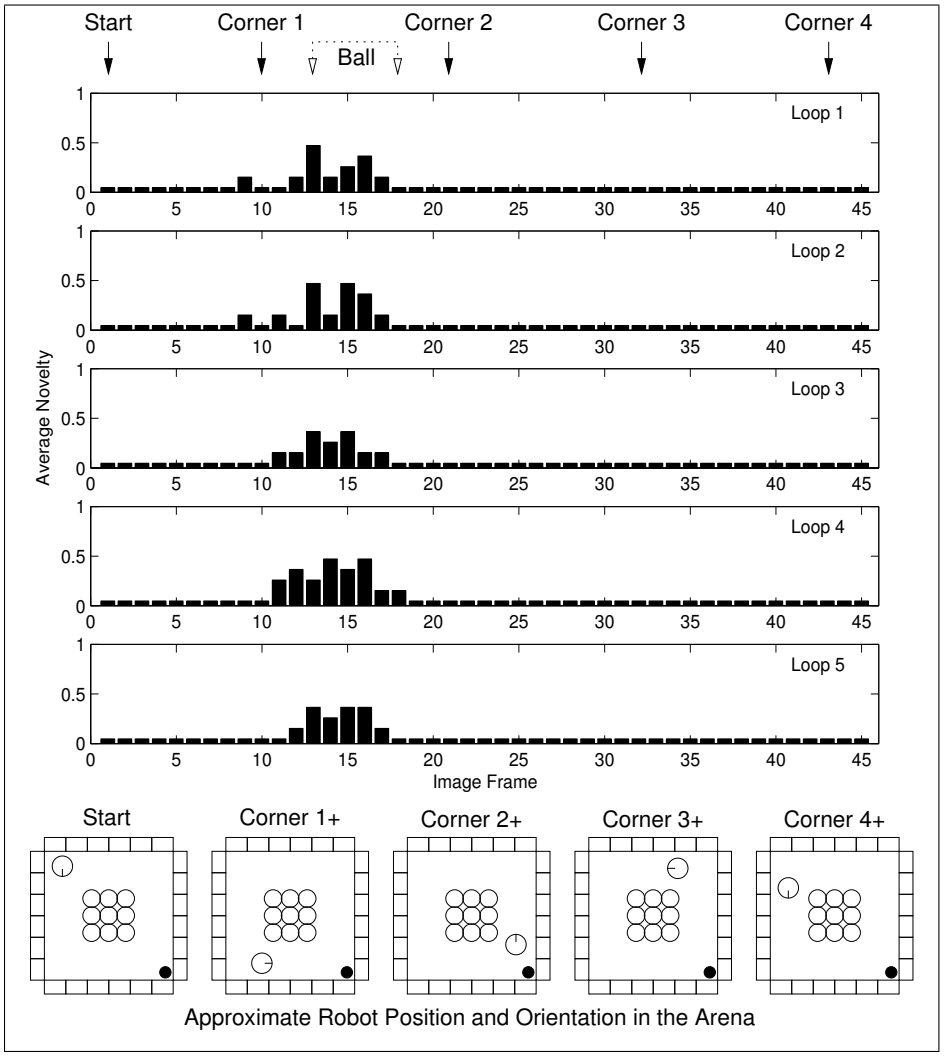



Figure 4.10: Inspection of the arena with the orange football (novel stimulus) using local colour histograms (32 bins) as image encoding scheme. The ball is clearly and consistently detected in every inspection loop around the arena.

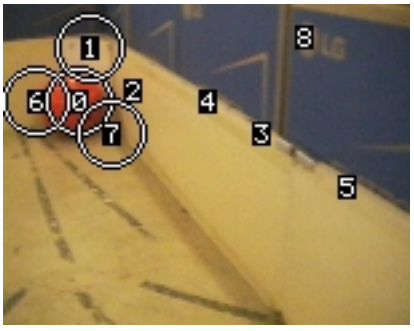



Figure 4.11: Output for an image where the orange football (novel stimulus) appears. The ball is clearly highlighted with white circles as being novel.

**Quantitative assessment.** It can be seen in Figure 4.11 that the system was able to indicate regions containing part of the orange ball, which was the

novel object present in the arena. Contingency table analysis through the $\chi^2$ test revealed statistical significance between the novelty filter response and actual novelty status ($p \leq 0.01$). The strength of this association was also measured by computing $V = 0.75$, $U = 0.53$ and $\kappa = 0.74$, revealing strong agreement between system response and actual novelty status.

### 4.2.2 Experiment 4: Local Colour Histograms with 64 Bins

A second experiment was conducted using local colour histograms with 64 bins, similar to what was done before in the experiments using global colour histograms. The novelty graphs obtained for the exploration phase of the empty arena are given in Figure 4.12.

Analysis of the novelty measure over time during the five exploration loops indicates that the GWR network was completely habituated to the arena. Using local colour histograms with 64 bins as input vectors resulted in the acquisition of nine nodes, one node more than the number acquired when using 32 bins.

**Qualitative assessment.** Qualitatively speaking, the results obtained for the inspection phase of the arena using 64 bins (shown in Figure 4.13) are very similar to the ones obtained using 32 bins for the local colour histograms (Figure 4.13).

Unlike the case of global colour histograms, a significant change in the resolution of local colour histograms showed not to have such a big impact in overall performance. We attribute this to the fact that the local approach uses histogramming to characterise colour distributions with much less samples ($24 \times 24 = 576$ pixels) than the global approach ($152 \times 120 = 18240$ pixels) and therefore less resolution is needed. The mechanism of attention clearly offers a contribution to the efficient representation of visual data

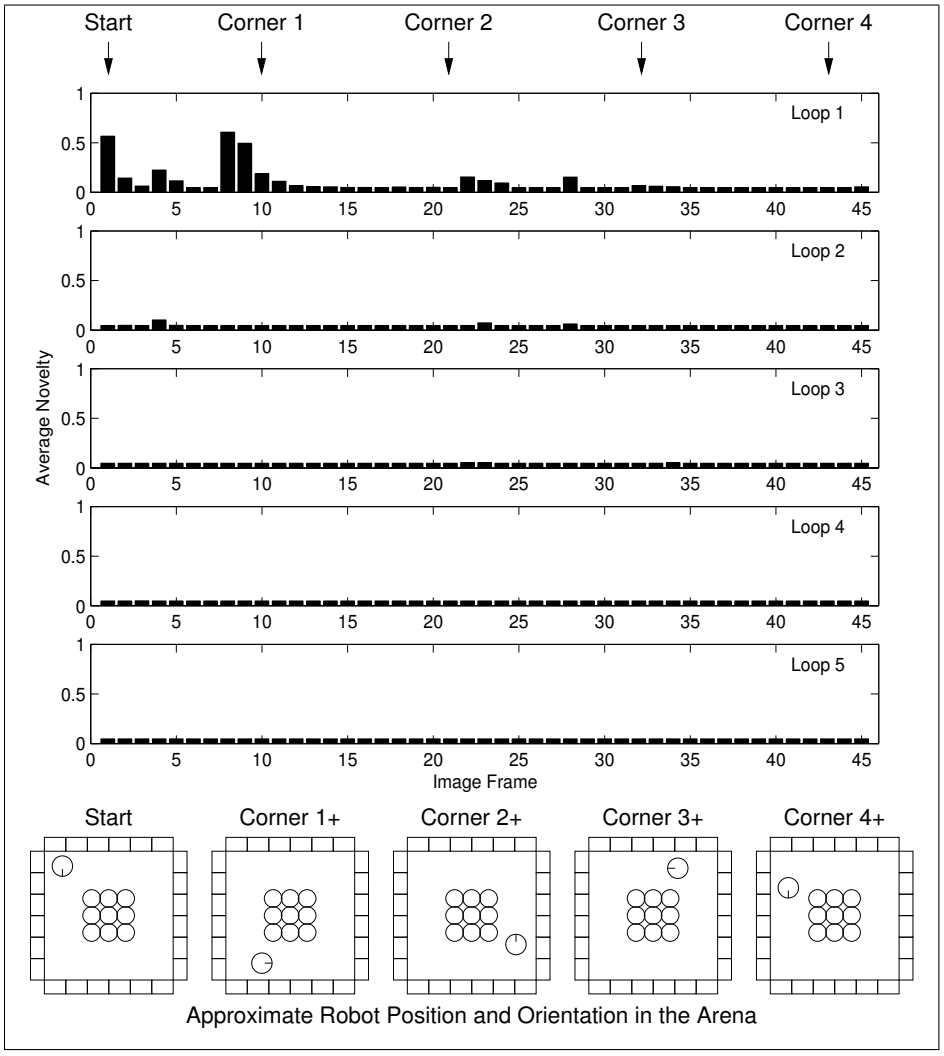



Figure 4.12: Exploration of the empty arena using local colour histograms (64 bins) as image encoding scheme. Nine salient regions were selected per image frame in order to assess local novelty and the graphs show the average novelty measure for each input image frame. Complete habituation was achieved after the second loop around the arena.

by splitting a relatively large image frame into several small image patches with high information contents (salient regions). Furthermore, detection of interest points using the saliency map model offers other important characteristics, such as robustness to translations, which will be discussed further later on.

**Quantitative assessment.** Because the results obtained using the attention model and local colour histograms were qualitatively very good, a quantitative assessment was also performed using contingency table analy-

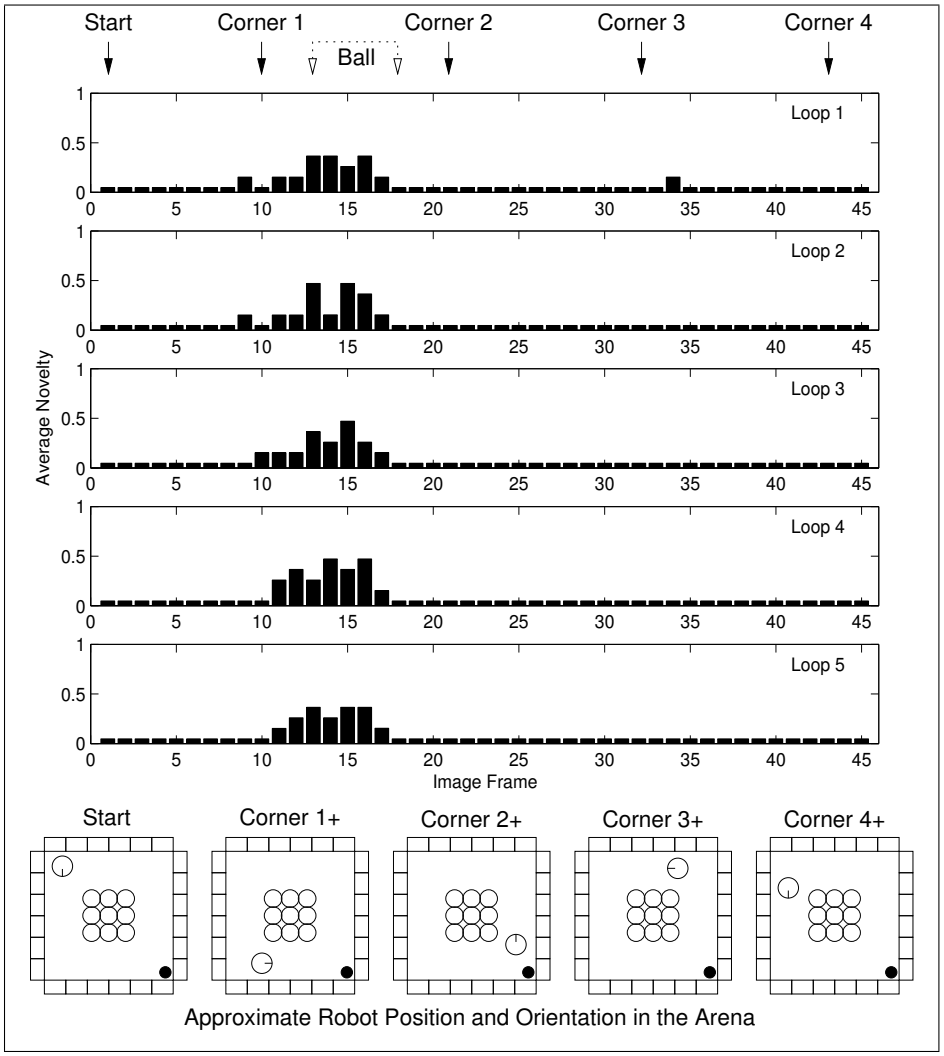



Figure 4.13: Inspection of the arena with the orange football (novel stimulus) using local colour histograms (64 bins) as image encoding scheme. The ball is clearly and consistently detected in every inspection loop around the arena.

sis. A $\chi^2$ analysis of the novelty filter responses (5 loops $\times$ 45 images $\times$ 9 patches = 2025 samples) revealed statistical significance ($p \leq 0.01$) for the association between system response and ground truth in both experiments. The quantitative performance comparison using Cramer's $V$, uncertainty coefficient $U$ and the $\kappa$ index of agreement is given in Table 4.2.

By comparing the values for $V$, $U$ and $\kappa$ in Table 4.2, one can notice that the performance of the system using local colour histograms with 32 bins was slightly better.

Table 4.2: Performance comparison using local colour histograms (2025 samples) while detecting the orange football as novel stimulus in the arena (Experiments 3 and 4). Both experiments resulted in strong agreement between novelty filter response and actual novelty status.

| | Experiment 3<br>32 Bins | Experiment 4<br>64 Bins |
|---|---|---|
| Orange ball | $V = 0.75$<br>$U = 0.53$<br>$\kappa = 0.74$ | $V = 0.68$<br>$U = 0.46$<br>$\kappa = 0.67$ |

It is interesting to point out that the use of colour histograms, which resulted in almost a complete failure when used in a global fashion, yielded very good results when applied locally in regions selected by the saliency map.

**False novelties.** Persistence in the detection of false novelties, particularly when the robot was close to the arena's corners, demanded an analysis of the corresponding images, which revealed the reason of such a problem. We analysed the pixel values of equivalent regions of images acquired in each corner and discovered that there were fluctuations, probably due to changes in illumination, caused by changes in viewpoint as the robot moved, or the camera's automatic gain control, which could not be disabled by the robot's control software.

The noise was observed in intensity and saturation, as hue alone is invariant to changes in sampled illumination. Eliminating saturation from the weighted colour histograms encoding scheme was not a solution because then the ability to discriminate between saturated and unsaturated colours — red and pink, for instance — would be lost. In fact, as one would expect, it is also necessary to include intensity information in the image encoding process so that shades of grey, which do not have defined hue or saturation, can also be detected and discriminated. Therefore, we needed an alternative colour-based image encoding procedure, invariant to changes in illumination, to minimise the problem of false novelties.

## 4.3 Experiments 5 to 8: Novelty Detection from Colour Angles

Sensitivity to illumination is identified by many researchers as a serious problem to colour matching and led them to develop colour constant image representation schemes (Finlayson et al., 1996, 1998; Funt and Finlayson, 1995; Gevers and Smeulders, 1997; Healey and Slater, 1994; Healey and Wang, 1995). Among those, the colour angular indexing proposed by Finlayson, Chatterjee, and Funt (1996) is particularly interesting because it provides a very compact colour constant representation based on statistics (colour channel covariances), which encodes the characteristics of the colour distribution as angles between colour channel vectors in image space.

The colour vectors **r**, **g** and **b** respectively contain all pixel values in the $R$, $G$ and $B$ colour channels of the image in scanning order. Hence, each colour vector has the dimensionality of the image area in pixels ($width \times height$). In order to compute colour angles, it is first necessary to obtain zero-mean colour vectors ($\mathbf{r}_0$, $\mathbf{g}_0$ and $\mathbf{b}_0$) by subtracting the corresponding average pixel value from each original colour vector:

$$\mathbf{r}_0 = \mathbf{r} - \bar{r}, \tag{4.7}$$

$$\mathbf{g}_0 = \mathbf{g} - \bar{g}, \tag{4.8}$$

$$\mathbf{b}_0 = \mathbf{b} - \bar{b}, \tag{4.9}$$

where $\bar{r}$, $\bar{g}$ and $\bar{b}$ are the average pixel values of the original colour vectors, **r**, **g** and **b**, respectively.

The next step is to normalise the zero-mean colour vectors to unitary length by dividing each one by their respective norm:

$$\mathbf{r}_N = \frac{\mathbf{r}_0}{\|\mathbf{r}_0\|}, \tag{4.10}$$

$$\mathbf{g}_N = \frac{\mathbf{g}_0}{\|\mathbf{g}_0\|}, \tag{4.11}$$

$$\mathbf{b}_N = \frac{\mathbf{b}_0}{\|\mathbf{b}_0\|}, \tag{4.12}$$

where $r_N$, $g_N$ and $b_N$ are the normalised zero-mean colour vectors.

Colour channel covariances are equivalent to dot products, which in this case geometrically correspond to the cosine of the angles between the corresponding unitary length colour vectors. These angles are invariant to changes in illumination and can be computed by the inverse cosine of colour vector dot products:

$$\phi_{rg} = \arccos(\langle \mathbf{r}_N, \mathbf{g}_N \rangle), \tag{4.13}$$

$$\phi_{gb} = \arccos(\langle \mathbf{g}_N, \mathbf{b}_N \rangle), \tag{4.14}$$

$$\phi_{rb} = \arccos(\langle \mathbf{r}_N, \mathbf{b}_N \rangle), \tag{4.15}$$

where $\langle \mathbf{a}, \mathbf{b} \rangle$ denotes the dot product of vectors $\mathbf{a}$ and $\mathbf{b}$.

The interesting aspect of this colour representation is that changes in the colour of the illuminant (the illuminating device) or changes in sampled illumination due to robot motion result in a rotation of the colour channel vectors for an image region. Nevertheless, the angles between those vectors remain the same and this is what grants robustness to illumination changes.

### 4.3.1 Experiment 5: Local Colour Angles

For the new set of experiments about to be presented we have used local colour angles to encode the colour distribution of the regions selected by the attention model. Therefore, the feature vectors fed to the GWR network in this case had only three dimensions ($\phi_{rg}$, $\phi_{gb}$ and $\phi_{rb}$). Figure 4.14 shows the novelty graphs obtained for the exploration of the empty arena.

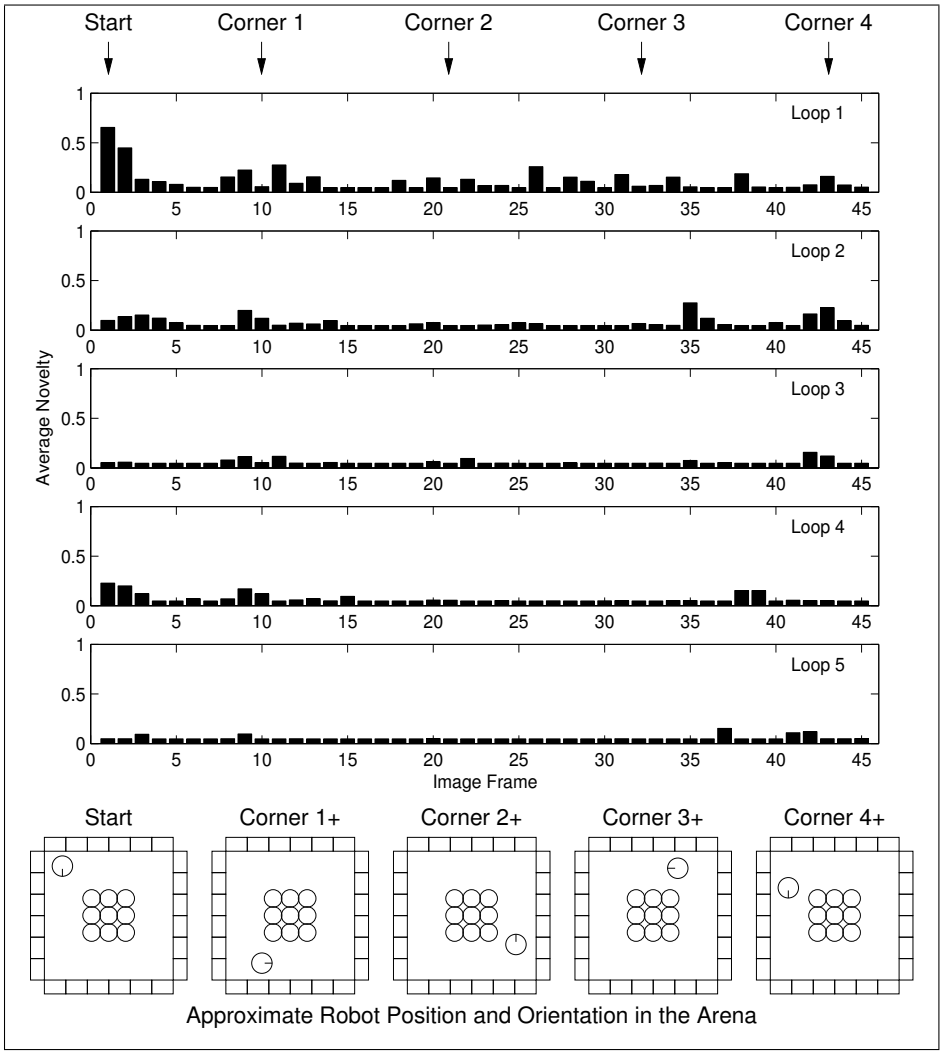



Figure 4.14: Exploration of the empty arena using local colour angles as image encoding scheme. The GWR network is sufficiently habituated to the environment by the end of the fifth exploration loop.

After exploration, 25 nodes were acquired by the GWR network, indicating the excellent ability of local colour angular encoding to discriminate colour distributions — network nodes increased three times when compared to the case using colour histograms. As can be seen in Figure 4.14, by the end of the fifth loop the network was still responding to the environment with novelty measures above the level of complete habituation. Inspection of node synaptic efficacies revealed that seven nodes had values above 0.1. Nevertheless, the level of habituation was considered sufficient and once more the acquired model was used to inspect the arena containing the orange ball. Results of the inspection phase are given in Figure 4.15.

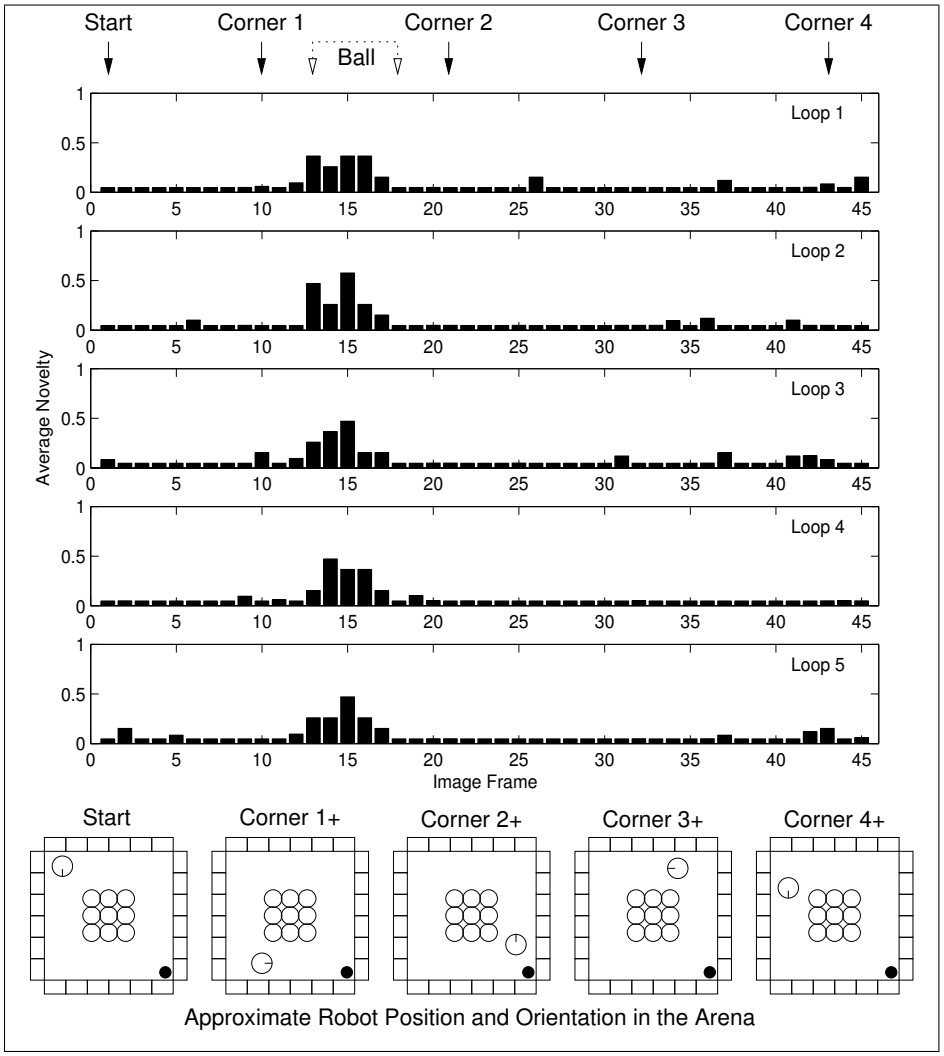



Figure 4.15: Inspection of the arena with the orange football (novel stimulus) using local colour angles as image encoding scheme. The ball is clearly and consistently detected in every inspection loop around the arena. Also, false novelties were minimised.

**Qualitative assessment.** Figure 4.15 shows some qualitative improvement in the system's performance in spite of the fact that not all of the GWR network nodes were completely habituated. Very few false novelties were detected, while the image frames where the ball appears were clearly identified as having high levels of novelty. One can also notice that the false novelties that were present in the previous experiments immediately after turning of the first corner (see Figure 4.10 on page 80 and Figure 4.13 on page 83) are now suppressed.

**Quantitative assessment.** The contingency table analysis of Experiment 5 revealed statistically significant association between novelty filter response and actual novelty status ($\chi^2$ test, $p \leq 0.01$) and yielded the following quantitative results: $V = 0.84$, $U = 0.73$ and $\kappa = 0.82$ (almost complete agreement between ground truth and system response).

### 4.3.2 Experiment 6: Global Colour Angles

In spite of resulting in input vectors with just three dimensions, encoding based on colour angles has shown an excellent ability to discriminate colour distributions. This property was confirmed by running experiments in which the attention mechanism was removed, *i.e.* by using colour angles in a *global* fashion, as in Experiments 1 and 2 presented earlier in this chapter.

Figure 4.16 shows the novelty graphs for the five exploration loops around the empty arena when using global colour angles as image encoding scheme. Like in previous experiments using global colour histograms, the *whole* input frame (152 $\times$ 120 pixels) was encoded and used to train the GWR network.

It can be noticed that learning was extremely fast — only one loop around the arena resulted in complete habituation of the GWR network. As a result of training the network with global colour angles, only two nodes were acquired, the same number as in the experiment using global colour histograms with 32 bins. However, unlike the experiments that used global colour histograms, the results obtained during the inspection phase were excellent, as Figure 4.17 demonstrates.

Figure 4.17 shows that only input frames where the orange football was visible to the robot's camera were correctly labelled as novel (high novelty value), except for frame 18 in every loop. However, visual inspection of image frame 18 in each loop reveals that in reality very little of the ball is visible, justifying the system response (see Figure 4.18).

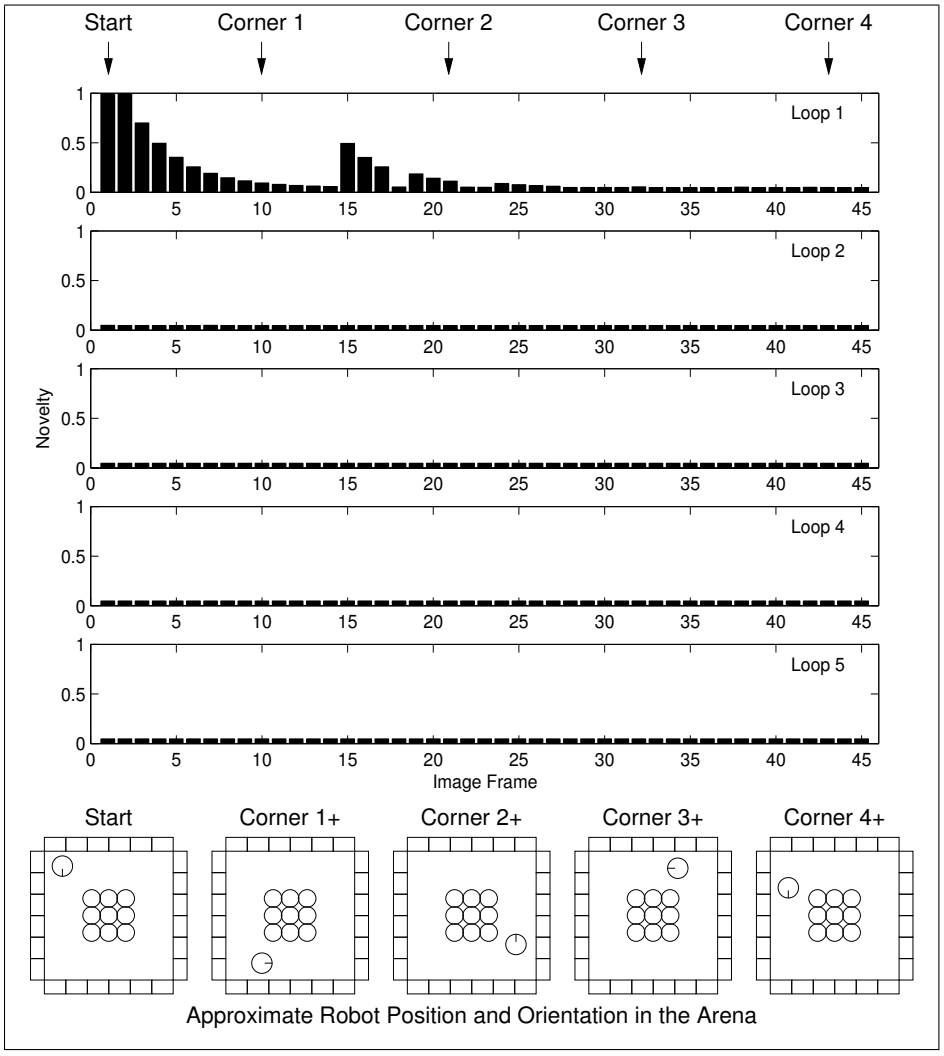



Figure 4.16: Exploration of the empty arena using global colour angles as image encoding scheme. The GWR network achieves complete habituation by the end of the first exploration loop around the arena.

In Figure 4.18 one can notice that the ball is largely out of the camera's field of view — the area occupied by the ball corresponds to 0.55% of frame 18 in the first loop. In frames that were correctly identified as novel (frames 13 to 17), the area occupied by the orange ball varies from 2.4% to 4.8%.

Table 4.3 shows a quantitative comparison between the results obtained in the experiments using local and global colour angular encoding. In both cases, the $\chi^2$ analysis revealed statistical significance ($p \leq 0.01$) between novelty detected and novelty actually present.

The local colour angle approach gave better results than its histogram-based counterparts (Experiments 3 and 4 — see Table 4.2 on page 84) and

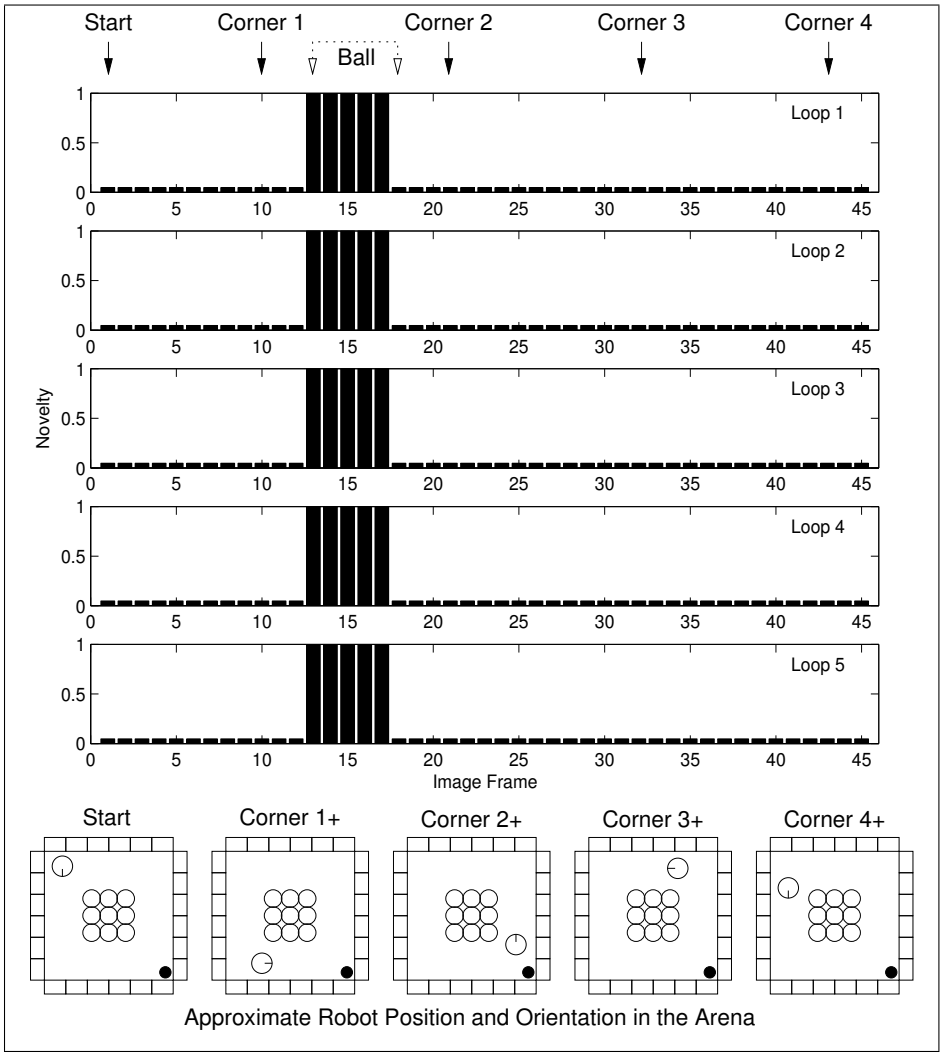



Figure 4.17: Inspection of the arena with the orange football (novel stimulus) using global colour angles as image encoding scheme. The ball is clearly and consistently detected in every inspection loop around the arena. Also, there are no false novelties.

(a) (b)

Figure 4.18: Image frame 18 (first loop) and its ground truth novelty map. The area occupied by the orange football corresponds to only 0.55% of the image frame.

Table 4.3: Performance comparison using local and global colour angles while detecting the orange football as novel stimulus in the arena (Experiments 5 and 6). Both experiments resulted in almost complete agreement between novelty filter response and actual novelty status.

| | Experiment 5<br>Local (2025 samples) | Experiment 6<br>Global (225 samples) |
|---|---|---|
| Orange ball | $V = 0.84$<br>$U = 0.73$<br>$\kappa = 0.82$ | $V = 0.89$<br>$U = 0.72$<br>$\kappa = 0.89$ |

the global colour angle method (Experiment 6) provided the best results so far. In fact, if frame 18 is deemed to have insufficient information to be labelled as novel, the global colour angular scheme yields perfect results ($V = 1$, $U = 1$ and $\kappa = 1$). However, as previously discussed, only the local approach is able to localise the novel colour features within the image frame.

Inspection of Table 4.3 also reveals an apparent contradiction between the statistics $V$, $U$ and $\kappa$ ($V$ and $\kappa$ for the global approach are greater than for the local approach, while $U$ is slightly less). The reason for these results is probably due to the large difference in the number of samples used in each approach (2025 samples for the local approach versus 225 samples for the global approach). We also consider the uncertainty coefficient $U$ to be more reliable than $V$ and $\kappa$ because of its foundation on information theory (entropy).

### 4.3.3 Experiment 7: Local Colour Angles Revisited

Having obtained successful results in the first arena with the orange ball, we decided to confirm them by conducting more experiments in a new setup. Therefore, we built another engineered environment, this time in the new robotics research laboratory at the University of Essex. The new arena was built with cardboard boxes only (no yellow wooden panels were used this time), resulting in its walls being mostly dark blue. Also, the floor of the

new arena was of shiny grey colour, as opposed to the matt yellow floor of the previous one.

As novel objects to be introduced in the new arena for the inspection phase of the experiments, besides the very conspicuous orange football, we have also used a much less conspicuous grey box (the colour of the box was very similar to the colour of the floor).

The idea behind the new experimental setup was to test the colour angular encoding in an environment which had predominantly grey and dark blue colours. We also wanted to check the system's ability to detect a novel object similar in colour to the environmental background and therefore not very salient.

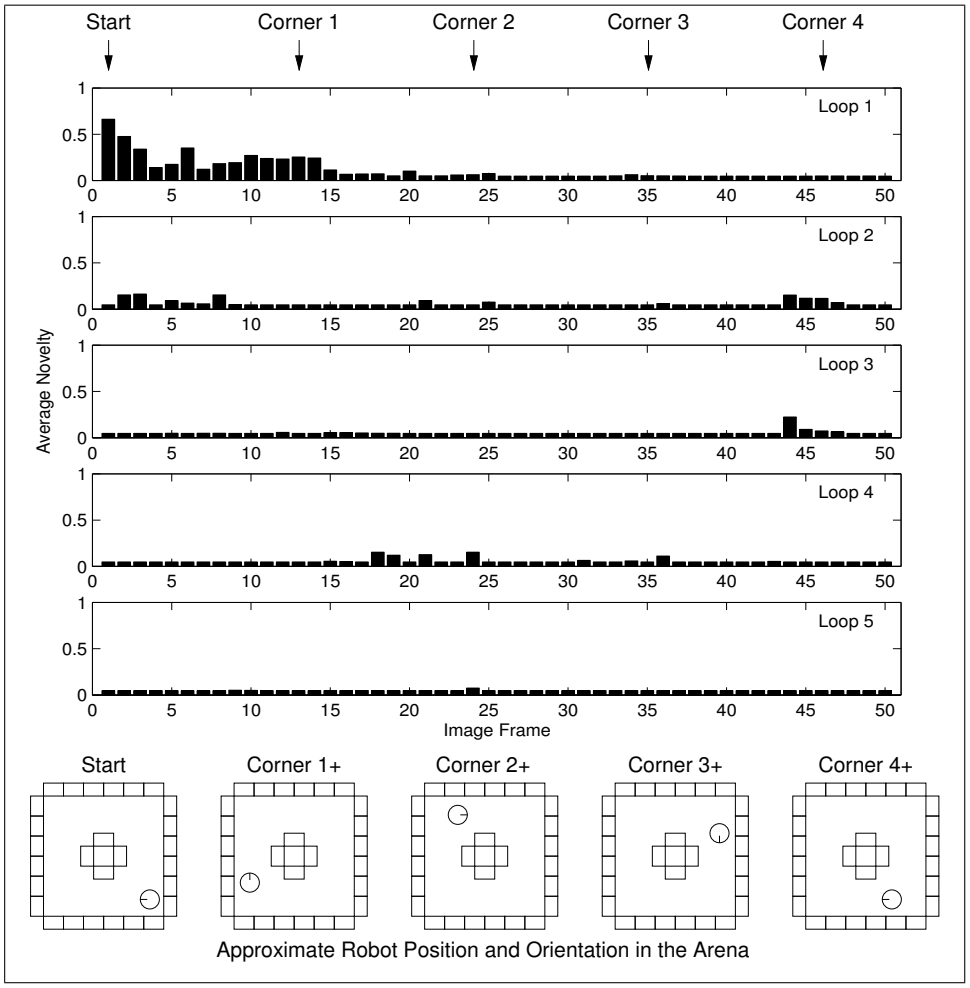



Figure 4.19: Exploration of the new empty arena using local colour angles as image encoding scheme. The graphs show that the GWR network is completely habituated to the new arena by the end of the fifth exploration loop.

**Normality model acquisition.** Training of the GWR network during exploration of the new empty arena was executed in the same fashion as in previous experiments and using the same parameters. However, exploration and inspection loops have now a total of 50 frames. Figure 4.19 shows the novelty graphs obtained for the exploration of the new arena, with the locations where the robot was turning corners indicated.

After the five loops around the new arena, the GWR network had completely habituated and acquired 18 nodes. Figure 4.20 shows the output image acquired at the robot's start position in the first exploration loop of the new arena.

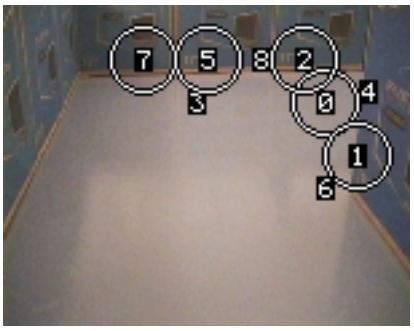



Figure 4.20: Output for the image at the robot's start position (first exploration loop of the new arena). As in previous examples, the numbers indicate the location of salient points in order of importance (0 corresponds to the most salient) and the white circles indicate that the region corresponding to a particular salient point was considered novel. Because this was the very first image to be presented to the robot, there are several regions that were marked as novel. As the robot explored the new arena and the GWR network habituated on it, fewer and fewer regions were labelled as novel.

**Novelty detection.** As usual, we used the trained GWR network as novelty filter during the inspection of the environment with a new object. The results of the inspection of the new arena with the orange ball are given in Figure 4.21.

The orange ball appeared in the camera's field of view after the robot turned the second corner of the new arena. All frames in which the football appeared are indicated in Figure 4.21, which also shows the successful detection of the novel object (see also Figure 4.22).

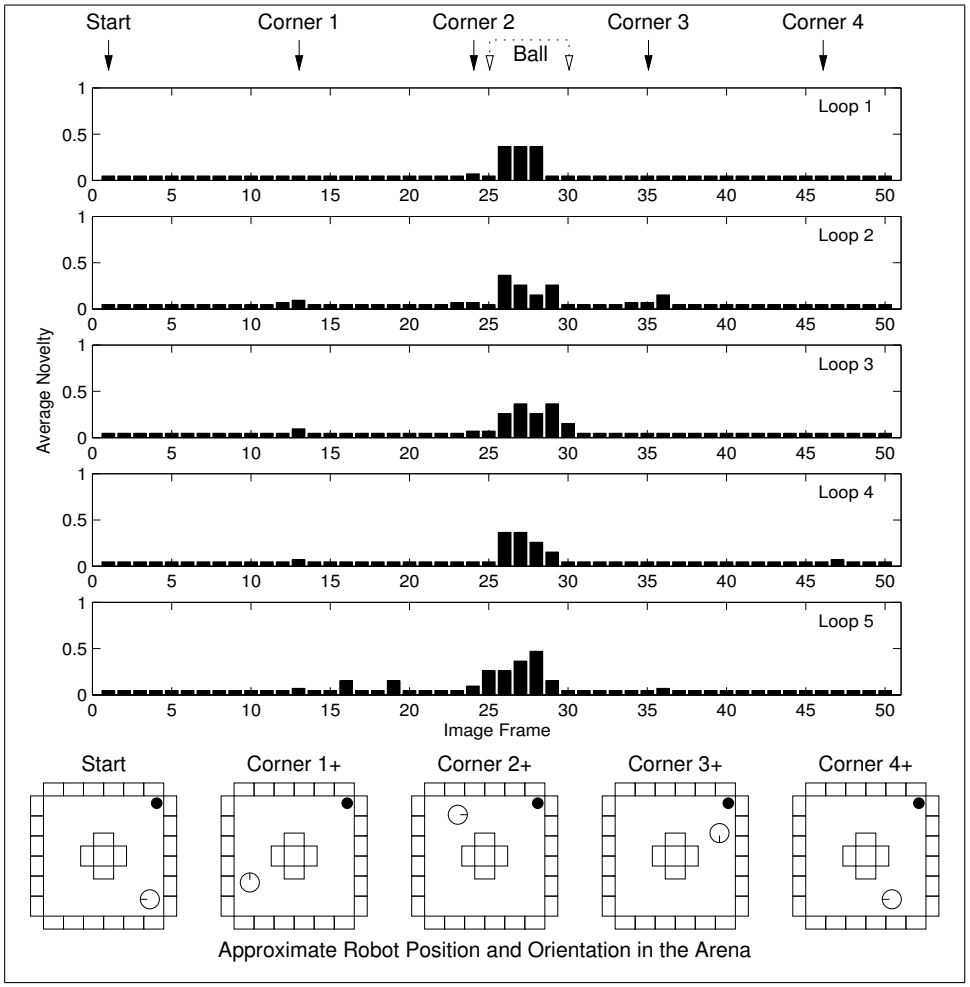



Figure 4.21: Inspection of the new arena with the orange football (novel stimulus) using local colour angles as image encoding scheme. The ball is clearly and consistently detected in every inspection loop around the arena.


Figure 4.22: Output for an image where the orange football appears in the new arena. The ball is correctly highlighted with white circles as the novel stimulus.

It can be seen in Figure 4.22 that the orange ball is correctly located and highlighted as the new object in the arena (numbers indicate salient locations found within the image frame, which appear circled when classified as novel by the GWR network).

A new inspection phase was performed in the new arena with the grey box and without the orange ball. The frames in which the grey box appeared in the camera's field of view (after the robot turned the first corner of the arena) are indicated with dotted arrows in Figure 4.23, where the results of this new inspection round are also given.

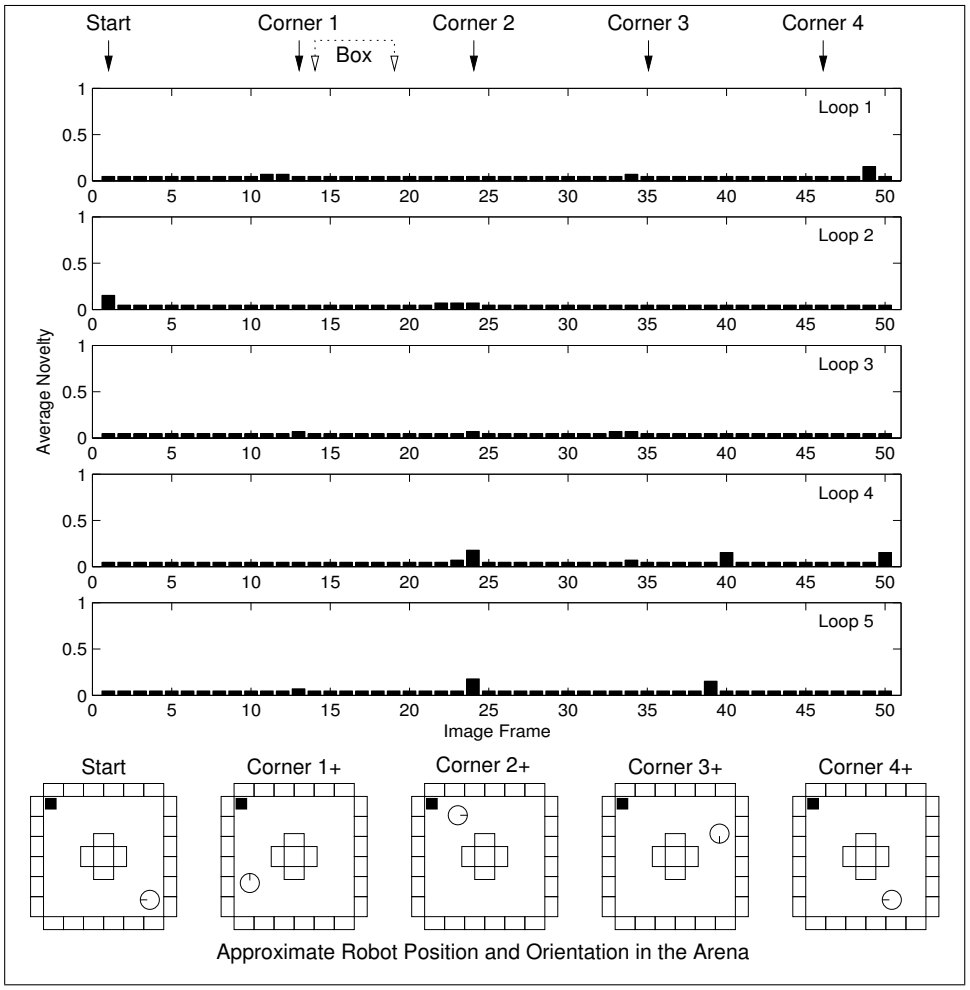



Figure 4.23: Inspection of the new arena with a grey box (novel stimulus) using local colour angles as image encoding scheme. The system fails to detect the novel stimulus.

One can notice in Figure 4.23 that the local colour angular encoding scheme has failed completely to highlight the grey box as novel object in the arena (see also Figure 4.24). This happens because regions containing only shades of grey have identical (or very similar) $R$, $G$ and $B$ colour components, meaning that all angles between the colour channel vectors are zero — shades of grey are literally invisible to colour angular encoding.

Figure 4.24 shows an image frame in which the grey box appears in the field of view but, despite having regions selected by the attention mechanism, is not highlighted as novel by the GWR network.


Figure 4.24: Output for an image where the grey box (novel stimulus) appears in the new arena (colour angular encoding). In spite of parts of the box being selected by the saliency map (regions 0 and 3), the system was unable to highlight them as novel stimuli.

The reason for not being able to detect the grey box is that colour angles cannot discriminate shades of grey. This situation happens because shades of grey have colour components $R = G = B$ and therefore colour angles always result in $\phi_{rg} = \phi_{gb} = \phi_{rb} = 0$. To solve this problem, it is necessary to include intensity information in our feature vectors. Experiments with colour angles with added information about intensity are reported in Section 4.4.

**Quantitative assessment.** Our contingency table analysis revealed statistical significance in the association between novelty filter response and ground truth data ($\chi^2$ test, $p \leq 0.01$) when the inspection involved the orange football. The strength of this association is measured by $V = 0.50$, $U = 0.37$ and $\kappa = 0.50$, meaning clear agreement between ground truth and system response. However, the $\chi^2$ test could not be performed when the inspection involved the grey box because the resulting contingency table was ill-conditioned (see Section 3.3). The quantitative analysis in this case resulted in $V = 0.07$, $U = 0.01$ and $\kappa = -0.04$, meaning no agreement between ground truth and novelty filter response.

### 4.3.4 Experiment 8: Global Colour Angles Revisited

For the sake of completeness and performance comparison, the experiments with the new arena were repeated using *global* colour angular encoding. Results for the exploration phase (empty arena) are given in Figure 4.25.

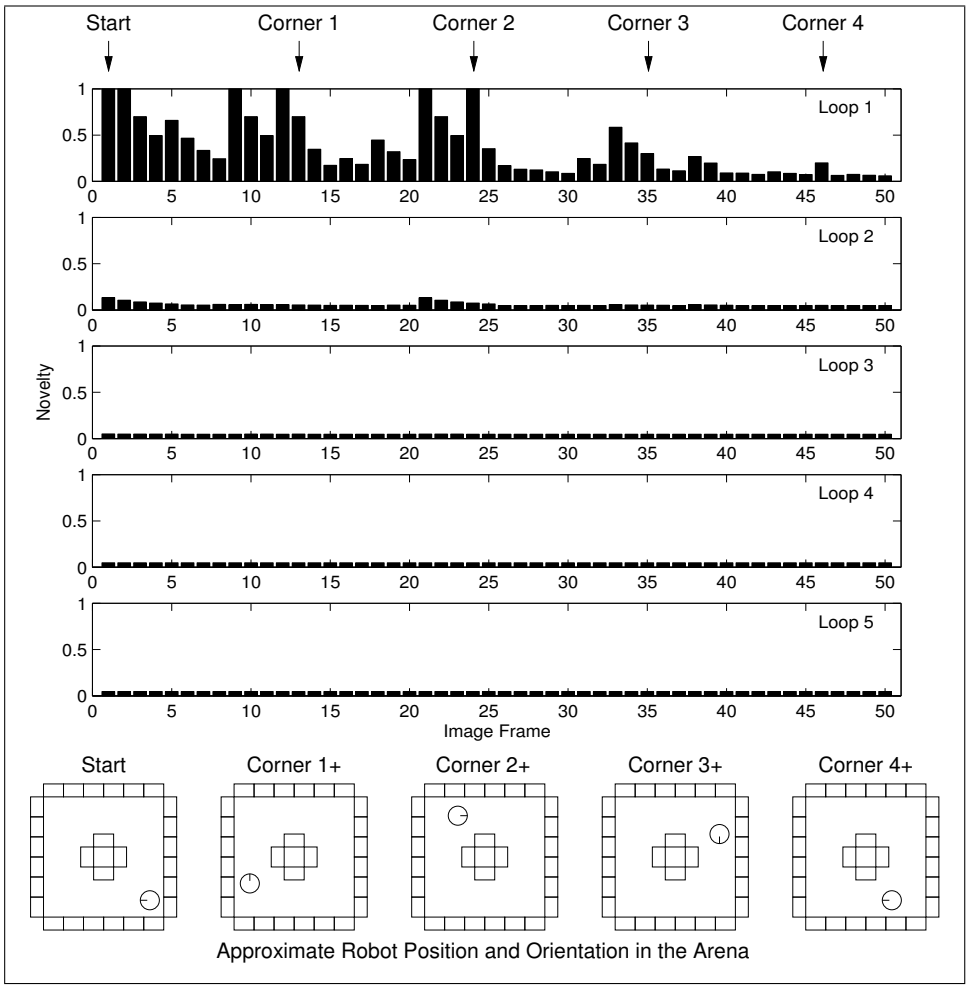



Figure 4.25: Exploration of the new empty arena using global colour angles as image encoding scheme. The GWR network was completely habituated by the end of the second exploration loop.

The GWR network acquired six nodes, all of which were completely habituated by the end of the second loop, which again demonstrates how fast the network training is when using such a compact and robust colour encoding scheme. More nodes were acquired than in the first arena (Experiment 6) and this is attributed to the fact that more details of the walls (drawings and inscriptions on the cardboard boxes) were visible this time — in the first arena, the wooden panels covered most of the details on the walls.

Figure 4.26 shows the successful detection of the orange ball as the new object in most of the frames where it appeared during the inspection phase, with no occurrence of false novelties. The novel object was only missed in frame 26 during the third inspection loop.

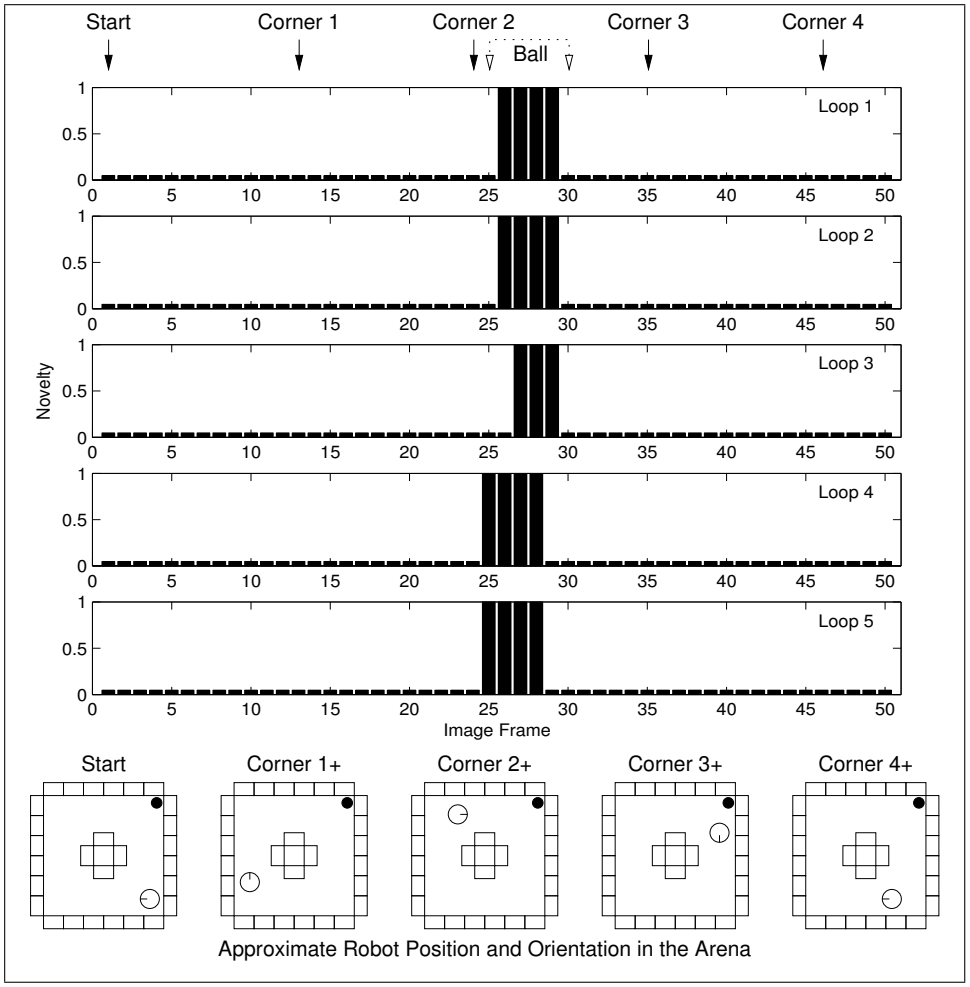



Figure 4.26: Inspection of the new arena with the orange football (novel stimulus) using global colour angles as image encoding scheme. The novel stimulus is correctly and consistently identified by the system.

Inspection was repeated for the arena containing the grey box, but the novelty filter has completely failed to highlight the new object (Figure 4.27) as in the case of local colour angular encoding.

Table 4.4 shows the results of the quantitative analysis for the inspection of the new arena with the orange football using both local and global encoding approaches. The overall results for each experiment, combining the results for both novel objects in a single contingency table are also shown.

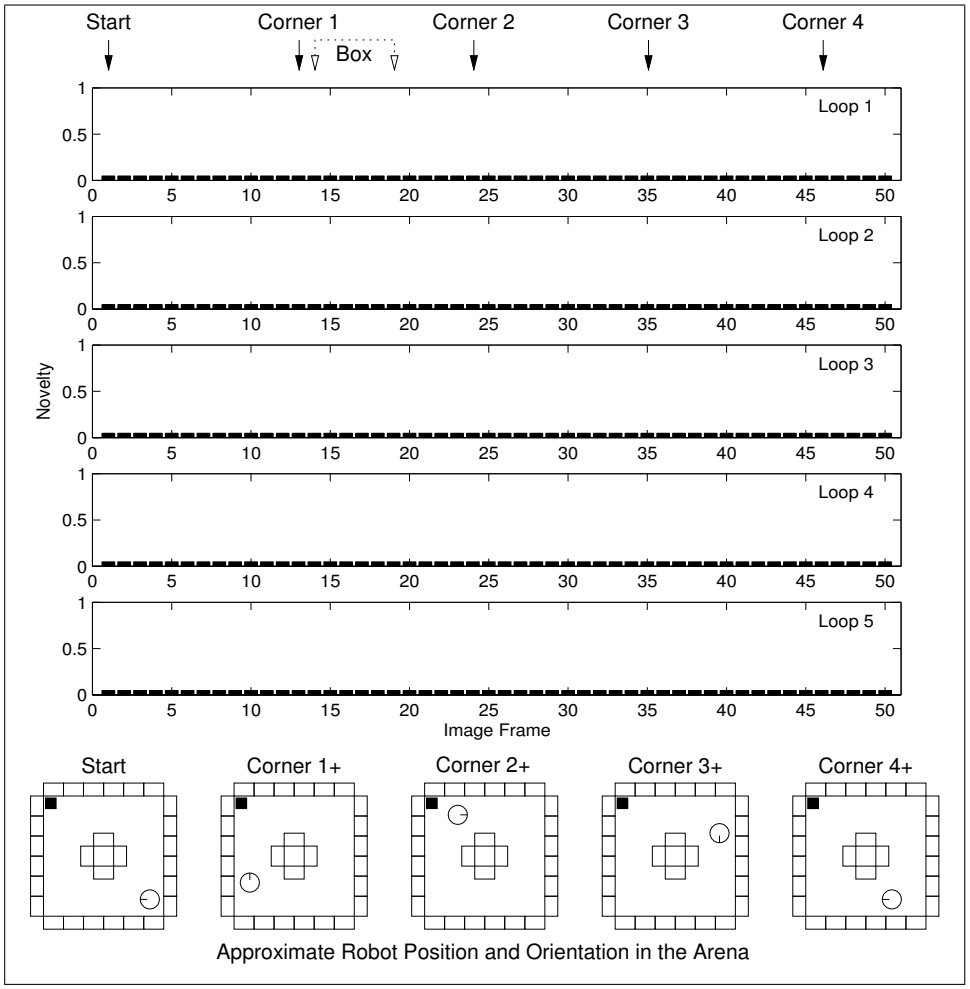



Figure 4.27: Inspection of the new arena with the grey box (novel stimulus) using global colour angles as image encoding scheme. The system fails to detect the novel stimulus.

The results reported in Table 4.4 are quantitatively poorer when compared to the results obtained for the first arena using the same image encoding strategy (see also Table 4.3 on page 92). Inspection of the arena containing the grey box in both approaches (local and global) yielded ill-conditioned contingency tables for the $\chi^2$ test. Failure to detect the grey box correctly has also impaired the "overall" performance (combined performance for the orange ball and the grey box) because colour angular encoding is unable to represent shades of grey, which are largely present in the the new arena. Therefore, an improved colour encoding scheme was needed in order to discriminate grey features in the environment.

Table 4.4: Performance comparison using local and global colour angles while detecting the orange football and the grey box as novel stimuli in the new arena (Experiments 7 and 8). Only inspections of the arena containing the orange ball resulted in statistically significant ($\chi^2$ test, $p \leq 0.01$) association between novelty filter response and actual novelty status.

| | Experiment 7<br>Local (2250 samples) | Experiment 8<br>Global (250 samples) |
|---|---|---|
| Orange ball | $V = 0.50$<br>$U = 0.37$<br>$\kappa = 0.50$ | $V = 0.81$<br>$U = 0.58$<br>$\kappa = 0.80$ |
| Grey box | $V = 0.07^*$<br>$U = 0.01$<br>$\kappa =$ -0.04 | $V = 0.00^*$<br>$U = 0.00$<br>$\kappa = 0.00$ |
| Overall | $V = 0.35$<br>$U = 0.10$<br>$\kappa = 0.33$ | $V = 0.52$<br>$U = 0.22$<br>$\kappa = 0.42$ |

*Ill-conditioned contingency tables for the $\chi^2$ test

## 4.4 Experiments 9 and 10: Novelty Detection from Colour Angles and Intensity Spread

From the previous experiments it became obvious that the angles between the colour vectors of a distribution, although robust to illumination conditions, cannot discriminate unsaturated colours, *i.e.* shades of grey whose $R$, $G$ and $B$ components have the same value. To overcome this situation, we decided to also include intensity statistics in the image encoding.

**Including intensity information.** The intensity information included in the colour representation should also be robust to changes in illumination. Therefore, we need to quantify the intensity distribution in relative terms. The mean intensity of a given region is an absolute measure and therefore is not robust with respect to illumination conditions. Brighter illumination, for example, would increase the mean intensity of the region in question —

what is wanted is an intensity measure that yields the same value for the same image region, as independent as possible of the actual illumination levels.

An adequate statistic in this case is the intensity standard deviation, which represents the variation around the mean intensity, a relative measurement. In fact, changes in illumination intensity would shift all sampled intensity levels, but their standard deviation would tend to remain the same, except when saturation occurs (under-exposure or over-exposure).

We decided to use the normalised standard deviation $\sigma_i$ of the intensity values (see equation 4.1 on page 63) of image regions as an additional element in our image encoding, resulting in feature vectors with four dimensions ($\phi_{rg}$, $\phi_{gb}$, $\phi_{rb}$ and $\sigma_i$).

The normalised standard deviation is defined here as the ratio between the actual standard deviation computed from the intensity distribution and its maximum theoretical value, which provides values ranging from zero to one. The maximum theoretical value for the standard deviation corresponds to the mean value of a distribution with equal number of samples divided between the extreme values (for example, an image containing half pure black and half pure white pixels). In our case, intensity values range from 0 to 255 and therefore the maximum theoretical value for the standard deviation is 127.5.

In mathematical terms, the normalised standard deviation is computed using the following equation:

$$\sigma_i = \frac{\sqrt{\sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} (I_{x,y} - \bar{I})^2}}{127.5XY}, \tag{4.16}$$

where $(x, y)$ are the pixel coordinates, $X$ and $Y$ are the image width and height, respectively, and $\bar{I}$ is the mean intensity value.

### 4.4.1 Experiment 9: Local Colour Angles and Intensity Spread

The data acquired in the new arena was used again for our experiments using local colour angular encoding with added intensity spread. Figure 4.28 shows the GWR network training results obtained for the exploration of the empty arena.

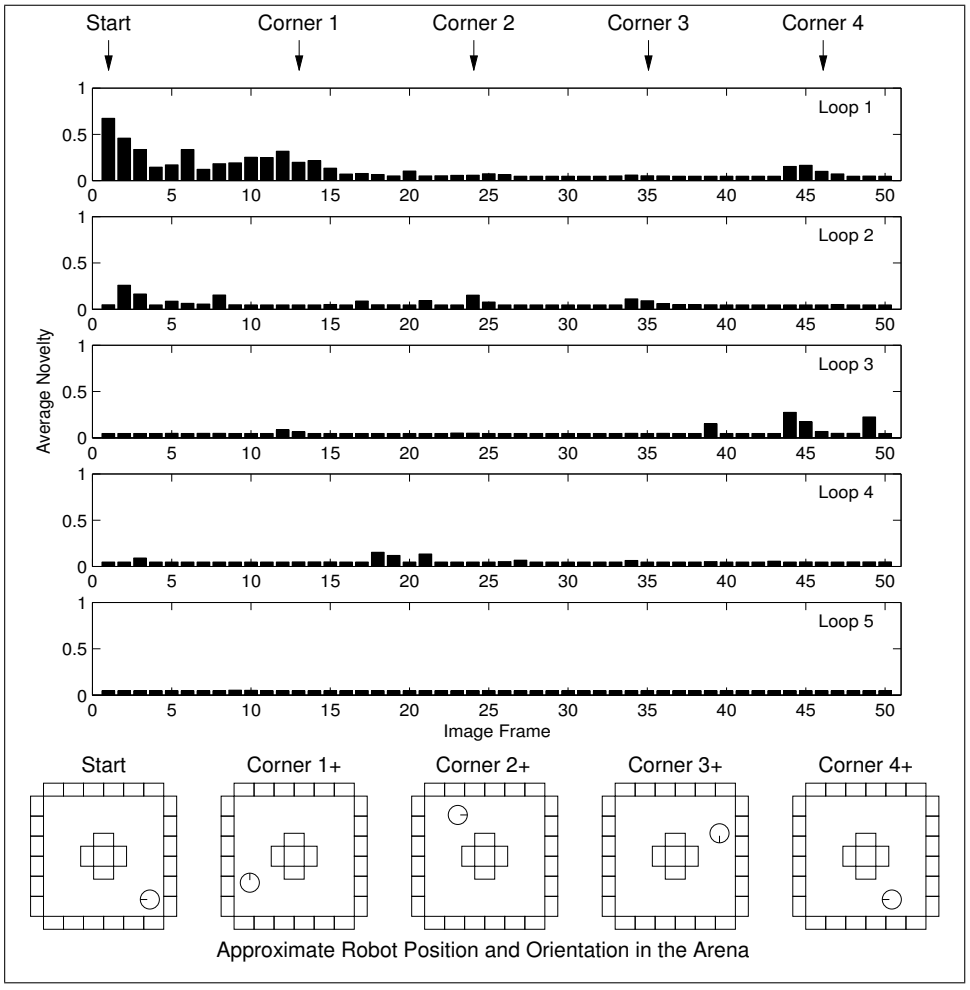



Figure 4.28: Exploration of the new empty arena using local colour angles and intensity spread as image encoding scheme. The GWR network is completely habituated to the environment by the end of the fourth exploration loop.

The GWR network acquired 21 nodes — three nodes more than the acquired using local colour angles alone — and had completely habituated on the environment after the fourth loop. The arena containing new objects was then inspected using the trained GWR network. Results of the inspec-

tion of the arena containing the orange football are given in Figure 4.29, while the results of the inspection of the arena containing the grey box are given in Figure 4.30.

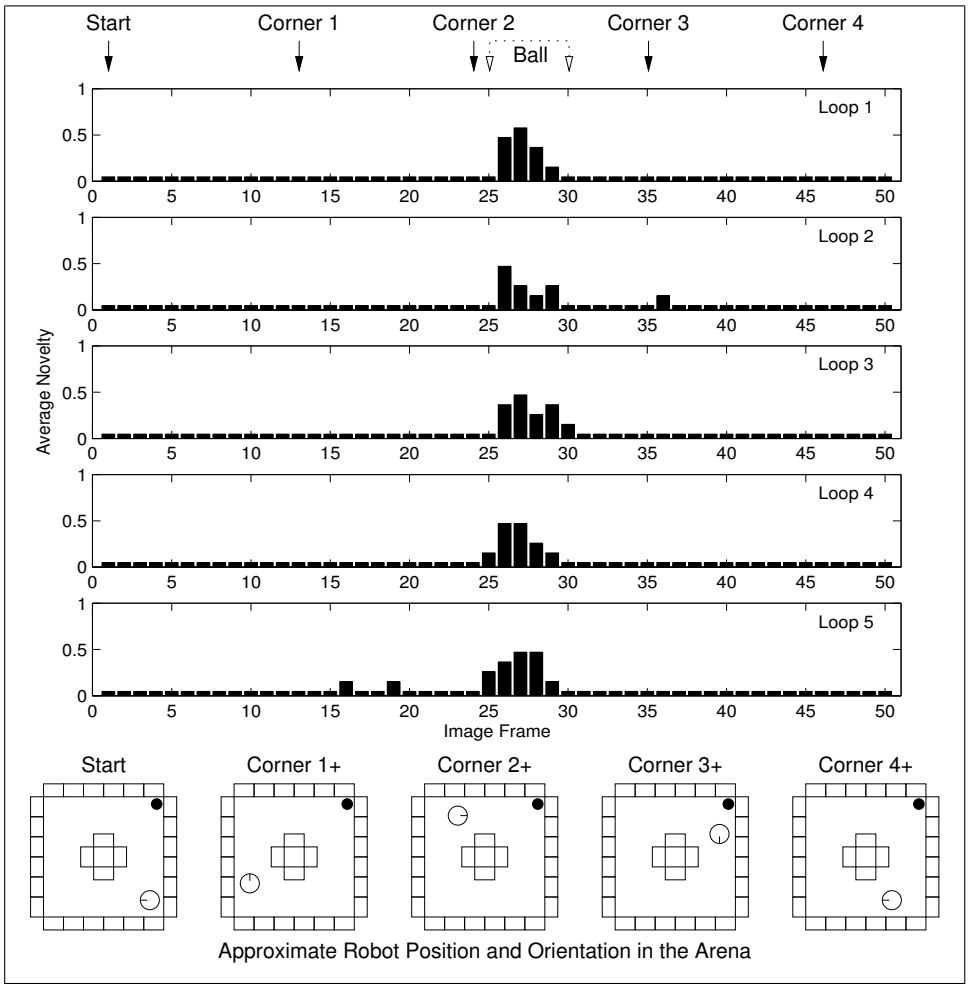



Figure 4.29: Inspection of the new arena with the orange football (novel stimulus) using local colour angles and intensity spread as image encoding scheme. The ball is clearly and consistently detected in every inspection loop around the arena. Also, there are very few false novelties.

Using local colour angular encoding with added intensity spread enabled not only the correct detection of the orange football, as in previous experiments, but also the correct detection of the grey box, which had been "invisible" in the experiments using colour angles alone. In Figure 4.29 the locations where the orange ball was detected are clearly shown and in Figure 4.30 the same happens for the locations where the grey box was detected.

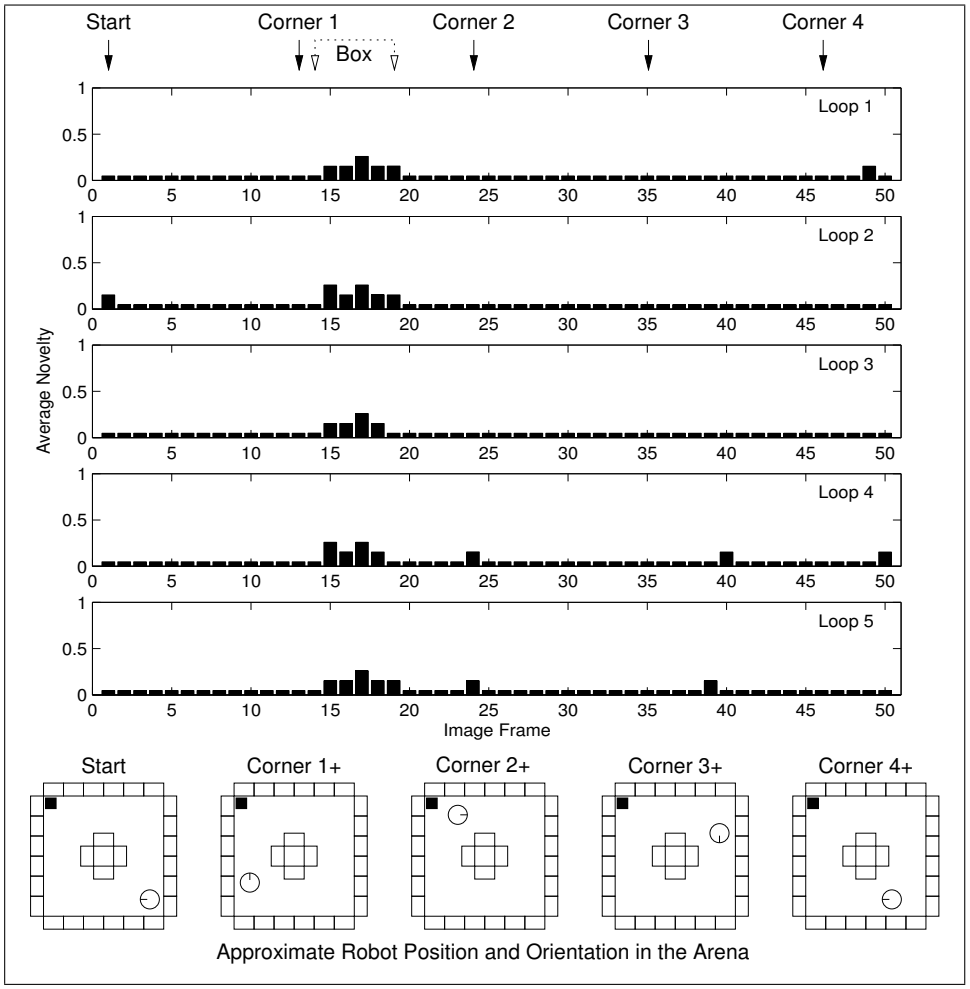



Figure 4.30: Inspection of the new arena with the grey box (novel stimulus) using local colour angles and intensity spread as image encoding scheme. The box is correctly and consistently detected in every inspection loop around the arena, with the occurrence of very few false novelties.

Figure 4.31 depicts an example output image in which the grey box is correctly highlighted as being the novel stimulus (compare with Figure 4.24 on page 97).

Contingency table analysis revealed statistically significant association between novelty filter response and actual novelty status ($\chi^2$ test, $p \leq 0.01$) for both orange ball and grey box during inspection. The quantitative analysis for the case concerning the orange ball yielded $V = 0.63, U = 0.32$ and $\kappa = 0.63$, revealing strong agreement between ground truth and system response. For the case concerning the grey box, the quantitative assessment resulted in $V = 0.66$, $U = 0.36$ and $\kappa = 0.65$, also revealing

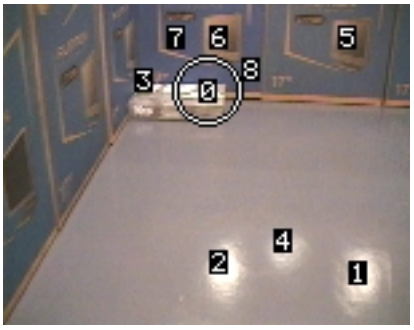



Figure 4.31: Output for an image where the grey box (novel stimulus) appears in the new arena ( colour angular encoding with added intensity spread). The system correctly identified region 0 as being a novel stimulus, but missed region 3.

strong agreement between ground truth data and novelty filter response. Therefore, the experiment using local colour angles and intensity spread for the image encoding stage was the first to yield consistent qualitative and quantitative results for both novel objects.

### 4.4.2 Experiment 10: Global Colour Angles and Intensity Spread

Our final experiments involving colour statistics were conducted using colour angles and intensity spread for the image encoding in a global fashion, *i.e.* without the use of the saliency map as attention mechanism. The same data acquired in the new arena (used in Experiments 7, 8 and 9) was used again in these experiments. Results obtained during the exploration of the arena can be seen in Figure 4.32.

Complete habituation of the GWR network was achieved after the second loop around the empty arena. Six nodes were acquired by the end of the GWR network training, the same amount acquired previously when using global colour angles alone (Experiment 8). A closer analysis of the GWR network trained with global colour angles and intensity spread revealed that the acquired nodes stored very similar information to the information stored in the nodes of the GWR network trained with global colour angles alone.

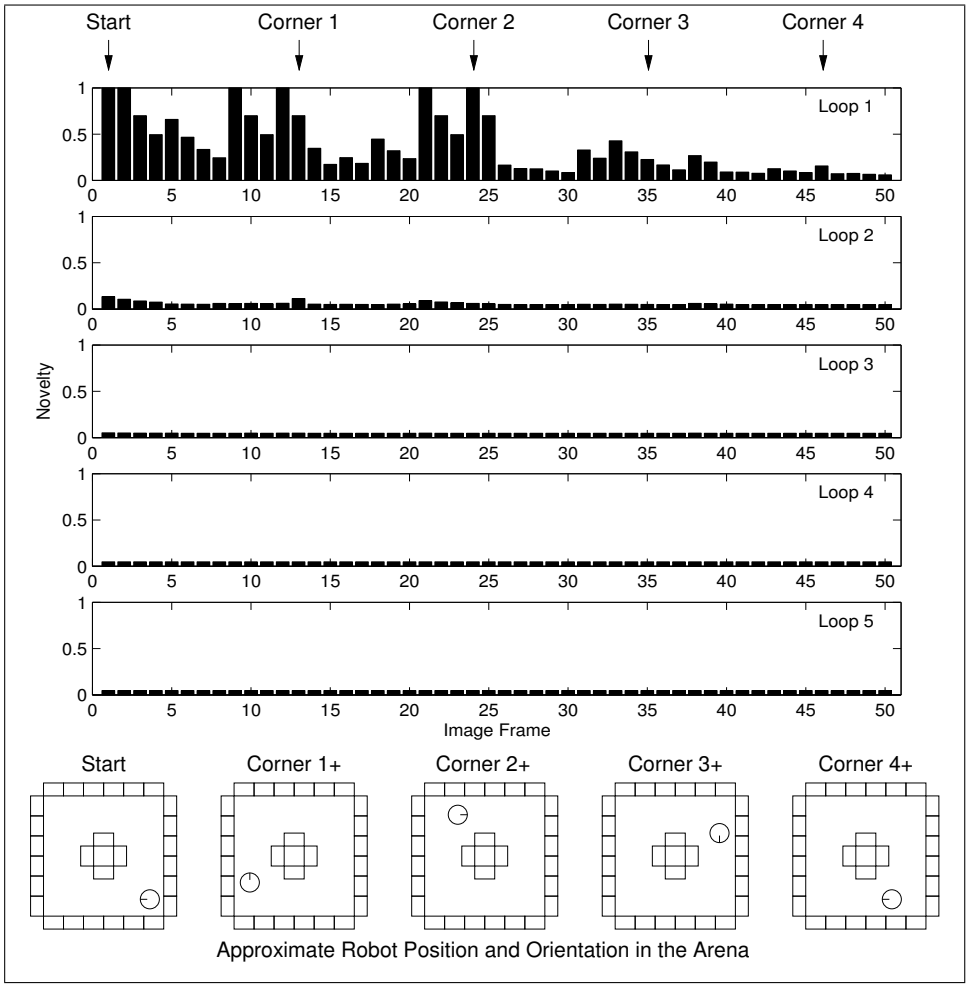



Figure 4.32: Exploration of the new empty arena using global colour angles and intensity spread as image encoding scheme. The GWR network is completely habituated to the environment by the end of the second exploration loop.

Figure 4.33 illustrates the results obtained during the inspection of the new arena containing the orange football. It can be noticed that the ball was correctly highlighted as the novel stimulus. The performance of the novelty filter was slightly better than when using colour angles alone in Experiment 8 (compare the system responses for frame 26 in the third inspection loop in Figure 4.33 and Figure 4.26 on page 99).

However, inspection of the new arena with the grey box revealed that the global approach to colour angular encoding with added intensity spread was unable to identify input image frames in which the novel object appeared (Figure 4.34). Although colour angles with added intensity spread were successful in this task when using the local approach (Experiment 9), the

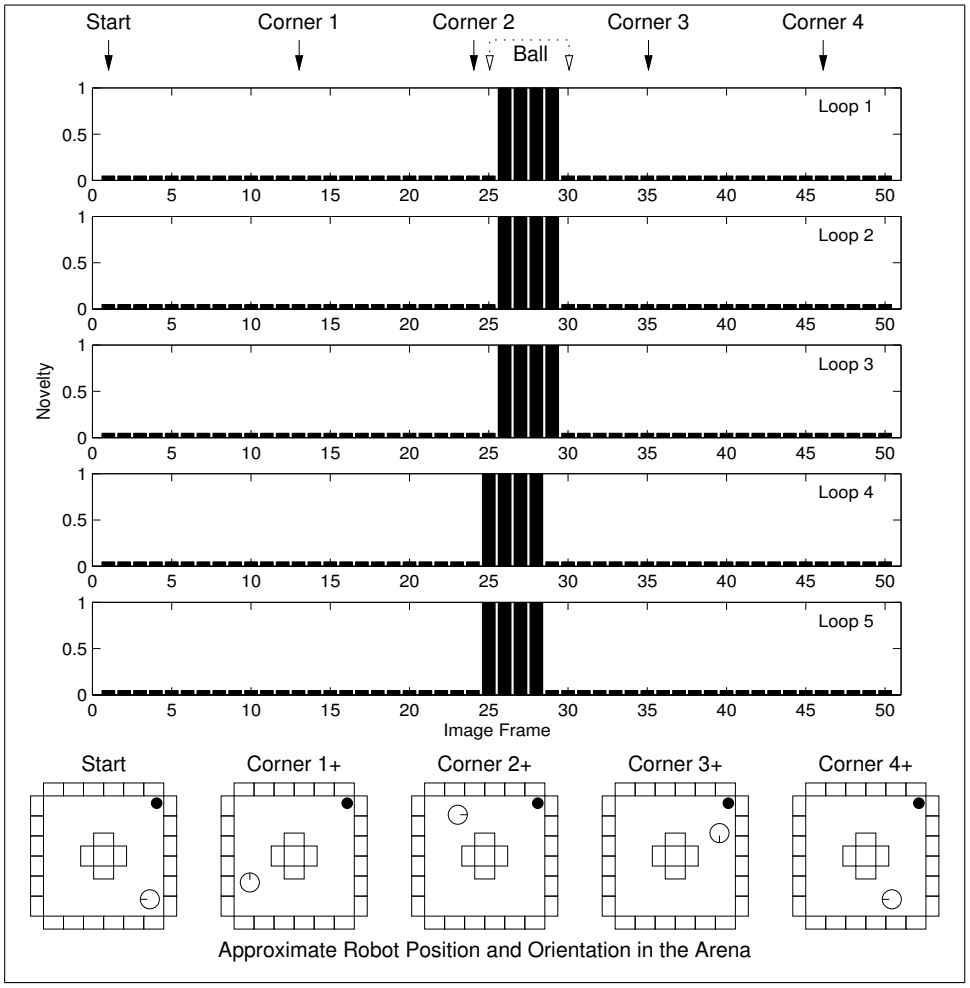



Figure 4.33: Inspection of the new arena with the orange football (novel stimulus) using global colour angles with added intensity spread as image encoding scheme. The ball is correctly detected in every inspection loop around the arena, with no occurrence of false novelties.

additional information provided by the normalised standard deviation of the intensity distribution was not sufficient to discriminate the presence of a relatively small grey object (4.3% of the pixels of the input frame) against a similar background colour.

Results of a quantitative analysis for the inspection of the new arena using the colour angular encoding with added intensity spread are given in Table 4.5. All experiments reflected statistical significance between novelty detected and actual novelty status ($\chi^2$ test, $p \leq 0.01$), except the one involving the global approach as encoding method and the grey box as novel stimulus.

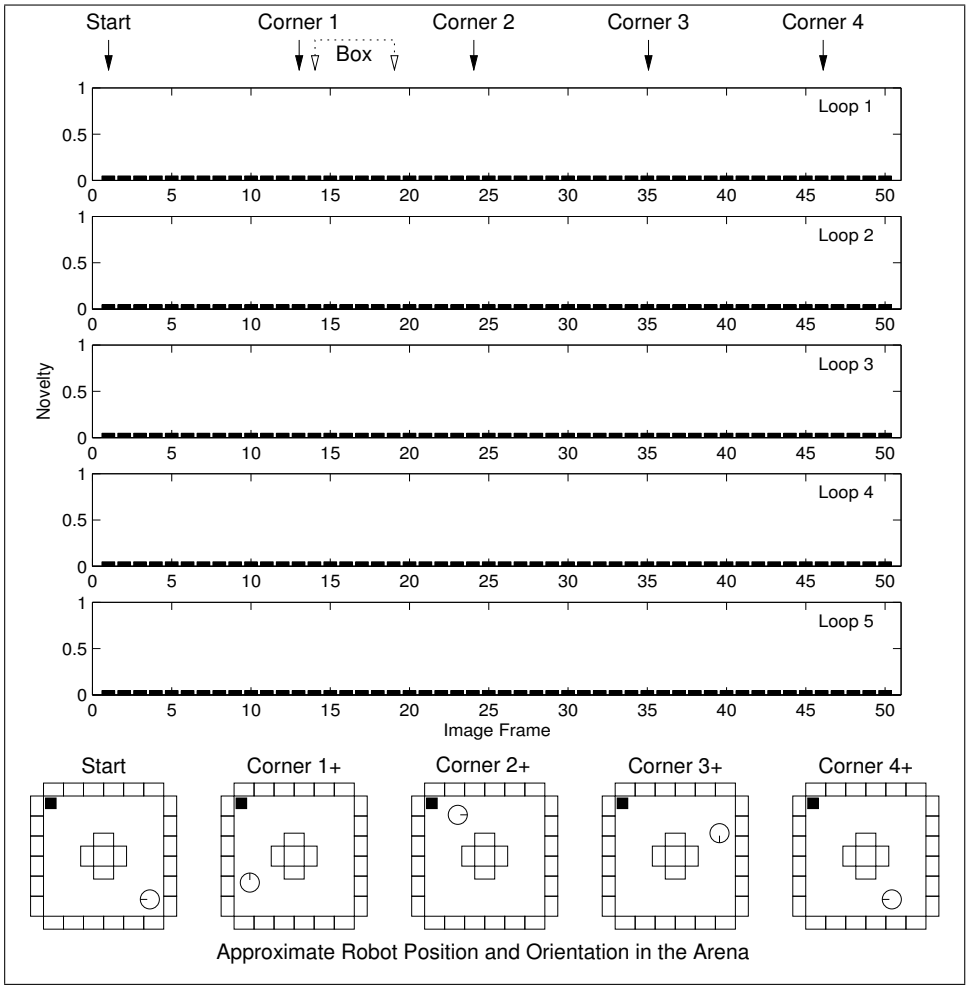



Figure 4.34: Inspection of the new arena with the grey box (novel stimulus) using global colour angles with added intensity spread as image encoding scheme. The system completely fails to detect the novel stimulus.

Table 4.5: Performance comparison using local and global colour angles with added intensity spread while detecting the orange football and the grey box as novel stimuli in the new arena (Experiments 9 and 10).

| | Experiment 9<br>Local (2250 samples) | Experiment 10<br>Global (250 samples) |
|---|---|---|
| Orange ball | $V = 0.63$<br>$U = 0.32$<br>$\kappa = 0.63$ | $V = 0.84$<br>$U = 0.63$<br>$\kappa = 0.83$ |
| Grey box | $V = 0.66$<br>$U = 0.36$<br>$\kappa = 0.65$ | $V = 0.00^*$<br>$U = 0.00$<br>$\kappa = 0.00$ |
| Overall | $V = 0.65$<br>$U = 0.35$<br>$\kappa = 0.65$ | $V = 0.54$<br>$U = 0.24$<br>$\kappa = 0.42$ |

*Ill-conditioned contingency table for the $\chi^2$ test

Although global encoding using colour angles and intensity spread provided the best quantitative results for inspection of the arena with the orange ball, it failed completely during inspection of the arena with the grey box. Local encoding using colour angles and intensity spread was the only approach investigated so far that yielded consistent qualitative and quantitative results (revealing strong agreement between actual novelty status and novelty filter response) for both orange football and grey box.

## 4.5 Summary and Discussion

In this chapter we have presented experiments using colour statistics to perform image encoding within our visual novelty detection framework. The image encoding stage is intended to reduce dimensionality of the input data prior to higher level processing by the actual GWR-based novelty filter.

Also, comparisons between global and local statistics through the use of an attention mechanism were made in order to evaluate the possibility of determining not only *which* image frames contained potential novel visual features, but also *where* these novel features were located within the frames. The use of salient regions to generate local colour statistics made the acquisition of a representation of the operating environment possible, without the installation of any specific *a priori* knowledge.

Experiments using *global* colour histograms have shown that the resolution of the histograms, *i.e.* their number of bins, have a great influence in the novelty filter's overall ability to generalise. On the other hand, histogram resolution did not play such an important role when *local* colour histograms were used in conjunction with the saliency map acting as an attention mechanism. This happens because local histograms are computed within image regions with much smaller number of data samples (pixels) than the entire image frame. When using local image encoding, generalisation is dependent on the number of candidate regions selected by the saliency map.

The colour histograms used here have shown to be useful image descriptors, but they also presented some disadvantages. From the image encoding point of view, they are sensitive to illumination conditions and also do not encode intensity — therefore the colour histograms defined here cannot discriminate shades of grey. Furthermore, from the clustering point of view, the Euclidean distance, used by the GWR network algorithm, is not the ideal metric to measure similarity between histograms. A more appropriate similarity measure would be something like the histogram intersection proposed by Swain and Ballard (1991). The histogram intersection was not used as similarity measure in the experiments reported here because we decided to use exactly the same GWR algorithm and parameters for every experiment, in order to allow consistent performance comparisons between different image encoding mechanisms.

Colour angular encoding (Finlayson et al., 1996) proved to be a very compact and descriptive colour distribution representation that is robust to changes in illumination. Even when colour angles were used in a global fashion, this resulted in successful novelty detection by the GWR network, demonstrating how descriptive this image encoding procedure is. However, colour angles suffer from the same problem as the colour histograms used in this thesis, as they are also unable to discriminate intensity variations.

We found a solution to the issue of discriminating shades of grey by adding the normalised standard deviation of intensity to the standard colour angular encoding scheme. This new approach was able to detect grey features present in the environment which were previously "invisible" to the other discussed image encoding techniques. The best and most consistent experimental results were achieved by computing the colour angular encoding with added intensity spread in the vicinity of locations determined by a saliency-based mechanism of visual attention to form feature vectors. The image encoding procedure as a whole (including the saliency map model)

has shown to be very stable and robust to changes in perspective and perceived illumination. Experiments conducted in engineered environments with a moving robot have demonstrated that this approach has the ability to highlight new, *arbitrary* objects based on their colour characteristics as soon as they first appear in the camera's field of view.

The visual novelty detection configuration discussed in this chapter was able to learn a representation of the robot's normal operating environment quickly. The acquired representation was later used to detect any unusual colour features that were introduced after training. Our implementation is capable of on-line learning at eight frames per second when running autonomously on the Essex Magellan Pro robot (*Radix*), which is equipped with an 850MHz Pentium III processor.

Finally, we compared the performance of our mechanism to manually-generated ground truth and observed statistically significant correlation between them ($\chi^2$ analysis, $p \leq 0.01$). Most results from our experiments interfacing visual stimuli to the GWR-based novelty filter were very good. Our framework proved to be stable and robust to general image transformations and to provide consistent novelty detection based on local colour statistics. Tables 4.6 and 4.7 summarise the quantitative results achieved for each image encoding strategy discussed in this chapter.

The results in Table 4.6 refer to the experiments conducted in the first arena with the orange football as novel stimulus. Except for the global encoding using colour histograms with 32 and 64 bins, all other approaches presented statistically significant correlation between system response and actual novelty status ($\chi^2$ analysis, $p \leq 0.01$). The best results (almost complete agreement between system response and novelty ground truth) were achieved with the use of local colour angles.

Results in Table 4.7 correspond to the experiments carried out in the second arena, having the orange football and the grey box as novel stimuli.

Table 4.6: General performance comparison using local and global approaches while detecting the orange football as novel stimuli in the first arena.

Orange Ball

| | Local (2025 samples) | Global (225 samples) |
|---|---|---|
| Colour Histograms (32 bins) | $V = 0.75$<br>$U = 0.53$<br>$\kappa = 0.74$ | $V = 0.08^*$<br>$U = 0.02$<br>$\kappa =$ -0.07 |
| Colour Histograms (64 bins) | $V = 0.68$<br>$U = 0.46$<br>$\kappa = 0.67$ | $V = 0.08^*$ ($0.34^*$)<br>$U = 0.01$ (0.09)<br>$\kappa = 0.08$ (0.24) |
| Colour Angles | $V = 0.84$<br>$U = 0.73$<br>$\kappa = 0.82$ | $V = 0.89$<br>$U = 0.72$<br>$\kappa = 0.89$ |

*Ill-conditioned contingency tables for the $\chi^2$ test, values in brackets correspond to exploration using only four loops — see subsection 4.1.2 for details

Table 4.7: General performance comparison using local and global approaches while detecting the orange football and the grey box as novel stimuli in the second arena.

Orange Ball

| | Local (2250 samples) | Global (250 samples) |
|---|---|---|
| Colour Angles | $V = 0.50$<br>$U = 0.37$<br>$\kappa = 0.50$ | $V = 0.81$<br>$U = 0.58$<br>$\kappa = 0.80$ |
| Colour Angles + Intensity Spread | $V = 0.63$<br>$U = 0.32$<br>$\kappa = 0.63$ | $V = 0.84$<br>$U = 0.63$<br>$\kappa = 0.83$ |

Grey Box

| | Local (2250 samples) | Global (250 samples) |
|---|---|---|
| Colour Angles | $V = 0.07^*$<br>$U = 0.01$<br>$\kappa =$ -0.04 | $V = 0.00^*$<br>$U = 0.00$<br>$\kappa = 0.00$ |
| Colour Angles + Intensity Spread | $V = 0.66$<br>$U = 0.36$<br>$\kappa = 0.65$ | $V = 0.00^*$<br>$U = 0.00$<br>$\kappa = 0.00$ |

*Ill-conditioned contingency tables for the $\chi^2$ test

Overall (Orange Ball and Grey Box)

| | Local (2250 samples) | Global (250 samples) |
|---|---|---|
| Colour Angles | $V = 0.35$<br>$U = 0.10$<br>$\kappa = 0.33$ | $V = 0.52$<br>$U = 0.22$<br>$\kappa = 0.42$ |
| Colour Angles + Intensity Spread | $V = 0.65$<br>$U = 0.35$<br>$\kappa = 0.65$ | $V = 0.54$<br>$U = 0.24$<br>$\kappa = 0.42$ |

All experimented approaches yielded statistically significant association between the system response and manually generated novelty ground truth ($\chi^2$ analysis, $p \leq 0.01$) for the cases in which the orange ball was the novel stimulus. However, for the cases in which the novel stimulus was the grey box, only the local colour angular encoding with added intensity spread was able to discriminate the novel object correctly.

# Chapter 5

# Experiments using Raw Image Data

The experiments using colour statistics to perform image encoding within our visual novelty detection framework have yielded successful results, both qualitatively and quantitatively, in an engineered environment. However, image encoding based on colour statistics does not take other important visual features into account — topological relationship between pixels, for example — and does not hold enough information to reconstruct the original image. In fact, if one desires to examine the weights of the GWR network nodes and analyse which visual aspects of the environment were acquired during learning, the best information that can be retrieved in this case is the relative amount of different colours present in a given image region.

Often, depending on the application not only colour, but also texture and shape need to be encoded. This would allow the system to differentiate between two objects with exactly the same colour distribution, but with differences in shape or texture, for instance. Obviously, by adding more information to the image encoding, some of its generalisation properties are lost and may need to be accounted for in alternative ways. Again, the degree of robustness to changes in appearance is application-specific.

Therefore, at this stage in our experiments we wanted to go one step further and encode image structure as well as colour information. Also, we wanted to obtain an image encoding procedure that allows image reconstruction so that examination of the GWR network nodes would provide useful information about which elements of the environment were actually learnt.

The design of an image encoding procedure that meets these requirements is not easy because it is not always clear which elements of the data are the most relevant. Hence, we decided to use normalised raw image patches as input vectors to the GWR network. For that, we rely on the assumption that the attention model provides reasonably stable and accurate interest points. Good accuracy is necessary to minimise alignment errors when matching image patches using the Euclidean distance metric, which is part of the GWR algorithm and operates in a pixel-by-pixel basis.

Using *RGB* image patches with $24 \times 24$ pixels in size — the same size used before during the experiments reported in Chapter 4, results in input vectors with $24 \times 24$ pixels $\times$ 3 colours = 1728 elements, which were normalised to unit length in order to even out lighting conditions. Also, as a side-effect, normalisation of input vectors reduces the input domain to the surface of a unit hyper-sphere in input space, which makes the task of classification easier for the clustering mechanism in use.

An input space of 1728 dimensions is large and may in fact not be necessary. We were therefore interested in reducing the input space dimensionality. One way to achieve this is to use Principal Component Analysis (PCA), which works by projecting the data onto principal axes (eigenvectors), in which variance is maximised. Because the principal axes are obtained from the data itself, PCA is effectively a bottom-up mechanism that automatically selects the most relevant parts of the data.

A particularly interesting characteristic of the incremental PCA algo-

rithm used in this work (Artač et al., 2002) is that the process of deciding if the current PCA model (eigenspace) needs to be incremented is by itself a novelty filter. If a given input is not well represented by the current eigenmodel and it needs to be updated, then this input must be novel. In the scope of incremental PCA, dimensionality reduction is achieved by exploiting the fact that the number of eigenvectors in the model are likely to be less in number than the dimensionality of the input vectors. Further dimensionality reduction can be achieved by keeping only the eigenvectors corresponding to the largest variances (eigenvalues) in the model at the expense of losses in reconstruction (and possibly in the overall recognition rate of the system). If all eigenvectors are kept in the model, perfect reconstruction of the original data is achieved. Reconstruction of the input image patches from the stored projected vectors provides valuable information about which aspects of the environment were learnt.

In this chapter we report experiments using raw image patches and two distinct novelty filters, one based on the GWR neural network and the other on incremental PCA. Their performances were compared while operating in engineered laboratory environments and also medium-scale real world environments. We maintained the same parameter values from previous experiments with the GWR network ($a_T = 0.9$, $h_T = 0.5$, $\eta = 0.1$, $\epsilon = 0.1$, $\tau = 3.33$, $\alpha = 1.05$, $h_0 = 1$, $S(t) = 1$ and $age_{max} = 20$) and, concerning the incremental PCA algorithm, we have set the residual error threshold to $r_T = 0.25$ and kept only the eigenvectors whose corresponding eigenvalues were larger than 1% of the largest eigenvalue in the model. The residual error threshold ($r_T$) establishes if the input vector reconstruction from the current PCA model is similar enough to the original input vector. In other words, it determines if the input vector is well represented by the PCA model or if it constitutes a novel input (more details are given in Subsection 2.1.2).

Figure 5.1 shows the block diagram of the approach followed in this

chapter. The saliency map was used as an attention mechanism to select raw image patches in the input frame. These patches were then normalised to unit-length vectors and fed to either of two distinct novelty filters, one of them based on the GWR neural network and the other based on incremental PCA, which indicated the presence or not of novelty.

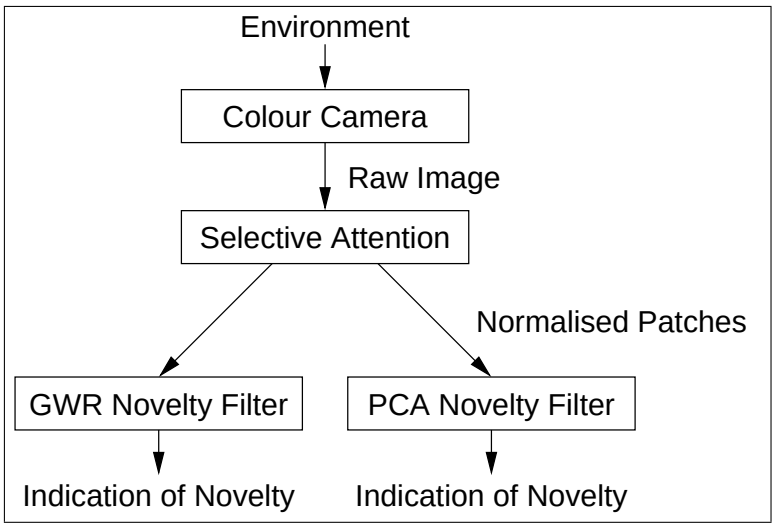



Figure 5.1: Novelty detection using raw image patches: the attention model selects candidate patches and feeds them (after normalisation) to a novelty filter based either on the GWR network or incremental PCA.

## 5.1 Experiment 11: Novelty Detection from Raw Image Patches

**Using the GWR network as novelty filter.** In this experiment using raw image patches, we used the same set of images used for the experiments using local colour angles and instensity spread (see Subsection 4.4), acquired from the arena built at the robotics research laboratory at Essex. As in previous experiments, learning of the GWR network occurred during the exploration of the empty arena in five loops. The graphs depicted in Figure 5.2 show the average novelty detected in each frame during exploration.

The novelty graphs in this chapter were plotted in a different way than

the previous ones presented in Chapter 4. Because the incremental PCA algorithm provides a binary indication of novelty as output — as opposed to the GWR network, which provides a measure of novelty in the range [0.05, 1] — the measure of novelty yielded by the GWR network needs to be thresholded in order to allow fair qualitative comparisons. Therefore, the output of the GWR network was thresholded using its own habituation threshold parameter $h_T$, resulting in a binary indication of novelty.

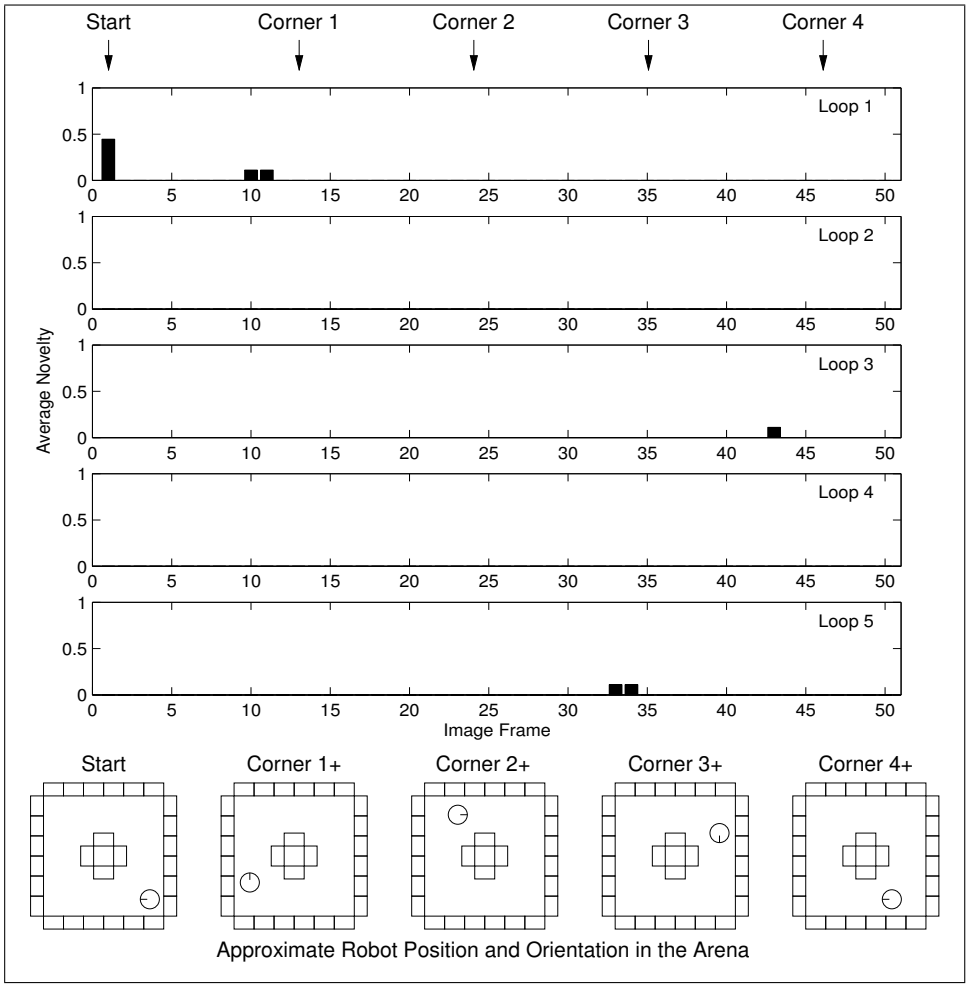



Figure 5.2: Exploration of the empty arena using raw image patches and the GWR network. Most of the novelty indications happen in the first image frame. By the end of the fifth exploration loop, the network acquired four nodes.

From Figure 5.2 it can be noticed that learning of the GWR network was very fast, with most of the novelty activity having happened in the first image frame. Only four nodes were acquired by the GWR network by the end of the fifth exploration loop, which is surprising given that 21

nodes were acquired when using local colour angles and intensity spread (see Subsection 4.4.1 on page 103). We expected that the use of raw image data, which is much more specific than colour statistics for image representation, would result in a larger number of acquired nodes.

The number of acquired nodes is related to the input data distribution and the choice of activation threshold $a_T$, which controls the size of clusters of the GWR network. However, increasing $a_T$ (*i.e.* reducing the size of GWR clusters) in order to increase the number of acquired nodes can also result in poor generalisation. In future research we intend to determine the size of clusters automatically from the input data distribution.

As in previous experiments, the acquired model of normality was used to inspect the arena containing an object that was not present during exploration. Results of inspection of the arena with the orange football are presented in Figure 5.3.

The graphs depicted in Figure 5.3 demonstrate the ability of the GWR network to detect consistently the locations in which the orange ball appeared, when using normalised raw image patches as input vectors. Inspection was repeated, but this time the orange football was removed from the arena and the grey box was inserted in a different corner. Figure 5.4 shows the results obtained for this new inspection.

The grey box was also correctly identified as novel by the GWR network, according to the graphs in Figure 5.4. However, unexpected novelty peaks also appeared consistently for image frame 46, which demands further analysis to find an explanation. Figure 5.5 shows the visual output for image frame 46 in the fifth inspection loop.

The output image frame depicted in Figure 5.5 shows that in this case the robot was turning a corner, very close to the wall of the arena. In this particular image frame, six of the nine most salient regions correspond to a large edge between two of the cardboard boxes that constitute the wall.

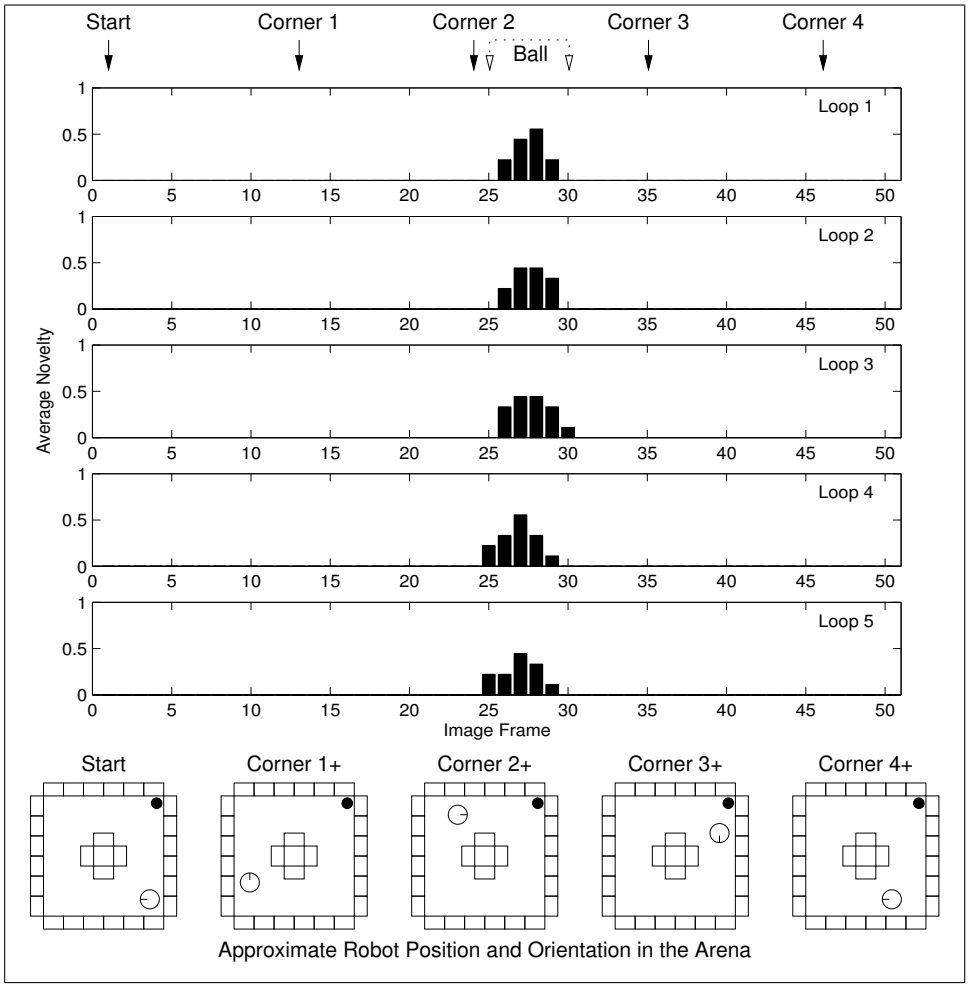



Figure 5.3: Inspection of the arena with the orange ball (novel stimulus) using raw image patches and the GWR network. The orange ball is clearly and consistently highlighted as novelty.

Although the robot was exposed before to edges between the cardboard boxes, it has never before been as close to them as happened in this case — this resulted in image patches containing edges larger in scale than the GWR network was habituated to, ultimately resulting in their classification as novel. In this same output image, it can be noticed also that regions 0, 1 and 5 are aligned with the darkest part of the edge, while regions 2, 4 and 7 are misaligned. Both robustness to scale and misalignment will therefore be addressed further in Chapter 6.

**Using incremental PCA as novelty filter.** In order to have a baseline to compare the performance of the GWR network using raw image

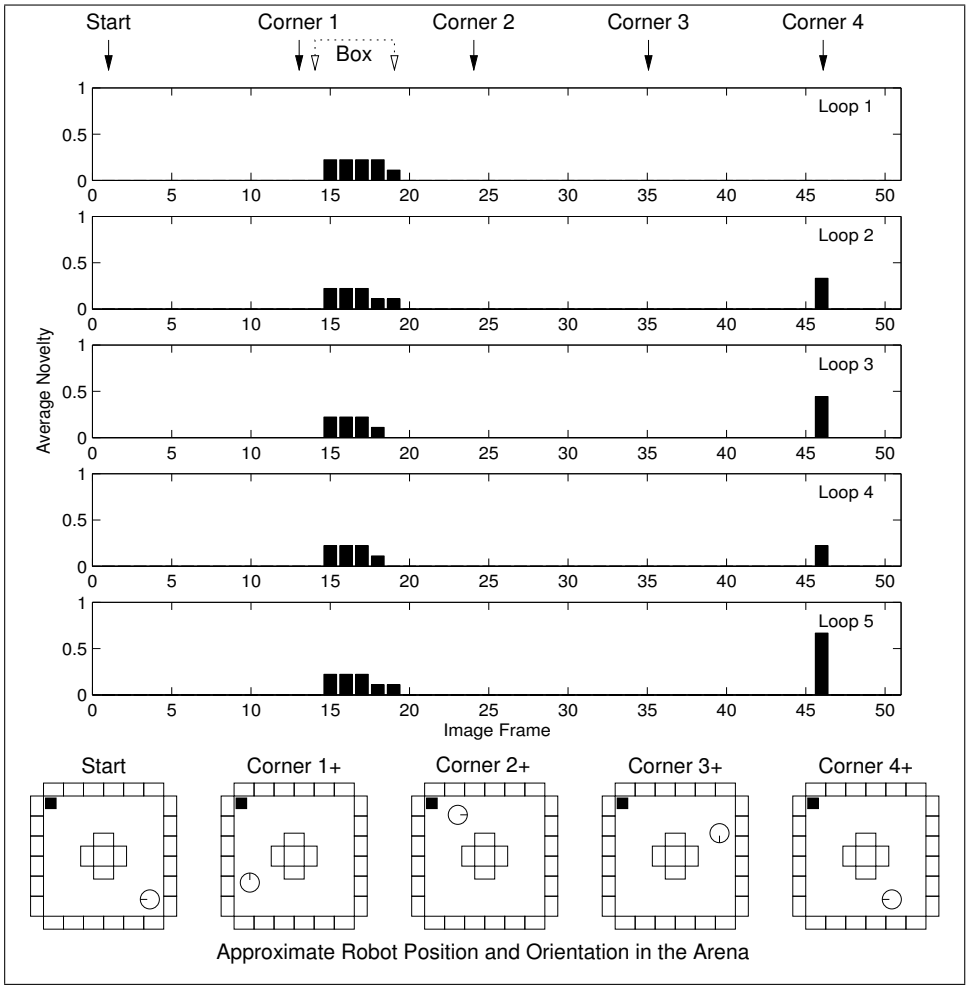



Figure 5.4: Inspection of the arena with the grey box (novel stimulus) using raw image patches and the GWR network. The grey box is clearly and consistently highlighted as novelty. Unexpected novelty indications also appeared consistently for image frame 46.

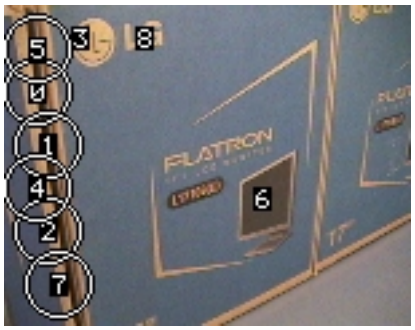


Figure 5.5: Image frame 46 (fifth loop of inspection of the arena with the grey box). A large edge between two cardboard boxes exposes weaknesses regarding feature scale and patch misalignment.

patches, we repeated the experiment using the incremental PCA approach. Figure 5.6 shows the novelty graphs obtained during the five loops of the exploration of the empty arena using incremental PCA.

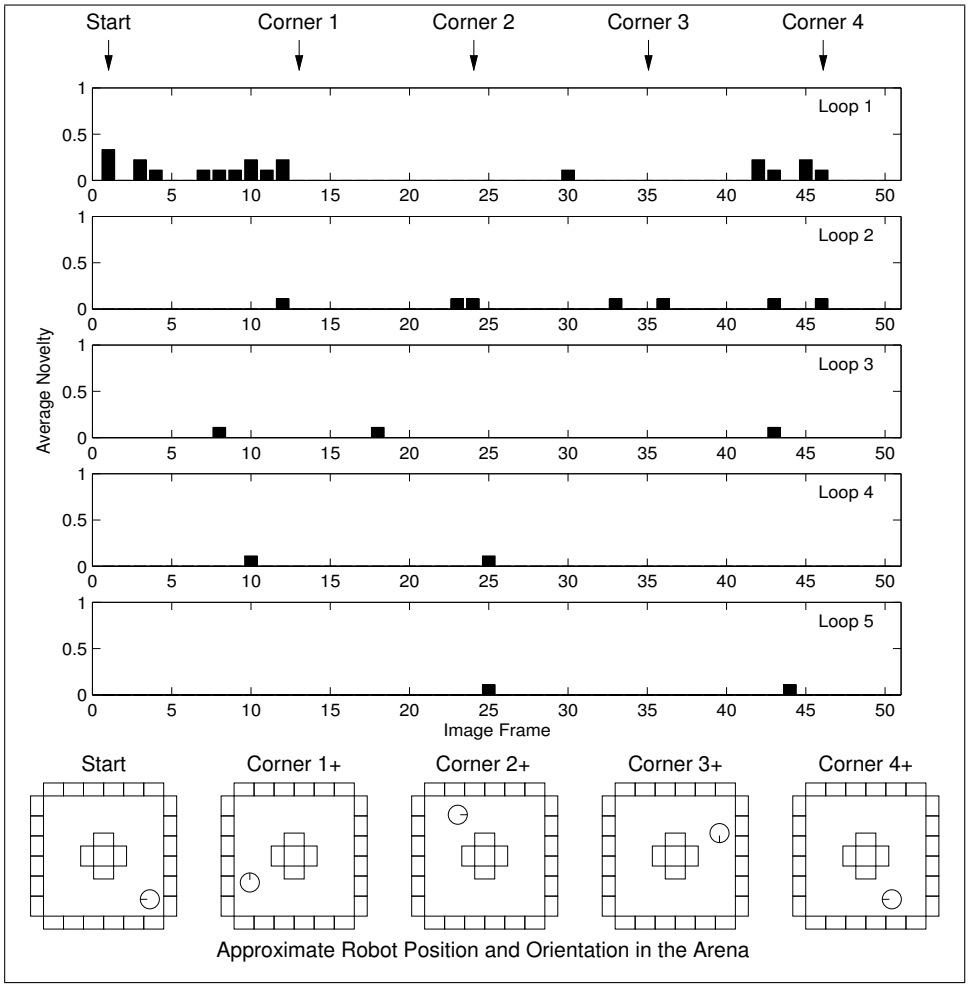



Figure 5.6: Exploration of the empty arena using raw image patches and incremental PCA — novelty activity decreases as the robot explores the arena.

It can be seen in Figure 5.6 that most of the eigenspace updates happened in the beginning of the first loop, becoming less frequent as the environment was explored. The incremental PCA approach has acquired 35 model vectors with 33 dimensions, representing a compression factor of more than 98%. However, an eigenvector with the original 1728 dimensions was also stored for each of the 33 dimensions of the acquired model vectors. Therefore, the total amount of memory used by incremental PCA was more than eight times larger than that used by the GWR network.

As before, the model learnt during the exploration phase was used to highlight novel visual features in the arena during the inspection phase, when the learning mechanism was disabled. The results obtained for the inspection of the arena with the orange ball are given in Figure 5.7.

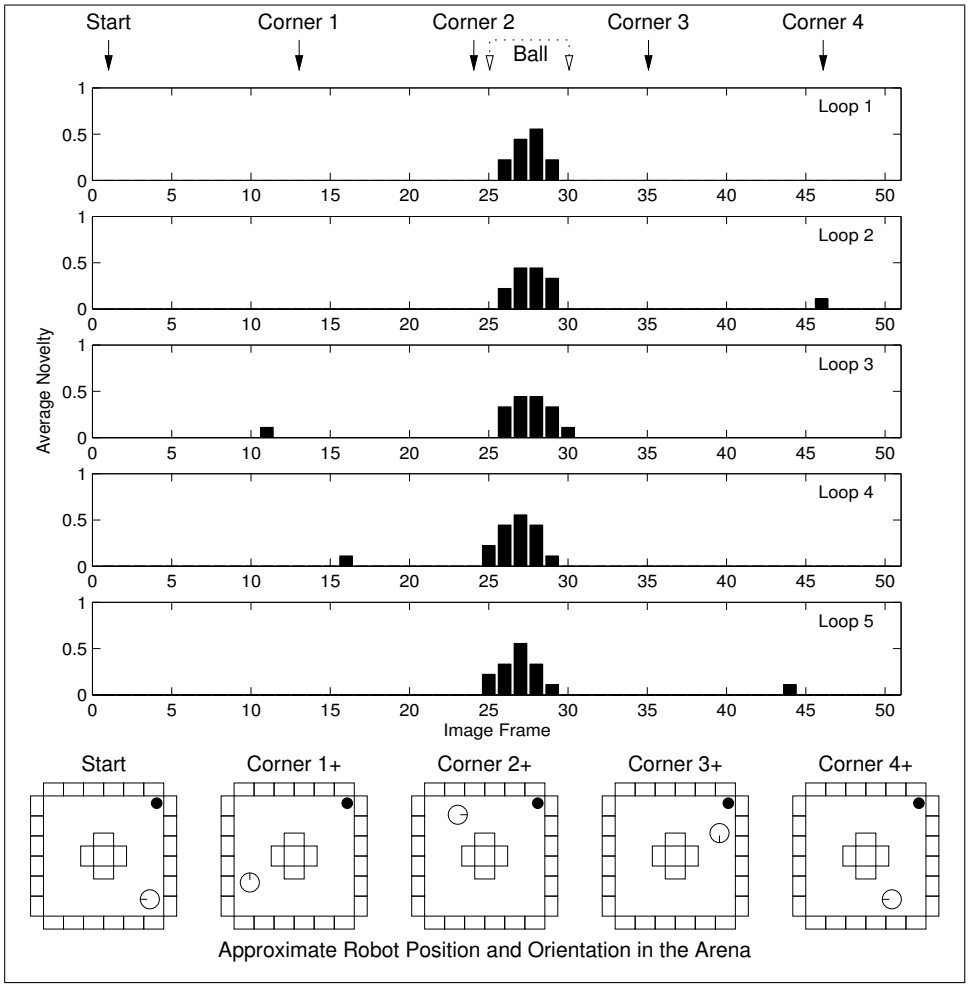



Figure 5.7: Inspection of the arena with the orange ball (novel stimulus) using raw image patches and incremental PCA. The orange ball is clearly and consistently highlighted with very few unexpected novelty indications.

The orange ball was correctly identified as the new entity in the environment as the graphs in Figure 5.7 show. Also, there were very few false indications of novelty. Inspection was repeated for the arena containing the grey box (the orange football was removed from the arena) and the results obtained are given in Figure 5.8.

The grey box was also correctly highlighted by the incremental PCA approach with very few spurious novelty indications. Incremental PCA coped better with the robot getting closer to the arena's walls (*e.g.* the large scale edge present in frame 46). This can be attributed to the choice of parameters which influence generalisation and does not necessarily mean that incremental PCA performs better than the GWR network in this respect.

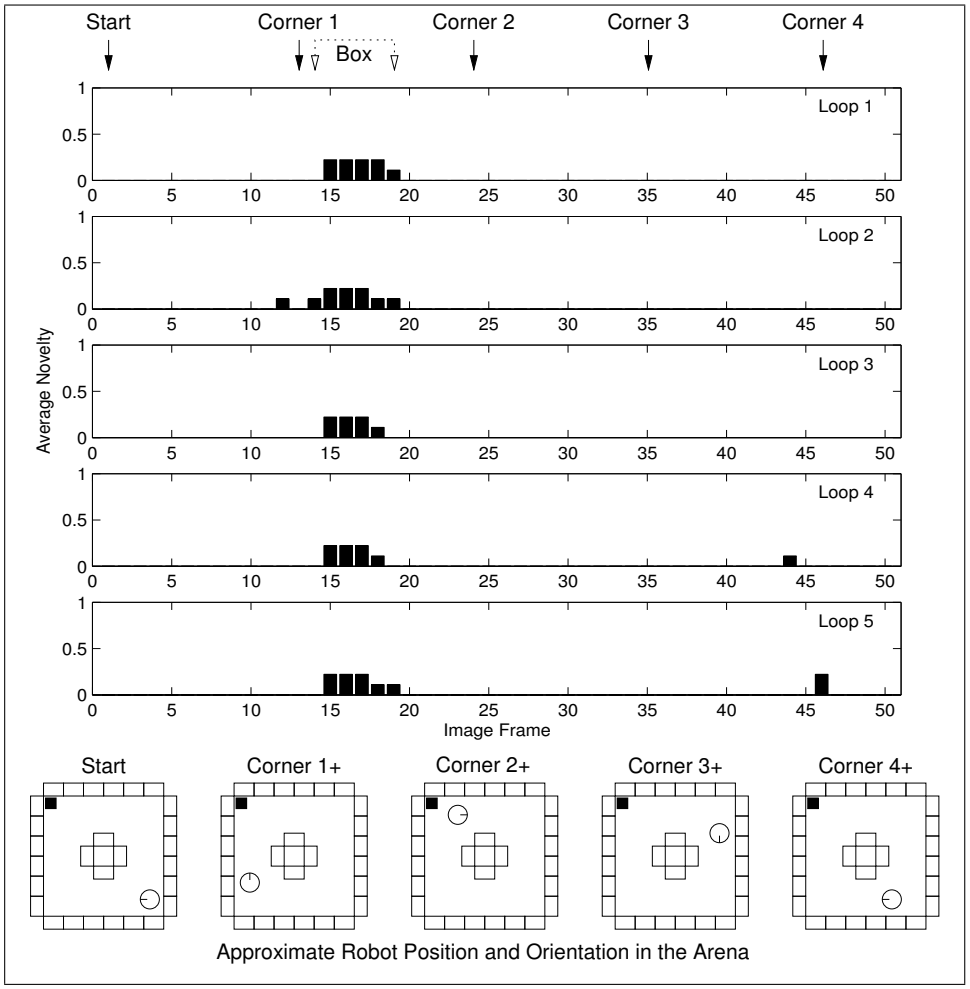



Figure 5.8: Inspection of the arena with the grey box (novel stimulus) using raw image patches and incremental PCA. The grey box is correctly highlighted as novelty with very few unexpected novelty indications.

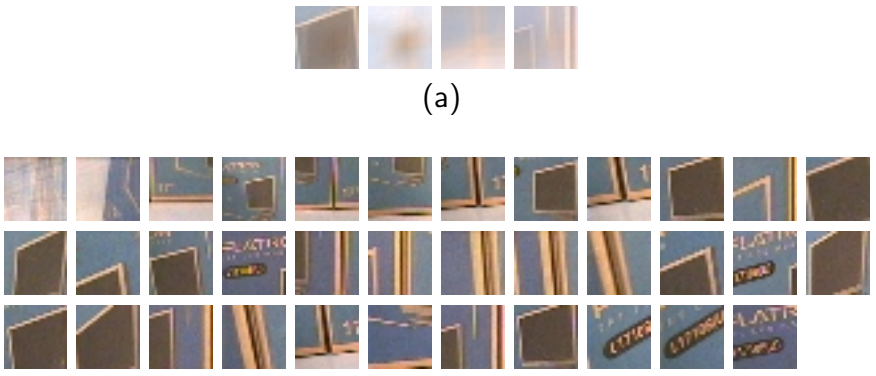
Table 5.1 shows a quantitative comparison between the results obtained using raw image patches for both the GWR network and the incremental PCA approach. All cases presented statistically significant correlation between novelty filter response and actual novelty status according to the $\chi^2$ analysis ($p \leq 0.01$).

Overall performances (combined performances for the orange ball and the grey box) of both approaches are quantitatively very similar (almost complete agreement between novelty filter response and ground truth), although the incremental PCA algorithm yielded slightly better and more consistent overall results.

Table 5.1: Performance comparison using raw image patches — Experiment 11 (2250 samples). All results correspond to statistically significant correlation between system response and actual novelty status ($\chi^2$ analysis, $p \leq 0.01$).

| | GWR Network | Incremental PCA |
|---|---|---|
| Orange ball | $V = 0.91$<br>$U = 0.74$<br>$\kappa = 0.91$ | $V = 0.86$<br>$U = 0.68$<br>$\kappa = 0.86$ |
| Grey box | $V = 0.70$<br>$U = 0.44$<br>$\kappa = 0.70$ | $V = 0.83$<br>$U = 0.60$<br>$\kappa = 0.83$ |
| Overall | $V = 0.82$<br>$U = 0.58$<br>$\kappa = 0.82$ | $V = 0.85$<br>$U = 0.64$<br>$\kappa = 0.85$ |

**Reconstruction of image patches from the model of normality.** Using raw image patches provides the extra functionality of being able to reconstruct the acquired image patches from the GWR network weight space or the PCA space, depending on the case. Figure 5.9 shows these reconstructed patches for both GWR and PCA approaches.



(a) (b)

Figure 5.9: Image patches acquired during Experiment 11: (a) reconstructed from the GWR network; (b) reconstructed from incremental PCA.

It can be noticed that the reconstruction of the weight vectors from the GWR network results in averaged image patches. This happens because of the learning procedure used in the GWR network algorithm (see equation 2.12 on page 23). On the other hand, the reconstructed images from the PCA space are very faithful. The incremental PCA also has provided a more detailed representation of the arena, including some models of large edges

(see image patches in the middle-right of the second row in Figure 5.9b), which are similar to the one present in frame 46 during the inspection of the arena with the grey box (see Figure 5.5). This explains why the incremental PCA performed better during the inspection phase of the arena with the grey box. The reconstructed image patches in Figure 5.9a show that the averaging due to the GWR node insertion and adaptation rules does not produce good results and in fact causes significant distortions in the original input patterns. An alternative on-line clustering approach that only uses "real" input data samples as cluster centres was proposed by Angelov (2004) and constitutes a good option for future research.

## 5.2 Experiment 12: Saliency versus Novelty

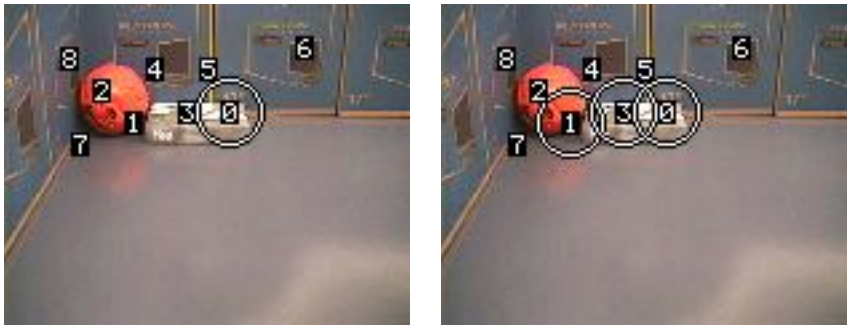
Having obtained successful results in the previous experiments, we wanted to establish that the task of novelty detection was actually achieved by the novelty filters, rather than by the attention mechanism, which was employed to make a pre-selection of image patches to be evaluated for possible novelty.

Therefore, we conducted a sixth round of experiments in which the robot explored the arena containing the conspicuous orange football and afterwards inspected it with the inclusion of the inconspicuous grey box. Two different situations were analysed:

- First, the grey box was placed in a different corner than the ball, fact that does not affect the performance of the attention mechanism observed in previous experiments (the two objects of interest are never going to be present in the same image frame). However, training the novelty filters with the orange ball present in the arena allows the assessment of performance degradation during inspection. In other words, we wanted to investigate if the orange ball was ignored during the inspection phase, while the grey box continued to be highlighted.

- Second, the grey box was placed next to the ball, in the same corner of the arena. This obviously affects the response of the attention mechanism because the two objects of interest are present at the same time in some of the image frames, competing for saliency. Once more, we wanted to evaluate if the grey box was correctly identified as new even in the presence of the more conspicuous but already known orange football, which should be ignored by the novelty filters.

Both GWR network and incremental PCA algorithm were able to detect the *novel* object successfully in each situation, regardless of the presence or not of the distractor (the orange ball) in the same image frame. This is illustrated in Figure 5.10, which shows an example of output image for each of the novelty filters used.




(a) (b)

Figure 5.10: Output images for Experiment 12: (a) GWR network; (b) incremental PCA. The grey box (novel stimulus) is correctly highlighted as novel regardless of the presence of the orange ball (normal stimulus).

From Figure 5.10 it can be noticed that the GWR network has correctly identified region 0 as novel. Incremental PCA highlighted all of the regions containing the grey box, although region 1 also contained a large portion of the orange ball. In this experiment, the GWR network has acquired 18 nodes while incremental PCA has acquired 45 vectors with 32 dimensions.

Results of the $\chi^2$ analysis revealed statistical significance ($p \leq 0.01$) in the associations between novelty filter response and manually generated ground truth. A quantitative comparison in terms of Cramer's $V$, uncertainty coefficient $U$ and $\kappa$ index of agreement is given in Table 5.2.

Table 5.2: Performance comparison using raw image patches — Experiment 12 (2250 samples). All results correspond to statistically significant correlation between system response and actual novelty status ($\chi^2$ analysis, $p \leq 0.01$).

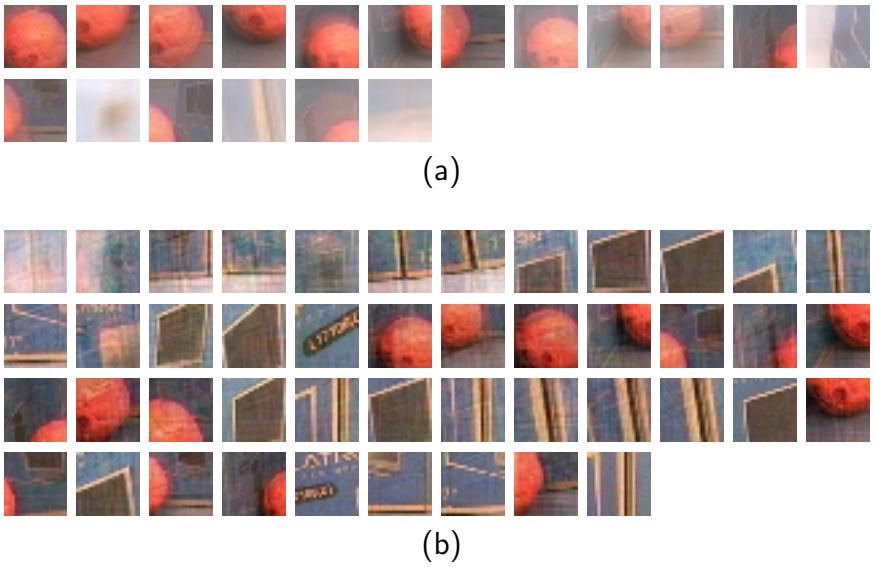
| | GWR Network | Incremental PCA |
|---|---|---|
| Grey box (alone) | $V = 0.73$<br>$U = 0.46$<br>$\kappa = 0.73$ | $V = 0.79$<br>$U = 0.52$<br>$\kappa = 0.78$ |
| Grey box (with ball) | $V = 0.37$<br>$U = 0.10$<br>$\kappa = 0.33$ | $V = 0.54$<br>$U = 0.23$<br>$\kappa = 0.53$ |

It can be noticed that the performance of both systems in the case when the grey box appears alone — *i.e.* in a different location than the ball — was not affected very much by the inclusion of the orange ball during the exploration phase (compare "grey box (alone)" in Table 5.2 with "grey box" in Table 5.1). However, performance of both systems deteriorated noticeably when the orange ball was present in the same frame as the grey box.

The fact that performance remained virtually the same when the grey box appeared alone (*i.e.* in similar conditions as in Experiment 11) indicates that having the orange football present in the arena during training did not impair the ability of both novelty filters to discriminate between classes correctly. Therefore, we concluded that the combination of attention mechanism and image encoding is the one to be blamed for the reduction in performance when the orange ball was present at the same location as the grey box. Causes of the poor performance obtained when both objects of interest were in the same image frame are possibly related to accuracy and stability in the location of salient points, which will be investigated in Chapter 6.

The reconstructed image patches for both GWR network and incremental PCA are shown in Figure 5.11, where the averaging effect discussed in Experiment 11 can be noticed again in some of the image patches acquired by the GWR network (see, for example, the last image patch in Fig-

ure 5.11a). Concerning the incremental PCA approach, the reconstructed image patches show some deterioration of the PCA space (compare the quality of reconstructed patches in Figures 5.11b and 5.9b on page 126) because of the inclusion of the orange ball during training. However, the incremental PCA model shows very good consistency (all image patches in Figure 5.9b are still present plus some more corresponding to the ball), contrasting to the results obtained for the GWR network.



(a)

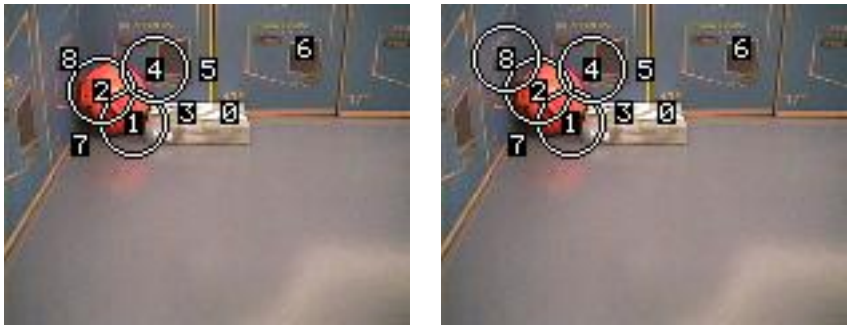
(b)

Figure 5.11: Image patches acquired during Experiment 12: (a) reconstructed from the GWR network; (b) reconstructed from incremental PCA.

## 5.3 Experiment 13: Saliency versus Novelty Revisited

In order to obtain confirmation of the results obtained previously, we inverted the roles of the objects of interest in Experiment 12. A new experiment, in which the robot explored the arena containing the inconspicuous grey box and afterwards inspected it with the inclusion of the conspicuous orange ball, was conducted. In this new context, the grey box became the known object and the orange football became the novel object. Again, two different situations were analysed, when the ball and the box were at the same corner and also when they were in different corners of the arena.

Both the GWR network and the incremental PCA algorithm were able to detect the orange ball successfully as the novel object in each situation, regardless of the presence or not of the known grey box in the same image frame. Figure 5.12 shows an example of output image for each of the novelty filters used, where one can notice that both novelty filters correctly identified the orange ball as being novel and ignored the grey box. The GWR network acquired 11 nodes in this experiment and the incremental PCA acquired 47 vectors with 29 dimensions.




(a) (b)

Figure 5.12: Output images for Experiment 13: (a) GWR network; (b) incremental PCA. The orange ball (novel stimulus) is correctly highlighted as novel regardless of the presence of the grey box (normal stimulus).

The $\chi^2$ analysis once more revealed statistical significance ($p \leq 0.01$) in the associations between novelty filter response and manually generated ground truth. Table 5.3 shows the quantitative performance comparison using Cramer's $V$, uncertainty coefficient $U$ and $\kappa$ index of agreement.

Table 5.3: Performance comparison using raw image patches — Experiment 13 (2250 samples). All results correspond to statistically significant correlation between system response and actual novelty status ($\chi^2$ analysis, $p \leq 0.01$).

| | GWR Network | Incremental PCA |
|---|---|---|
| Orange ball (alone) | $V = 0.81$<br>$U = 0.55$<br>$\kappa = 0.81$ | $V = 0.84$<br>$U = 0.65$<br>$\kappa = 0.83$ |
| Orange ball (with box) | $V = 0.88$<br>$U = 0.68$<br>$\kappa = 0.88$ | $V = 0.79$<br>$U = 0.56$<br>$\kappa = 0.78$ |

The performance of the incremental PCA approach in the case when the orange football appears alone was not affected by the inclusion of the grey box during the exploration phase (compare "orange ball (alone)" in Table 5.3 with "orange ball" in Table 5.1 on page 126). However, performance of the GWR network deteriorated noticeably in this case. Surprisingly, in the case where the ball was present in the same corner as the box performance was better for the GWR network and worse for the incremental PCA. A possible explanation for the deterioration of incremental PCA performance is the fact that in this experiment the dimensionality of the PCA space was reduced to 29 eigenvectors.

Figure 5.13 shows the reconstructed image patches for the GWR network and the incremental PCA approach, which now also include fragments of the grey box that was present during training, clearly showing that both systems acquired knowledge about the grey box's appearance.

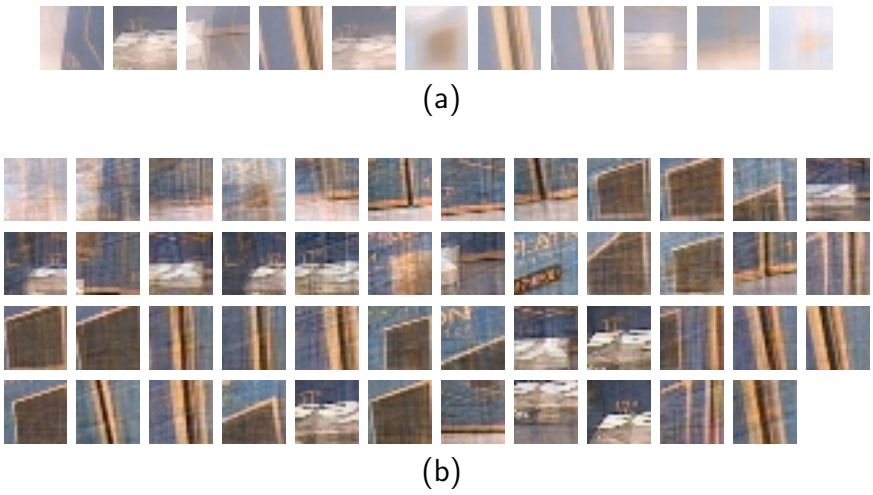



Figure 5.13: Image patches acquired during Experiment 13: (a) reconstructed from the GWR network; (b) reconstructed from incremental PCA.

As in Experiment 12, reconstruction also shows some deterioration of the PCA space (compare the quality of reconstructed patches in Figures 5.13b and 5.9b on page 126) because of the increase in environmental complexity due to the inclusion of extra visual features from the grey box.

## 5.4 Experiment 14: Novelty Detection in a Real World Environment

After the successful results obtained in experiments conducted in laboratory environments, it was time to test the proposed visual novelty detection approach in a medium-scale real world environment.

The ideal scenario would be to send the robot down a sewer pipe to inspect for cracks, tree roots and other types of faults. However, our research robot is too large and also not fit to operate in such an environment. Furthermore, rigorous analysis and assessment of the system's behaviour in this type of situation would be very difficult to perform due to the lack of knowledge and control of environmental characteristics — construction of the novelty ground truth, for instance, would be a hard task.

**Experimental setup.** Hence, we decided to conduct experiments in one of the corridors at the Network Centre building at Essex. The robot navigated along the corridor using the same navigation behaviour previously used in the laboratory experiments, acquiring one image frame per second, which resulted in the acquisition of 50 images per journey along the corridor. Differently from previous laboratory experiments, the camera's pan-tilt unit was driven to its home position (facing straight towards the forward direction of the robot) for the experiments in the corridor.

Exploration was performed in the "empty" corridor to acquire a model of normality, as in previous laboratory experiments, but limited to three journeys along the corridor. Finally, the learnt model of normality was used to inspect the corridor for unusual visual features that were manually inserted *a posteriori*.

We placed three different novel objects in the corridor at different times: a black rubbish bag, a dark brown bin and a yellow wooden board. These objects appeared in the robot's field of view immediately after the traversal

of a door, which was present in the corridor. Figure 5.14 shows examples of images in which these objects appeared, along with their novelty ground truth images.

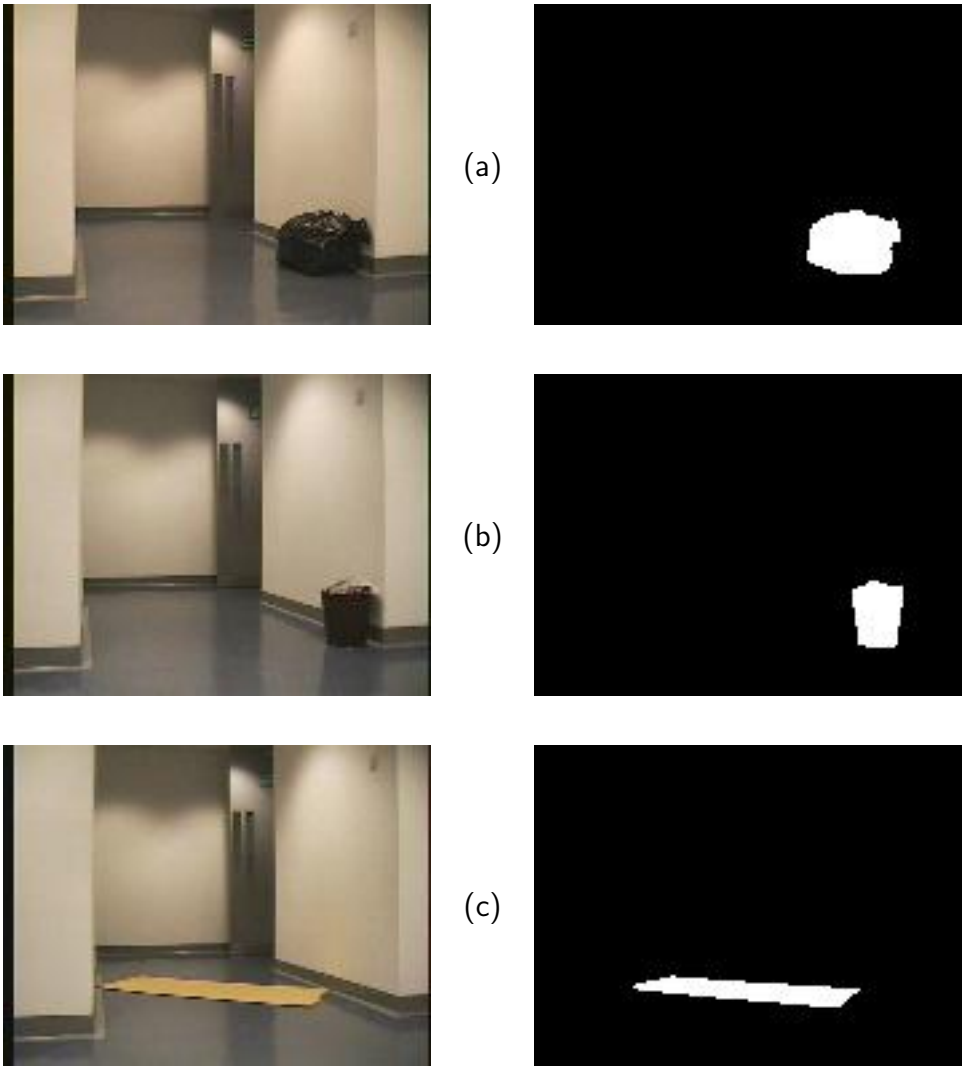



Figure 5.14: Input images and their novelty ground truth: (a) black rubbish bag; (b) dark brown bin; (c) yellow wooden board.

**Results.** After three exploration journeys along the empty corridor, the GWR network acquired 48 nodes, while the incremental PCA acquired 80 vectors with 19 dimensions. Apart from the wooden board, the chosen novel objects are dark and therefore to some extent similar to the dark areas of the normal environment.

Contrast in the images acquired in the corridor was generally poor because no extra illumination was used, just the weak lighting already present.

In spite of this fact, both GWR network and incremental PCA algorithm were able to correctly highlight the novel objects in the corridor during inspection, as shown in Figure 5.15.

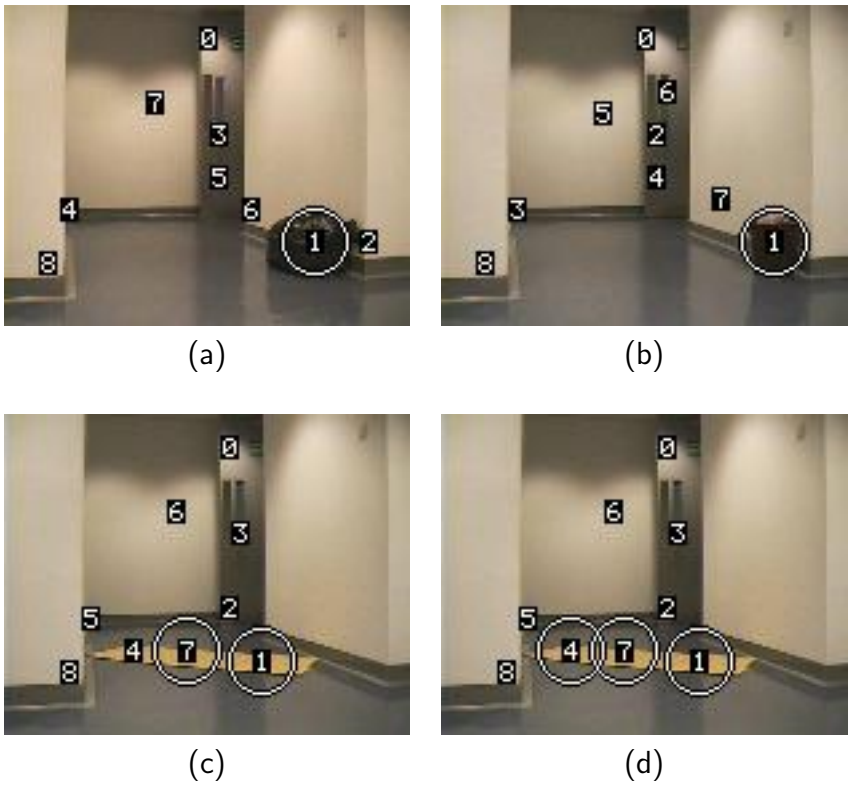
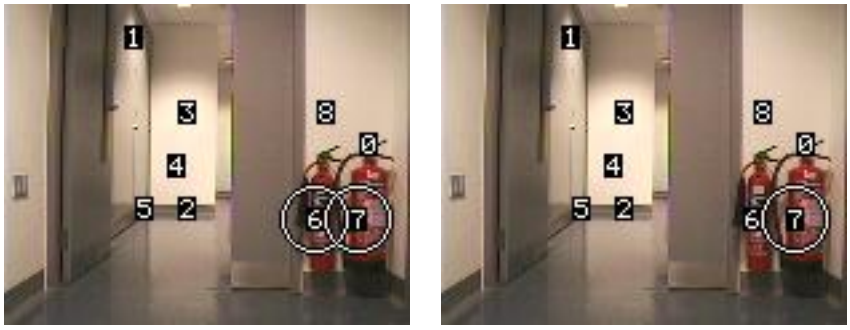



Figure 5.15: Output images for Experiment 14: (a) black rubbish bag (GWR and PCA); (b) dark brown bin (GWR and PCA); (c) yellow wooden board (GWR only); (d) yellow wooden board (PCA only). The white circles correctly indicate regions containing novel features.

However, both novelty filters also responded with false novelty indications for a pair of fire extinguishers that were present in the corridor. These novelty indications were unexpected because the fire extinguishers were present during the exploration phase and therefore should be part of the acquired model of normality. An example of such false responses is given in Figure 5.16.

Initially we attributed these false novelties to the relatively small number of exploration journeys, which may not have been sufficient for the acquisition of a good model of normality. Indeed, closer examination of the nodes acquired by the GWR network revealed that 21 out of 48 nodes were




(a) (b)

Figure 5.16: False novelty responses for the fire extinguishers: (a) GWR network; (b) incremental PCA. The white circles indicate regions erroneously labelled as novel.

not completely habituated. We therefore forced these nodes to be habituated completely by manually altering their synaptic efficacies and repeated inspection of the corridor using the resulting GWR network.

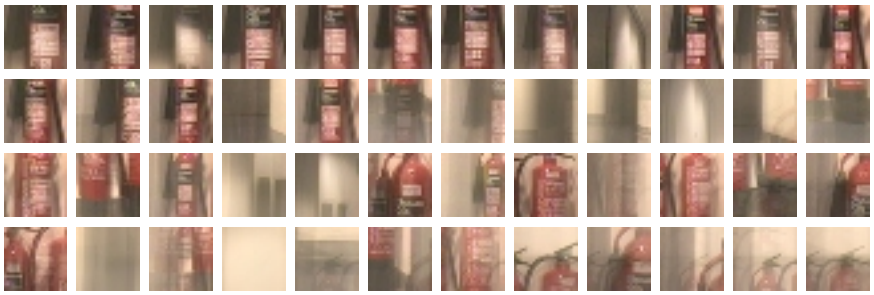
However, forcing habituation had only a minor impact on the number of false positives. We hypothesise that improvements in accuracy of the salient locations and the use of an image encoding method robust to changes in scale would contribute to reduce false novelty indications and enhance general performance of the visual novelty filter. Experiments concerning the improvement of accuracy in the location of interest points and automatic scale selection are reported in Chapter 6.

Table 5.4 shows a performance comparison in terms of Cramer's $V$, uncertainty coefficient $U$ and $\kappa$ index of agreement. All results showed statistically significant correlation between system response and actual novelty status ($\chi^2$ analysis, $p \leq 0.01$). Overall performance (combined results for all three novel objects) indicates strong agreement between system response and actual novelty status. The GWR network presented the most consistent results.

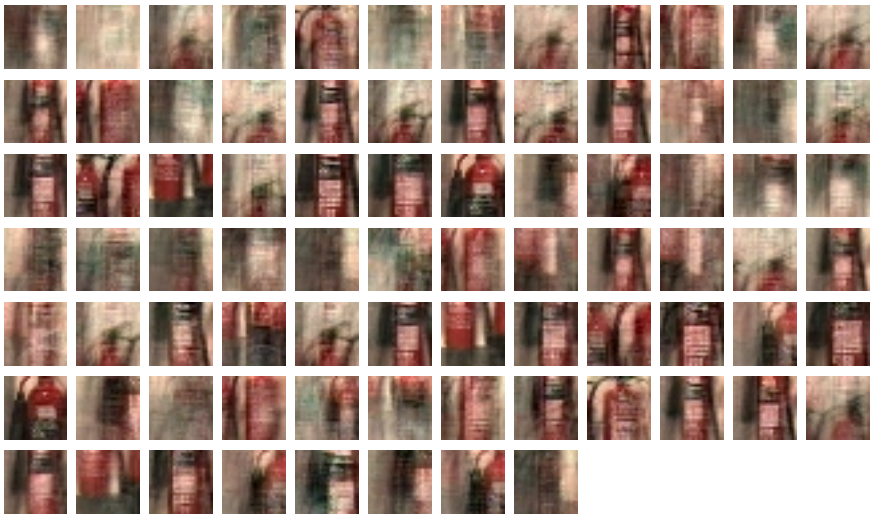
Figure 5.17 shows the reconstructed image patches acquired by both approaches during the exploration phase of this experiment, where one can notice that the fire extinguishers are widely represented in the models acquired by both approaches.

Table 5.4: Performance comparison using raw image patches — Experiment 14 (2250 samples). All results correspond to statistically significant correlation between system response and actual novelty status ($\chi^2$ analysis, $p \leq 0.01$).

| | GWR Network | Incremental PCA |
|---|---|---|
| Bag | $V = 0.63$<br>$U = 0.35$<br>$\kappa = 0.63$ | $V = 0.65$<br>$U = 0.37$<br>$\kappa = 0.65$ |
| Bin | $V = 0.64$<br>$U = 0.38$<br>$\kappa = 0.64$ | $V = 0.50$<br>$U = 0.23$<br>$\kappa = 0.50$ |
| Wood | $V = 0.67$<br>$U = 0.37$<br>$\kappa = 0.67$ | $V = 0.84$<br>$U = 0.69$<br>$\kappa = 0.84$ |
| Overall | $V = 0.65$<br>$U = 0.36$<br>$\kappa = 0.65$ | $V = 0.70$<br>$U = 0.44$<br>$\kappa = 0.70$ |



(a)



(b)

Figure 5.17: Image patches acquired during Experiment 14: (a) reconstructed from the GWR network; (b) reconstructed from incremental PCA.

Nevertheless, neither the GWR network nor the incremental PCA algorithm were able to avoid false novelties due to the fire extinguishers, indicating poor generalisation by both learning mechanisms. In fact, the large amount of image patches corresponding to similar features of the fire extinguishers is already a consequence of poor generalisation by the clustering mechanisms.

### 5.4.1 Application Illustration: Air Duct Inspection

We were also interested in testing the performance of our framework in real applications, such as sewer pipe or air conditioning duct inspection. A Danish duct cleaning company has kindly provided us with some images acquired from real air conditioning ducts. Our interest in this case was to test if our visual novelty detection framework would be able to highlight dirty sections of the duct while inspecting it. The GWR network was trained with images acquired from clean ducts using a remotely controlled robot, driven by a human operator. After training, images acquired from dirty ducts were presented to the novelty filter. Some examples of the results obtained can be seen in Figure 5.18.

Although the experiment using air duct data can be considered successful because the system was able to highlight dirty sections of the duct, it is hard to be assessed. First, as it can be noticed from Figure 5.18, the images from the clean duct (a) are rather different than the images from dirty ducts (b, c and d). Furthermore, the lack of a controlled experimental setup makes it difficult to establish ground truth data in order to evaluate performance. Nevertheless, this is an experiment that illustrates a potential real world application.

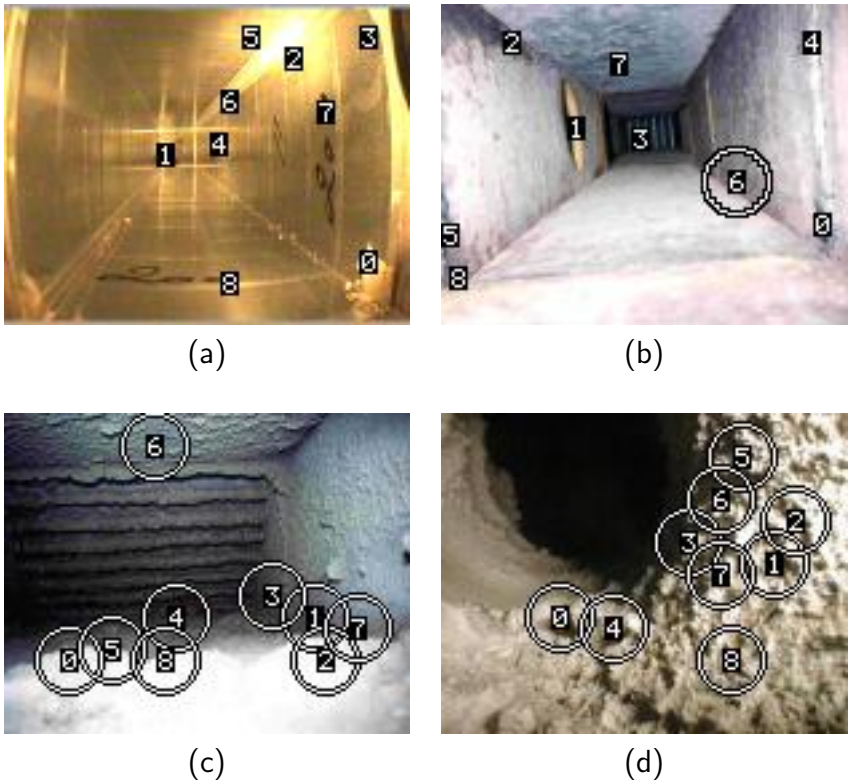


Figure 5.18: Output images for the air duct experiment: (a) clean air duct; (b) dirty air duct 1; (c) dirty air duct 2; (d) dirty air duct 3.

## 5.5 Experiment 15: Influence of the Navigation Trajectory

After the successful results obtained by the previous experiments, a question arose concerning the robustness of our approach to variations in the robot's trajectory when exploring and inspecting the environment. As one of the potential applications for the system is the inspection of ducts and sewers, possibly using remote controlled robots — and therefore subject to larger variations in trajectory than automated navigation behaviours — another set of laboratory experiments was designed to address this question.

The experiments were conducted in the laboratory mainly because the use of an overhead camera was necessary to track the robot's trajectory while exploring and inspecting the environment. Again, a square arena was built with cardboard boxes and used as the robot's operating environment. A major difference to the environments used in previous experiments is that

the floor of the arena in this case was metallic and hence very reflective, which poses extra difficulties to robot vision algorithms.

Instead of changing the robot's behaviour, we decided to slightly change the environment in order to alter the robot's trajectory in the arena. This was possible thanks to the force-field obstacle avoidance behaviour that was used in all previous experiments. Changes in the robot's path were done by removing the central cardboard boxes in the arena which previously forced the robot to follow a uniform path around the arena. It is important to mention that the central obstacles were not sensed by the robot's camera (only by the laser range sensor which controlled navigation) and therefore the act of adding or removing them to the environment did not affect the robot's visual world (in the sense of more objects being visually detected).

Figure 5.19 illustrates the effect of the presence or not of the central boxes in the final trajectory followed by the robot. The paths shown were plotted from data logged using the overhead camera at the robotics research laboratory at Essex.

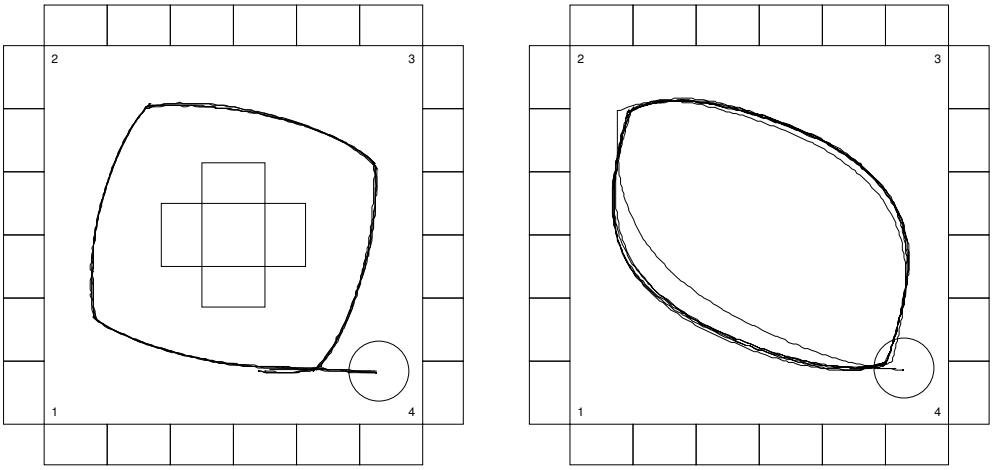



Figure 5.19: Trajectories around the arena: (a) "square" trajectory that results from the presence of obstacles in the centre of the arena; (b) "oval" trajectory that results from the absence of obstacles in the centre of the arena. When central obstacles were present, the robot was repositioned at the starting point for every loop in a total of five loops in order to obtain consistent novelty graphs for qualitative assessment. When central obstacles were absent, the five loops were logged continuously without stopping or repositioning the robot.

In Figure 5.19a it can be noticed that the robot approaches the four corners in a similar fashion when the central boxes are present ("square" trajectory), and in Figure 5.19b the robot gets closer to corners 2 and 4 while keeping distance from corners 1 and 3 ("oval" trajectory). These trajectories are noticeably different and hence result in different visual experience of the environment. Not having the obstacles in the centre allows a wider range of scale transformations due to the larger variation of the distance from the walls, for instance.

We started the trajectory experiments by exploring the empty arena using the "square" trajectory. After five loops of exploration, the GWR network acquired 15 nodes, which were used to inspect the arena and highlight any novel object. Two new objects were inserted in the arena for the inspection phase: the orange football already used in previous experiments, and a yellow cone. However, differently from previous experiments, the central obstacles were removed from the arena during the inspection phase, resulting in an "oval" inspection trajectory.

Figures 5.20 and 5.21 show that despite the differences in trajectory, the GWR network was able to identify the orange football correctly as the new object in the arena. The GWR network also produced some false novelty indications, which were due to images acquired from points of view previously unexplored during training. The "oval" trajectory offers the robot more variety of affine transformations (changes in perspective) because it is not as regular as the "square" trajectory.

The output image depicted in Figure 5.21 is particularly interesting because it shows that the GWR network correctly highlighted the orange ball as the novel object and also its reflection on the shiny metallic floor. Although the reflection is not considered as novel in the ground truth image, it certainly can be considered as a novel visual feature in the environment and was correctly identified by the novelty filter. However, for the purposes

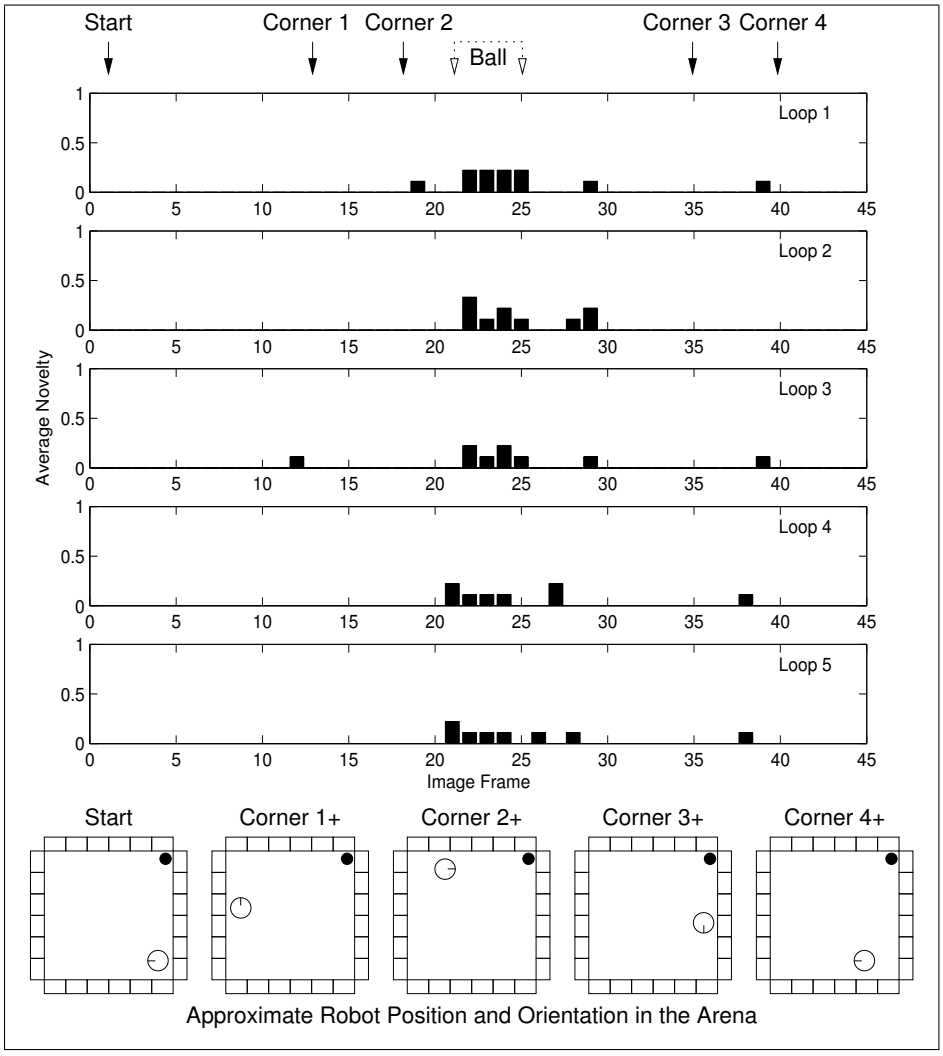



Figure 5.20: Inspection of the arena with the orange football (novel stimulus) using the GWR network as novelty filter. Exploration was done using the "square" trajectory while inspection was done using the "oval" trajectory. The novel stimulus is consistently identified with few false novelty indications.


(a) (b)

Figure 5.21: Output and ground truth images for the "oval" trajectory inspection when the orange ball was introduced in the arena. The ball (region 0) was correctly identified as being the novel stimulus. Also, the ball's reflection on the shiny metallic floor (region 5) was highlighted as novel.

of quantitative performance assessment of these experiments (computation of $V$, $U$ and $\kappa$), we did not consider any reflections of novel objects on the shiny floor as being novel.

A second inspection round was conducted using a different new object, the yellow cone. Figures 5.22 and 5.23 show that the GWR network was able to identify the new object correctly, although some false novelties occurred.

One can notice in Figure 5.23 that the reflection of the yellow cone on the floor (region 8) was considered novel by the novelty filter, similarly to what happened in the case of the orange ball.

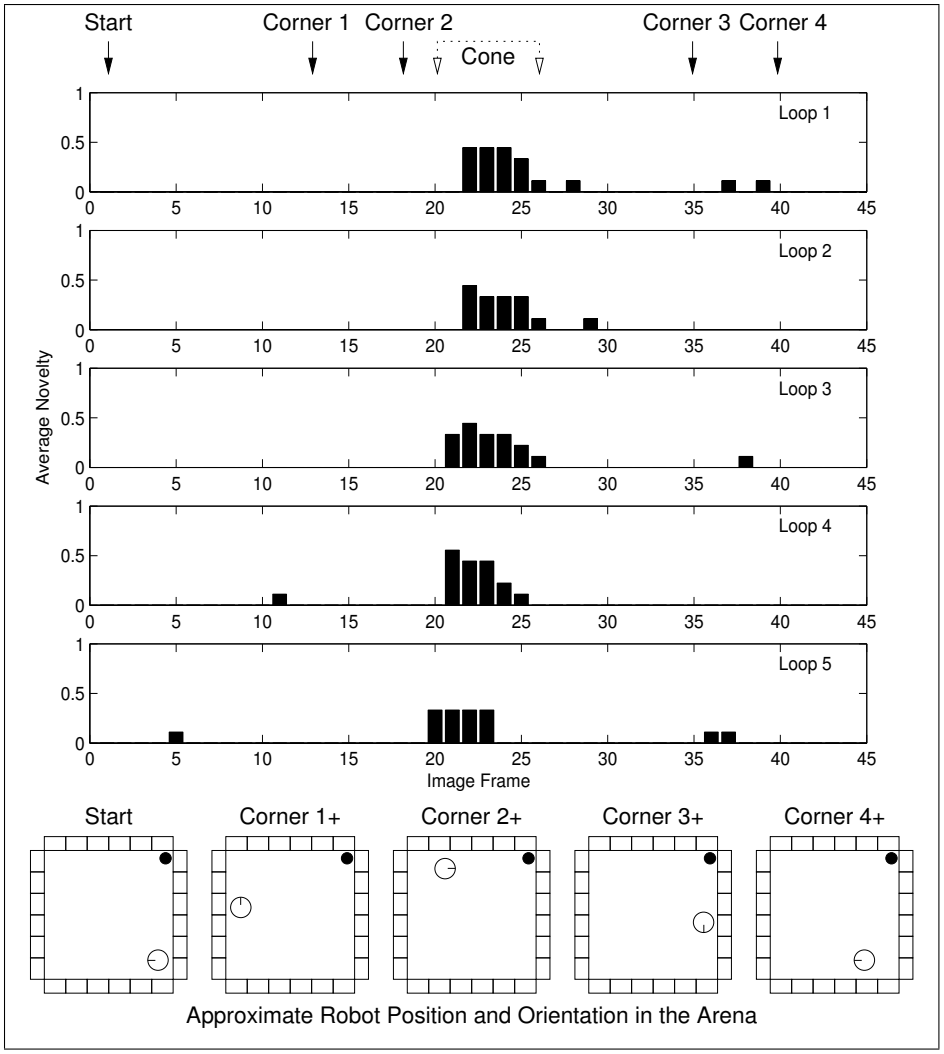



Figure 5.22: Inspection of the arena with the yellow cone (novel stimulus) using the GWR network as novelty filter. Exploration was done using the "square" trajectory while inspection was done using the "oval" trajectory. The novel stimulus is correctly and consistently identified.


(a) (b)

Figure 5.23: Output and ground truth images for the "oval" trajectory inspection when the yellow cone was introduced in the arena. The cone (regions 0, 2 and 7) was correctly identified as being the novel stimulus. Also, the cone's reflection on the shiny metallic floor (region 8) was highlighted as novel.

It can be argued that the "oval" trajectory provides more varied samples of the environment — at different perspectives and affine transformations — than the "square" trajectory, because the former is not as regular as the latter. Therefore, it seems natural to expect that exploring the arena using the "oval" trajectory would result in a more general model of normality, reducing the amount of false novelties when using a different, more regular, inspection trajectory.

Therefore, a new exploration phase was conducted in the empty arena using the "oval" trajectory in order to test this hypothesis. Surprisingly, the GWR network acquired only 8 nodes, in contrast to the 15 nodes acquired previously using the "square" trajectory. Hence, the acquired model using the "oval" trajectory can be considered less varied than the one acquired using the "square" trajectory, contradicting our expectations.

As before, the acquired model was used to filter out abnormal visual features during the inspection phase of the experiment. Figure 5.24 shows the qualitative results obtained during inspection of the arena with the orange football using the "square" trajectory.

It can be noticed in Figure 5.24 that the ball was correctly highlighted as being a new object, but also there were unexpected consistent indications of a second new entity in the arena. A closer examination of the corresponding

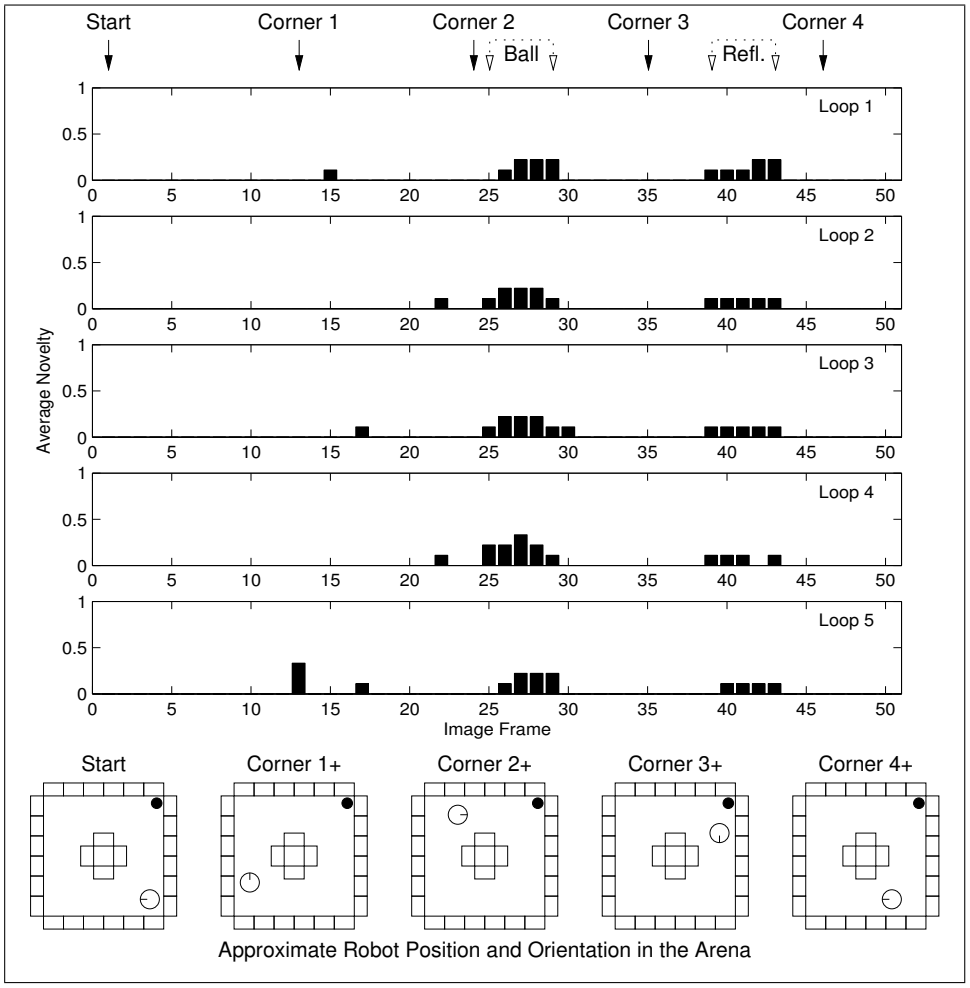
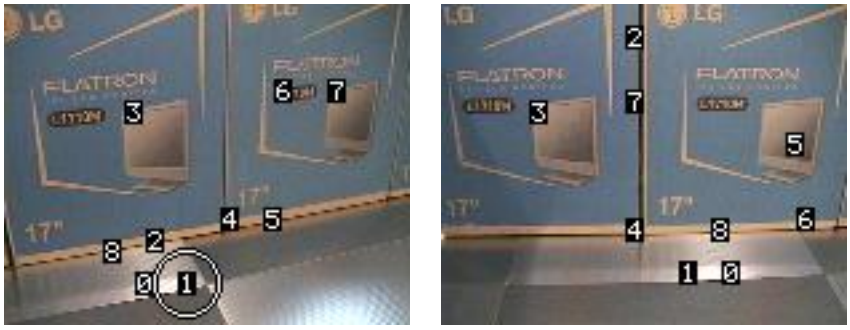



Figure 5.24: Inspection of the arena with the orange football (novel stimulus) using the GWR network as novelty filter. Exploration was done using the "oval" trajectory while inspection was done using the "square" trajectory. The ball is consistently identified, but also a bright spot due to a reflection on the metallic floor.

image frames revealed that this unexpected new entity was caused by a bright reflection on the shiny floor (see Figure 5.25a).

The explanation is that this bright reflection was never seen — at least not with this appearance — during the exploration phase. The output image obtained during the exploration phase corresponding to the same environment location is shown in Figure 5.25b, where one can notice the displacement of the reflection — and therefore the displacement of the corresponding salient points — due to changes in perspective that resulted from the different navigation trajectories.

A new inspection round was conducted in the arena containing the yel-




(a) (b)

Figure 5.25: Output images for the bright reflection on the floor: (a) during inspection; (b) during exploration. The resulting image patches in the vicinity of regions 1 in (a) and 0 in (b) are noticeably different.

low cone and the corresponding qualitative results obtained are depicted in Figure 5.26. Once more, the trained GWR network was able to differentiate the novel object, but also consistently indicated the same bright reflection described above as a novel entity.

The quantitative performance assessment according to the trajectories used for exploration and inspection is presented in Table 5.5. Results obtained for the cases in which both exploration and inspection were conducted using the same trajectory are also presented for comparison. All cases resulted in statistically significant association between system response and actual novelty status ($\chi^2$ analysis, $p \leq 0.01$). The $\chi^2$ analysis was computed from 2250 samples (5 loops $\times$ 50 images $\times$ 9 salient regions) in the case of inspection with the "square" trajectory and 1980 samples (5 loops $\times$ 44 images $\times$ 9 salient regions) in the case of inspection with the "oval" trajectory.

Table 5.5 shows that differences in the trajectory used for exploration and inspection have a negative impact in the overall performance of the proposed visual novelty detection framework — the best performances are obtained when the same navigation trajectory is employed for both exploration and inspection phases. Nevertheless, the results obtained with different navigation trajectories are still statistically significant and detect novel objects correctly.

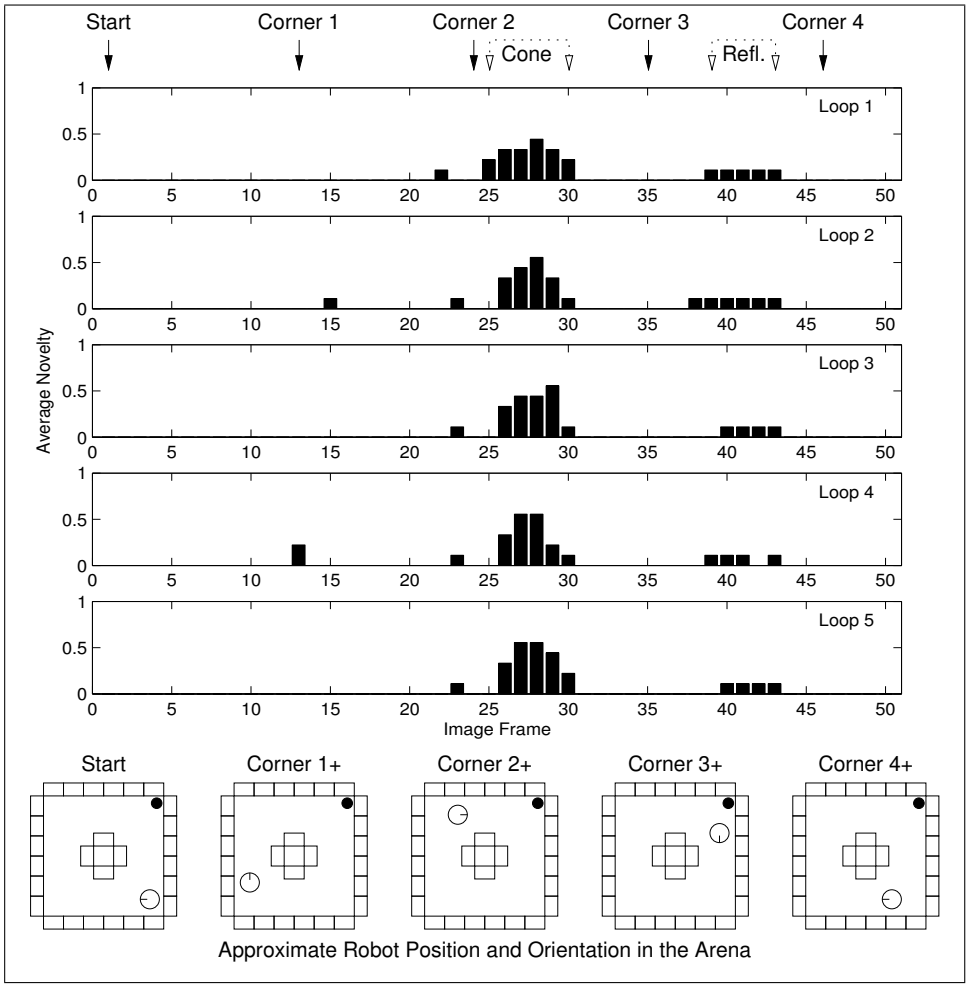



Figure 5.26: Inspection of the arena with the yellow cone (novel stimulus) using the GWR network as novelty filter. Exploration was done using the "oval" trajectory while inspection was done using the "square" trajectory. The cone is consistently identified, but also the bright spot due to a reflection on the metallic floor.

As an extra baseline for comparisons, we also conducted the same trajectory experiments using the incremental PCA algorithm, whose results are given in Table 5.6.

The incremental PCA approach acquired 35 vectors with 28 dimensions when the exploration trajectory was "square" and 36 vectors with 30 dimensions when the trajectory was "oval", which shows no big difference in the generalisation ability of the models due to changes in exploration trajectory.

Table 5.5: Performance comparison between different exploration and inspection trajectories using the GWR network. All results correspond to statistically significant correlation between system response and actual novelty status ($\chi^2$ analysis, $p \leq 0.01$).

| | Square vs. Oval | Oval vs. Square | Square vs. Square | Oval vs. Oval |
|---|---|---|---|---|
| Orange ball | $V = 0.74$<br>$U = 0.66$<br>$\kappa = 0.71$ | $V = 0.69$<br>$U = 0.61$<br>$\kappa = 0.64$ | $V = 0.80$<br>$U = 0.66$<br>$\kappa = 0.78$ | $V = 0.83$<br>$U = 0.74$<br>$\kappa = 0.82$ |
| Yellow cone | $V = 0.83$<br>$U = 0.70$<br>$\kappa = 0.81$ | $V = 0.82$<br>$U = 0.70$<br>$\kappa = 0.80$ | $V = 0.91$<br>$U = 0.81$<br>$\kappa = 0.91$ | $V = 0.82$<br>$U = 0.68$<br>$\kappa = 0.81$ |
| Overall | $V = 0.80$<br>$U = 0.68$<br>$\kappa = 0.78$ | $V = 0.77$<br>$U = 0.66$<br>$\kappa = 0.74$ | $V = 0.87$<br>$U = 0.75$<br>$\kappa = 0.87$ | $V = 0.83$<br>$U = 0.71$<br>$\kappa = 0.81$ |

Table 5.6: Performance comparison between different exploration and inspection trajectories using incremental PCA. All results correspond to statistically significant correlation between system response and actual novelty status ($\chi^2$ analysis, $p \leq 0.01$).

| | Square vs. Oval | Oval vs. Square | Square vs. Square | Oval vs. Oval |
|---|---|---|---|---|
| Orange ball | $V = 0.79$<br>$U = 0.70$<br>$\kappa = 0.76$ | $V = 0.69$<br>$U = 0.61$<br>$\kappa = 0.65$ | $V = 0.83$<br>$U = 0.69$<br>$\kappa = 0.82$ | $V = 0.86$<br>$U = 0.77$<br>$\kappa = 0.85$ |
| Yellow cone | $V = 0.82$<br>$U = 0.68$<br>$\kappa = 0.80$ | $V = 0.86$<br>$U = 0.75$<br>$\kappa = 0.86$ | $V = 0.97$<br>$U = 0.90$<br>$\kappa = 0.97$ | $V = 0.84$<br>$U = 0.71$<br>$\kappa = 0.82$ |
| Overall | $V = 0.81$<br>$U = 0.69$<br>$\kappa = 0.79$ | $V = 0.80$<br>$U = 0.68$<br>$\kappa = 0.78$ | $V = 0.92$<br>$U = 0.80$<br>$\kappa = 0.91$ | $V = 0.85$<br>$U = 0.73$<br>$\kappa = 0.84$ |

## 5.6 Summary and Discussion

All experiments reported in this chapter produced statistically significant correlations between system response and actual ground truth. In other words, all of them worked as desired — Tables 5.1, 5.2, 5.3, 5.4, 5.5 and 5.6 illustrate the strength of this correlation.

Although the strength of the association measured by Cramer's $V$ and the uncertainty coefficient $U$ are not all that close to the ideal value of one in some situations, these results have to be understood as "very conservative"

for two reasons. First, if the consistency of novelties detected between successive image frames is taken into account, it is possible to rule out most false positives (novelty detected but not present). And second, most false negatives (novelty present but not detected) can be eliminated using the fact that a single image patch within the new object labelled as novel is enough to characterise the entire object as novel. Nevertheless, the values of $V$ and $U$ serve well for the purpose of comparing performances of different methods.

**Comparison between GWR and PCA novelty filters.** Results given by the GWR network and the incremental PCA approach are similar in performance, although the size of the models acquired by each are very different for the set of parameters used. The smaller amount of vectors learnt by the GWR had always the original input dimensionality (1728 elements), while the dimensionality of the vectors acquired by incremental PCA varied from 19 to 33 elements. However, every dimension of the projected vectors acquired by the incremental PCA approach corresponds to an eigenvector with 1728 elements, resulting in the allocation of more memory. Also, on average, the GWR-based novelty filter performed twice as fast as the incremental PCA algorithm. Throughout all experiments, the incremental PCA algorithm proved to be more expensive in terms of memory and computing power.

Dimensionality issues become important when we consider that the Euclidean metric was used to determine similarity between vectors. When Euclidean distance is used, a small difference between two high-dimensional vectors tend to be large in value, making it difficult to establish thresholds of similarity for high-dimensional spaces, as it is the case with the vectors acquired by the GWR network.

Despite its clear weaknesses concerning computational cost, the PCA algorithm offers some advantages over the GWR mechanism. Initially, in

the incremental PCA approach similarity between inputs is performed by the residual error in reconstruction from the projected space. Moreover, substitution of the Euclidean distance by the Mahalanobis distance can be easily implemented in the incremental PCA approach once the covariance matrix of the stored projected vectors is available as a sub-product of the method. The Mahalanobis distance normalises the contribution of vector elements according to the covariance matrix of the data:

$$d_{\mathbf{xy}} = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathsf{T}}\mathbf{C}^{-1}(\mathbf{x}-\mathbf{y})}, \tag{5.1}$$

where $d_{\mathbf{xy}}$ is the Mahalanobis distance between the column vectors $\mathbf{x}$ and $\mathbf{y}$ and $\mathbf{C}$ is the covariance matrix of the data. Euclidean distance corresponds to the special case where $\mathbf{C}$ is the identity matrix.

Another advantage of the incremental PCA approach is the ability to reduce dimensionality automatically, allowing optimal reconstruction of the original input image patch (principal component projection minimises the squared reconstruction error). By using raw image patches the user can evaluate exactly which parts of the environment were actually learnt by the system through image patch reconstruction. It was possible to notice that reconstruction of the GWR network nodes resulted in averaged image patches because of the network learning procedure.

We made further experiments with the GWR network by increasing the activation threshold $a_T$ in order to acquire a number of nodes as close as possible to the number of vectors acquired by the incremental PCA. This resulted in better reconstruction from the network nodes, but also decreased the overall novelty detection performance of the GWR-based system noticeably. As one would expect, the number of false negatives decreased in this case, but on the other hand the number of false positives increased immensely. We attribute this effect to the use of the Euclidean distance in a high-dimensional space.

The difficulty in evaluating similarity between inputs in high dimensions using Euclidean distance normally forces the system designer to use an additional preprocessing stage, such as the use of colour statistics (as in the experiments reported in Chapter 4), for dimensionality reduction when using the GWR network. On the other hand, the GWR approach offers the functionality of constructing a topological relationship between inputs.

Future investigations should therefore aim at combining the embedded dimensionality reduction feature of the incremental PCA with the topological construction algorithm of the GWR network using the Mahalanobis distance as a measure of similarity between patterns. Extensions to the incremental PCA algorithm that make it robust to partial occlusions in the image patches (Skočaj and Leonardis, 2003) are also attractive for future investigations.

Considering the general system functionality, the attention mechanism plays an important role in the system's ability to generalise by providing image patches that are robust to translations and therefore reducing the number of acquired nodes or vectors. Alternatives to the saliency map as attention model, which offer invariance to scale (Mikolajczyk and Schmid, 2001) and affine transformations (Mikolajczyk and Schmid, 2002) may improve the generalisation ability of the system, helping to reduce the number of vectors or nodes in the models of normality. Experiments concerning invariance to scale are reported in the next chapter.

# Chapter 6

# Visual Attention and Automatic Scale Selection

The attention mechanism plays an important role in the general performance of the proposed framework for visual novelty detection. Its use not only allows the localisation of novel visual features within an input image frame, but also contributes to the reduction in the dimensionality of the feature vectors that are fed to the novelty filter stage. In cases where relative positioning of pixels is important to the image encoding scheme, the accuracy of the interest points located by the attention mechanism is crucial to avoid misalignment errors. Also, image encoding procedures which are robust to changes in scale and affine transformations are desirable in order to improve generalisation and reduce the number of acquired concepts (model nodes or vectors) by the learning mechanism.

In Chapters 4 and 5 we reported experiments using the saliency map (Itti et al., 1998) as the mechanism of visual attention. This biologically-inspired model uses multi-scale image representations (image pyramids) to detect variations in intensity, colour and edge orientation. Conspicuity maps of each feature type are normalised and combined into a single saliency map (see Subsection 2.2.1 for details), whose values are a measure of the degree

of saliency of the corresponding image region. Normalisation is necessary in order to combine intensity, colour and orientation conspicuity maps with different dynamic ranges into a single saliency map and, as a result, gives more weight to unusual features in the input image frame.

Figure 6.1 shows examples of typical images acquired from a robot arena and their respective pseudo-coloured saliency maps (low saliency values appear in blue and high saliency values appear in red), each of them with predominant features that determined the highest saliencies.

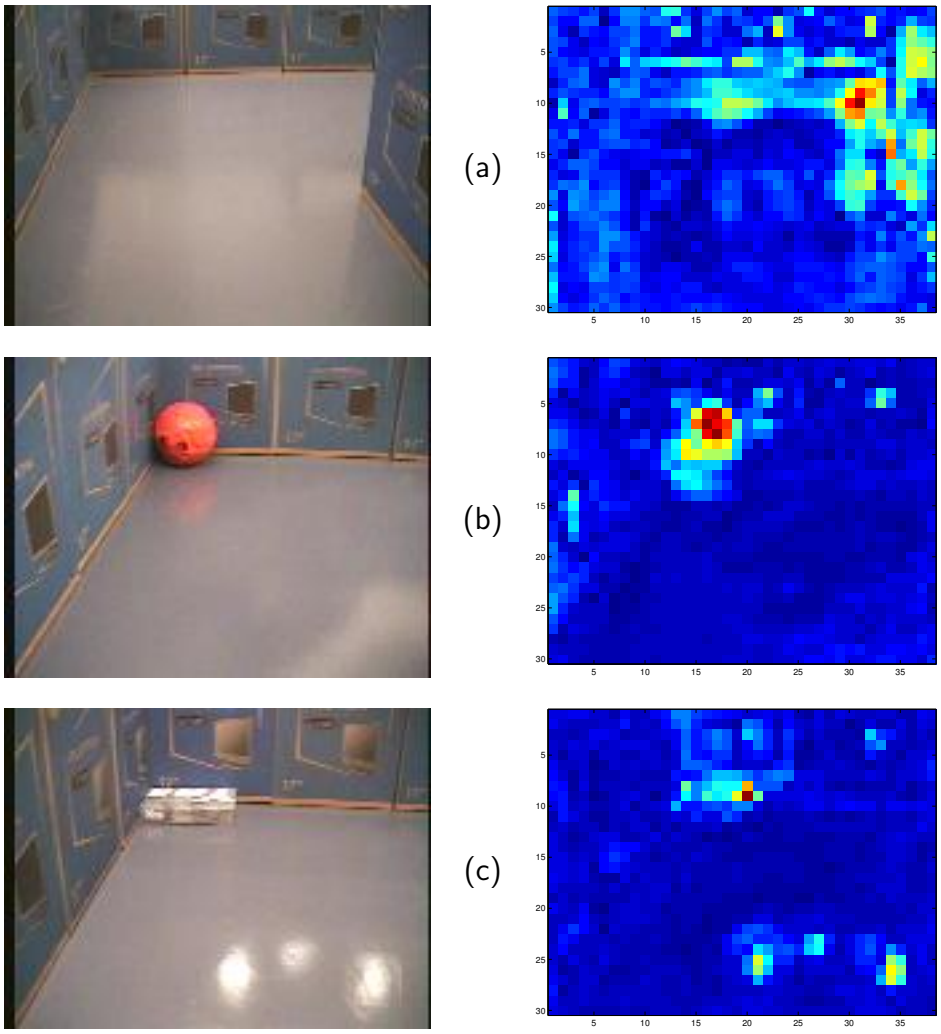



Figure 6.1: Saliency map: (a) predominant orientation features, with the most salient region corresponding to the edges at the top-right corner of the image; (b) predominant colour features, with the most salient region corresponding to the orange object in the image; (c) predominant intensity features, with the most salient regions corresponding to bright spots in the image. Low saliency values appear in blue and high saliency values appear in red.

It is important to notice in Figure 6.1 that because the saliency map is computed in scale 2 of the image pyramids in our implementation (see Subsection 2.2.1), a 1:4 reduction occurs in each image dimension (input images have $152 \times 120$ pixels and therefore yield $38 \times 30$ saliency maps).

The highest value within the saliency map needs to be searched in order to determine the location of the first focus of attention, then the second highest value needs to be found to establish the location of the second focus of attention and so on. To perform the search for saliency values in descending order, it is necessary to mark the salient locations which were previously found in the map, so that the focus of attention can shift towards the next highest saliency value. Once the model has attended a salient location, this location needs to be eliminated from further searches. This mechanism of avoiding locations that have already been attended to is known as inhibition of return (Itti et al., 1998).

Our implementation of the inhibition of return is very simple. It consists in setting saliency values to zero around the currently attended location in the saliency map. Therefore, when searching the saliency map again for the next highest value, the inhibition of return assures that previously visited locations (and a small circular region around them) will not be selected again. This effect is illustrated in Figure 6.2, where the three most salient regions of an input image are highlighted and the effect of the inhibition of return in the corresponding pseudo-coloured saliency map is shown. In this example, the rank of saliency ranges from 0 (most salient) to 2 (less salient), also indicating the order in which the regions were attended (from 0 to 2).

In Figure 6.2 and also in all the experiments reported so far in this thesis, the location of the salient points in the image frame was obtained from the coordinates in the saliency map multiplied by four to compensate for the 1:4 reduction in each dimension. The simplicity of this approach has

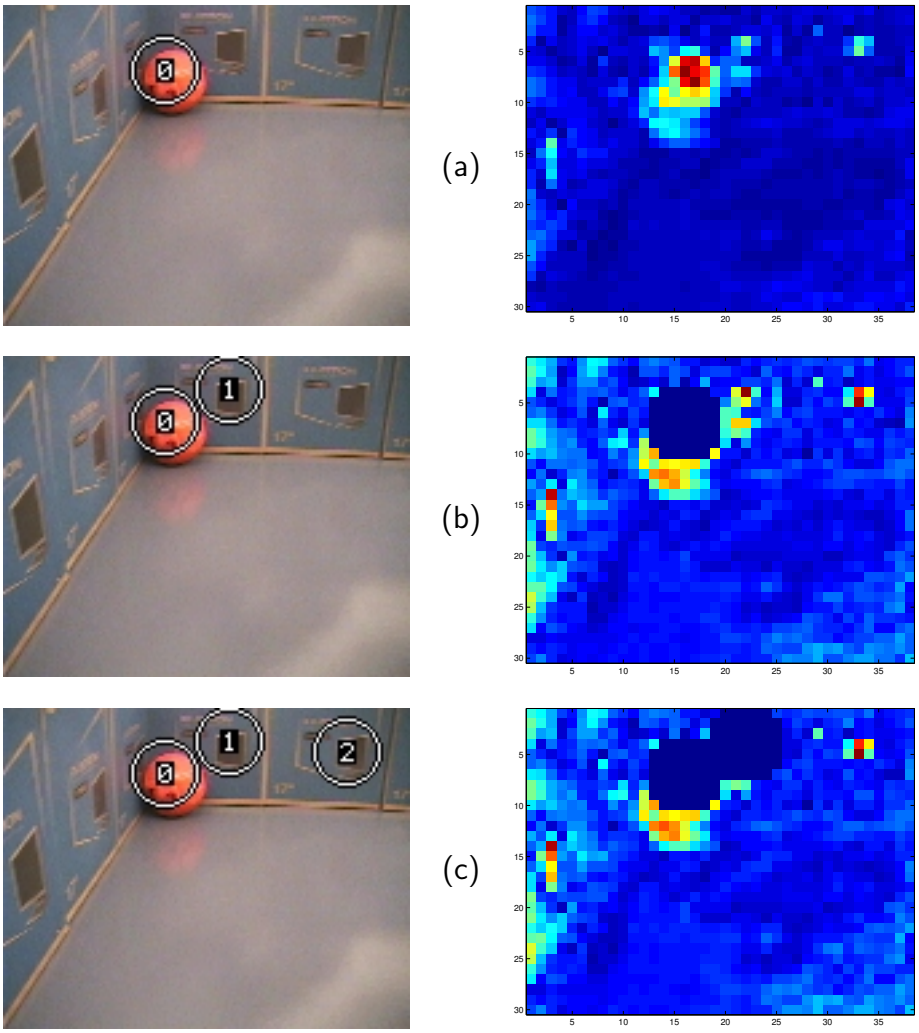



Figure 6.2: Inhibition of return: (a) first salient location with no inhibition of return; (b) second salient location with inhibition of the first location; (c) third salient location with inhibition of the first and second locations.

a serious shortcoming: the resulting resolution for the location of salient points in this case is four pixels (the coordinates of salient locations will be always multiples of four). A solution to this problem is to interpolate the location of local maxima in the saliency map to subpixel accuracy using a Taylor expansion up to the second order term because the saliency function is smooth. However, this approach cannot be followed with such a simplistic implementation of the inhibition of return because it changes the original values in the saliency map, which are needed for the interpolation.

Therefore, in this chapter we conduct experiments using a different way to select and inhibit salient regions in the saliency map, which allows inter-

polation of saliency maxima. We also compare results with the multi-scale Harris detector (see Subsection 2.2.2), which is based on the search for extrema in a Laplacian pyramid (Lowe, 2004; Mikolajczyk and Schmid, 2001). Our aim is to investigate the influence of the attention mechanism in the overall performance of the visual novelty detection system and also the possibility of implementing automatic scale selection, *i.e.* determining the size of interest regions automatically based on the characteristic scale property suggested by Lindeberg (1998).

## 6.1 Experiment 16: Accuracy and Stability

In order to study the influence of accuracy in the location of salient points — a very important issue when raw image patches are used because misalignment can lead to misclassification — we devised and implemented an alternative way of locating points in the saliency map. In this approach, the number of salient locations is determined automatically.

The saliency map is searched to determine its average saliency value ($\bar{S}$) and also its maximum saliency value ($\hat{S}$), which corresponds to the first location to be attended by the attention mechanism. These two values determine a saliency threshold ($S_T$) for the selection of salient points:

$$S_T = \bar{S} + k(\hat{S} - \bar{S}), 0 \leqslant k \leqslant 1 \tag{6.1}$$

where $k$ is a constant that determines the number of salient points. The lower the value of $k$ is, the larger is the resulting number of salient points — here we have used $k = 0.25$.

Salient locations are then determined by a search for local maxima whose values are above the saliency threshold $S_T$. Attended locations are marked in a separate inhibition map to avoid corrupting the values in the saliency map as happened in the previous implementation. The determined coordinates and their neighbours are then used to interpolate the location of the

maxima with subpixel accuracy using a second order Taylor expansion:

$$\hat{x} = -\frac{S_x}{S_{xx}} = \frac{S(x-1,y) - S(x+1,y)}{2[(x+1,y) - 2S(x,y) + S(x-1,y)]}, \tag{6.2}$$

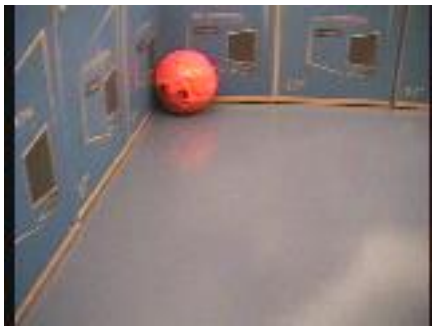
$$\hat{y} = -\frac{S_y}{S_{yy}} = \frac{S(x,y-1) - S(x,y+1)}{2[S(x,y+1) - 2S(x,y) + S(x,y-1)]}, \tag{6.3}$$

where $S_x$ and $S_y$ are the first partial derivatives and $S_{xx}$ and $S_{yy}$ are the second partial derivatives of the saliency function $S$ relative to coordinates $x$ and $y$, respectively.

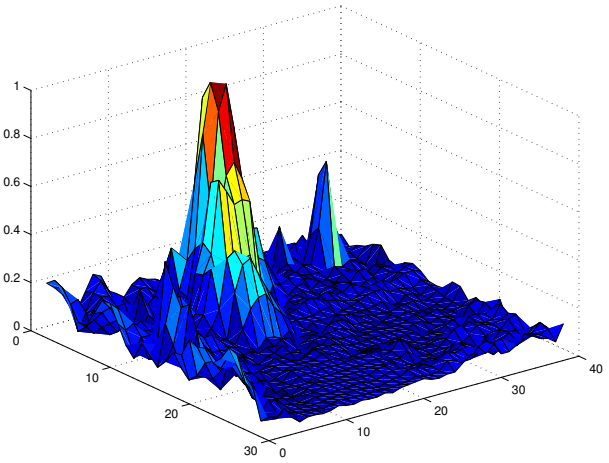
Equations 6.2 and 6.3 fit a parabola to the local saliency function in order to find the offset $(\hat{x}, \hat{y})$ to be added to the coordinates of the salient point previously found. A parabola is sufficient to interpolate a more accurate location for local maxima since the saliency function is reasonably smooth. Figure 6.3 illustrates a typical smooth landscape of the saliency function for an input image containing a salient orange ball.

The multi-scale Harris detector (Mikolajczyk and Schmid, 2001) was implemented as an alternative interest point selection strategy to the saliency map. As previously explained in Subsection 2.2.2, this algorithm basically consists in building an intensity Laplacian pyramid from the input image and then searching it for extrema, which are stable in both space and scale and therefore correspond to interest points (Lowe, 2004).

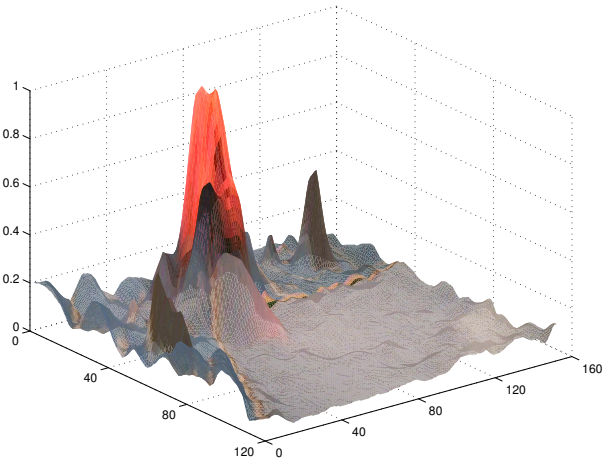
In order to compare performances of different strategies to select interest points, we conducted experiments using normalised raw image patches in the image encoding stage (the same approach used throughout Chapter 5). Raw image patches were used in this case because the overall performance of the visual novelty detection system is sensitive to patch misalignment, which obviously depends on the accuracy of the attention mechanism being used. Therefore, an attention mechanism that provides better interest point accuracy and stability is expected to also provide better overall performance when using raw image patches.



(a)




(b)




(c)

Figure 6.3: Typical saliency function landscape: (a) input image containing a very salient orange ball; (b) 3D plot of the corresponding saliency map; (c) saliency landscape with superimposed input image, showing that the highest peak corresponds to the orange ball.

For the experiments in this chapter we used the same dataset used in Chapter 5, *i.e.* images from the laboratory arena containing an orange football and a grey box. Also as in previous experiments, a fixed scale size of $24 \times 24$ pixels was used for the image patches. As mechanisms of attention, we have used the saliency map with interpolated locations as described above (referred to as "interpolated saliency") and the multi-scale Harris detector. Both of these approaches automatically decide the number of salient points to be selected within the input image.

The baseline for comparison of results was the basic saliency map with inhibition of return used in previous experiments (referred to as "coarse saliency"), in which we searched for a fixed number of three salient points. Additionally, we also ran experiments in which three interest points were selected at random, using the same inhibition of return strategy in order to avoid interest points being placed too close to each other. Because only three image patches are selected per image frame in these cases, only $3 \times 24 \times 24 = 1728$ pixels were processed from a total of $152 \times 120 = 18240$ pixels in the image frame. In other words, only 9.5% of the input image is actually analysed by the novelty filter and, therefore, a sensible selection of regions of interest is very important to detect any novelties with success. After these considerations, the performance of the random interest point selection is expected to be worse than the performance of the other attention mechanisms.

**Assessment of the attention mechanisms.** As novelty ground truth images were available for the objects of interest (the orange ball and the grey box), we have computed the percentage of selected interest regions ($P_A$) that contained at least 10% of pixels belonging to the objects in image frames where they appeared. This measurement evaluates the ability of each attention mechanism to select patches containing part of the objects of interest. The higher the percentage of regions containing part of the objects

of interest, the better are the chances of the corresponding attention mechanism to select those particular objects. We expect that a good attention mechanism selects at least one region containing the object of interest per frame, *i.e.* $P_A > 33.3\%$ when using three interest points per image frame.

However, the simple percentage of regions contemplating the object of interest is not always a faithful performance measurement for the attention mechanism. For example, it may happen that in a given image frame all selected patches contain the object of interest while in other frames the object is completely missed, but still yielding an average above the desired minimum of 33.3% when using three interest points. It is therefore also necessary to measure the percentage of image frames in which at least one selected region contains part of the object of interest ($P_1$). Good attention mechanisms are expected to result in $P_1$ being close to 100% of the image frames in which at least one region contains the object of interest.

Table 6.1 shows the percentages of regions containing part of the object of interest ($P_A$ and $P_1$) for each of the four experimented attention mechanisms (in total, 40 image frames were used to compute these percentages: 20 containing the orange football and 20 containing the grey box). Overall percentages (combined results for the orange football and the grey box) are also presented.

Table 6.1: Percentage of regions (fixed scale) containing part of the object of interest ($P_A$) and the percentage of image frames in which at least one selected region had part of the object of interest ($P_1$).

| | Random Selection | Coarse Saliency | Interpolated Saliency | Multiscale Harris Det. |
|---|---|---|---|---|
| Orange ball | $P_A = 11.7\%$<br>$P_1 = 30.0\%$ | $P_A = 65.0\%$<br>$P_1 = 100.0\%$ | $P_A = 60.0\%$<br>$P_1 = 90.0\%$ | $P_A = 44.1\%$<br>$P_1 = 100.0\%$ |
| Grey box | $P_A = 6.7\%$<br>$P_1 = 20.0\%$ | $P_A = 45.0\%$<br>$P_1 = 100.0\%$ | $P_A = 64.2\%$<br>$P_1 = 100.0\%$ | $P_A = 38.2\%$<br>$P_1 = 95.0\%$ |
| Overall | $P_A = 9.2\%$<br>$P_1 = 25.0\%$ | $P_A = 55.0\%$<br>$P_1 = 100.0\%$ | $P_A = 62.1\%$<br>$P_1 = 95.0\%$ | $P_A = 41.2\%$<br>$P_1 = 97.5\%$ |

As expected, Table 6.1 shows that the worst performance was yielded by the random selection of interest regions, which of course did not follow any specific criterion other than pure chance. Interestingly, the coarse saliency approach performed better than the interpolated saliency for the orange ball, but the situation inverted for the grey box. This effect is attributed to the fact that the grey box presents a larger colour distribution (from grey to white) than the more uniform orange ball, causing the interpolated saliency map to detect more salient points on the grey box because there is no inhibition of return in this case.

Regarding the multi-scale Harris detector, its performance was slightly worse than the performance of the interpolated saliency map concerning the percentage of regions embracing the object of interest ($P_A$), but similar concerning the percentage of frames in which at least one region contained the object of interest ($P_1$).

**Using the GWR network as novelty filter.** In order to assess the impact of the attention mechanism on the overall visual novelty detection performance, a GWR network with the same parameters as in previous experiments was trained with the raw image patches selected from the empty arena. Also as in previous experiments, the acquired model of normality of the empty arena was used to filter out any abnormal perceptions during the inspection of the environment. Inspection was conducted with the presence of the aforementioned novel objects in the arena and the results obtained with each attention mechanism are given in Table 6.2, including the sizes of the acquired models. The "overall" performance shown corresponds to the analysis of the results for the orange football and the grey box combined in a single contingency table.

All experiments resulted in statistically significant correlation between novelty ground truth and the classification made by the system ($\chi^2$ analysis, $p \leq 0.01$), except the ones using random selection of regions because the

Table 6.2: Visual novelty detection performance comparison using different interest point selection methods (fixed scale) and the GWR network.

| | Random Selection | Coarse Saliency | Interpolated Saliency | Multiscale Harris Det. |
|---|---|---|---|---|
| Model Size | 2 nodes | 3 nodes | 5 nodes | 4 nodes |
| Orange ball | $V = 0.84^*$<br>$U = 0.64$<br>$\kappa = 0.84$ | $V = 0.91$<br>$U = 0.73$<br>$\kappa = 0.91$ | $V = 0.93$<br>$U = 0.81$<br>$\kappa = 0.92$ | $V = 0.89$<br>$U = 0.73$<br>$\kappa = 0.89$ |
| Grey box | $V = 0.47^*$<br>$U = 0.23$<br>$\kappa = 0.47$ | $V = 0.77$<br>$U = 0.57$<br>$\kappa = 0.75$ | $V = 0.76$<br>$U = 0.53$<br>$\kappa = 0.73$ | $V = 0.59$<br>$U = 0.30$<br>$\kappa = 0.51$ |
| Overall | $V = 0.70^*$<br>$U = 0.45$<br>$\kappa = 0.70$ | $V = 0.84$<br>$U = 0.62$<br>$\kappa = 0.83$ | $V = 0.81$<br>$U = 0.56$<br>$\kappa = 0.80$ | $V = 0.75$<br>$U = 0.50$<br>$\kappa = 0.72$ |

*Ill-conditioned contingency tables for the $\chi^2$ test

resulting contingency tables were ill-conditioned for the $\chi^2$ test (the corresponding tables of expected values had entries with values below 5, see Section 3.3). Nevertheless, the statistics $V$, $U$ and $\kappa$ were still computed in order to assess the *classification* ability of the GWR network when using the random image patch selection strategy.

In fact, the strength of the association between the GWR network response using random selection of regions and ground truth data was still very reasonable. However, this result must be interpreted with care. It is also important to remember how often the novel features are selected by the attention mechanism. When random selection is used, there is no guarantee that novel features are ever going to be candidates (see Table 6.1). This is illustrated in Figure 6.4, where one can notice that both the coarse and interpolated saliency approaches and the multi-scale Harris detector have successfully selected image patches that contain part of the objects of interest, while the random selection has failed completely in the cases shown. Nevertheless, the GWR network was able to classify the three randomly selected regions shown in Figure 6.4a correctly as being non-novel.

It can also be noticed in Figure 6.4 that the saliency map prefers blobs and straight edges, while the multi-scale Harris detector prefers edges with

high curvature. The location of interest points selected by the multi-scale Harris detector are assigned with crosses instead of numbers because there is no saliency rank in this case. In fact, the same applies to random selection

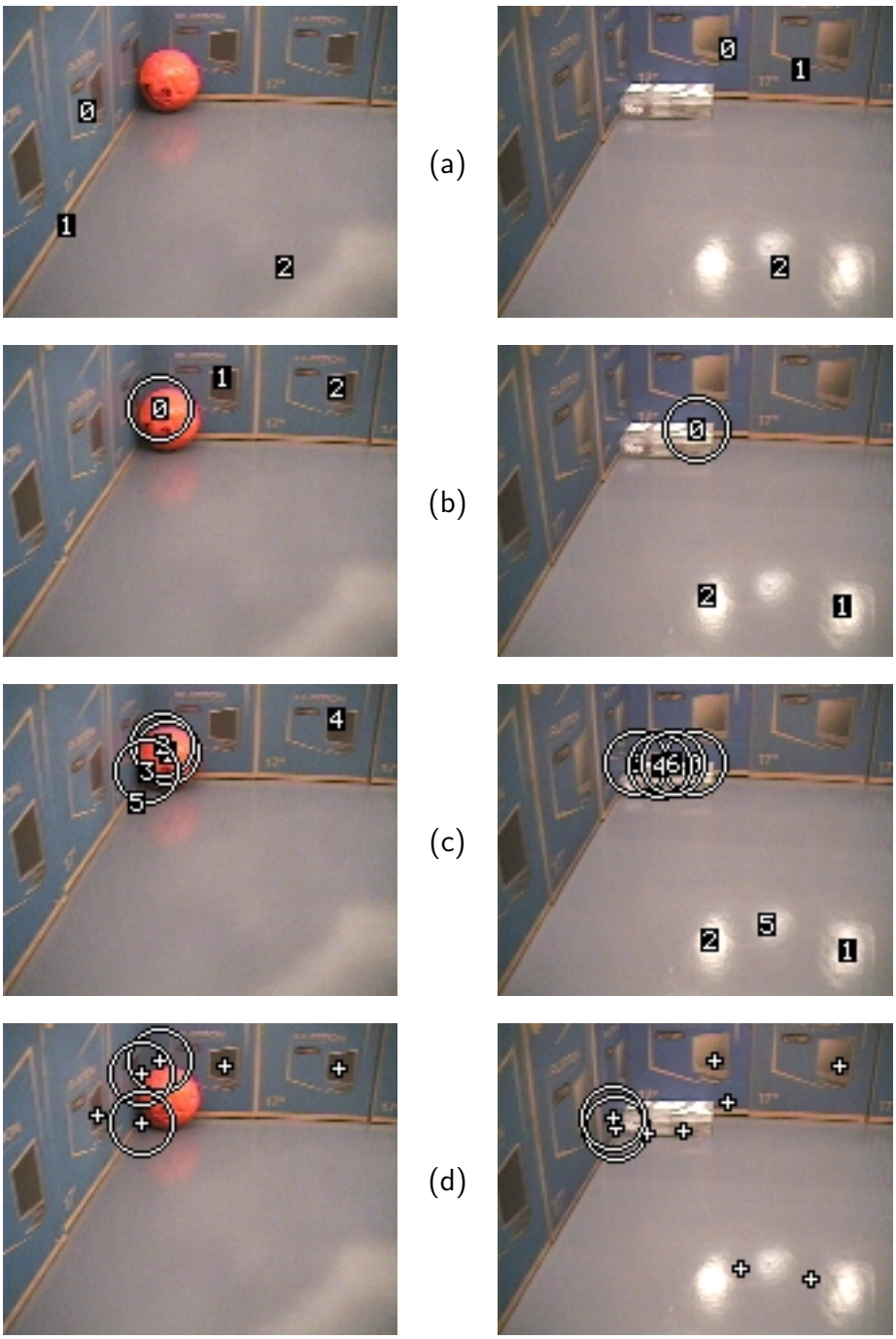



Figure 6.4: Output images (fixed scale, GWR network): (a) random selection with inhibition of return; (b) coarse saliency map with inhibition of return; (c) interpolated saliency map; (d) multi-scale Harris detector. Random selection is the only strategy that does not guarantee that interest points will lie within the objects of interest (the orange football on the left and images and the grey box on the right images).

of interest points, but the numbers serve the purpose of showing the lack of correlation between interest points in a sequence of frames. Image patches corresponding to interest points close to the borders of the input frame were discarded if their size was smaller than the fixed size of $24 \times 24$ pixels.

**Using incremental PCA as novelty filter.** For the sake of completeness, the same experiments were repeated using the incremental PCA algorithm as novelty filter. The results obtained are shown in Table 6.3 and include the sizes of the acquired models. Once more, the "overall" performance corresponds to the analysis of the results for both novel objects combined in a single contingency table.

Table 6.3: Visual novelty detection performance comparison using different interest point selection methods (fixed scale) and incremental PCA.

| | Random Selection | Coarse Saliency | Interpolated Saliency | Multiscale Harris Det. |
|---|---|---|---|---|
| Model Size | 21 vectors (20 dim.) | 24 vectors (22 dim.) | 30 vectors (28 dim.) | 28 vectors (27 dim.) |
| Orange ball | $V = 0.55^*$ $U = 0.45$ $\kappa = 0.47$ | $V = 0.87$ $U = 0.70$ $\kappa = 0.87$ | $V = 0.84$ $U = 0.61$ $\kappa = 0.84$ | $V = 0.94$ $U = 0.83$ $\kappa = 0.94$ |
| Grey box | $V = 0.29^*$ $U = 0.10$ $\kappa = 0.28$ | $V = 0.94$ $U = 0.84$ $\kappa = 0.94$ | $V = 0.75$ $U = 0.50$ $\kappa = 0.73$ | $V = 0.63$ $U = 0.31$ $\kappa = 0.62$ |
| Overall | $V = 0.46^*$ $U = 0.26$ $\kappa = 0.41$ | $V = 0.91$ $U = 0.76$ $\kappa = 0.90$ | $V = 0.79$ $U = 0.54$ $\kappa = 0.77$ | $V = 0.80$ $U = 0.53$ $\kappa = 0.79$ |

*Ill-conditioned contingency tables for the $\chi^2$ test

Once again, all results showed statistically significant association between system response and actual novelty status ($\chi^2$ analysis, $p \leq 0.01$) when incremental PCA was used as novelty filter, except the ones using random selection of regions (resulting contingency tables were ill-conditioned for the $\chi^2$ test). It can be noticed that the number of acquired vectors and their dimension is slightly larger for the cases in which the interpolated saliency approach or the multi-scale Harris detector were used, indicating

that a higher number of relevant eigenvectors are obtained when using these attention mechanisms. Results were much worse when random selection of image patches was used, showing the higher sensitivity of the incremental PCA algorithm to the accuracy and stability of the interest point detector in use when compared to the GWR network.

The coarse saliency map provided the best overall results and therefore, taking into account the important information about the percentage of selected regions containing part of the objects of interest (see Table 6.1), it is the best choice among the investigated attention mechanisms.

The reconstructed images from the acquired incremental PCA models using the interpolated saliency map and the multi-scale Harris detector as attention mechanisms are shown in Figure 6.5, where one can notice that the acquired models using either interest point detector are quite similar. Only reconstructed images from incremental PCA models acquired using the interpolated saliency map or the multi-scale Harris detector are shown because these are the only attention mechanisms that will be used in the experiments on automatic scale selection to be presented in the next section.

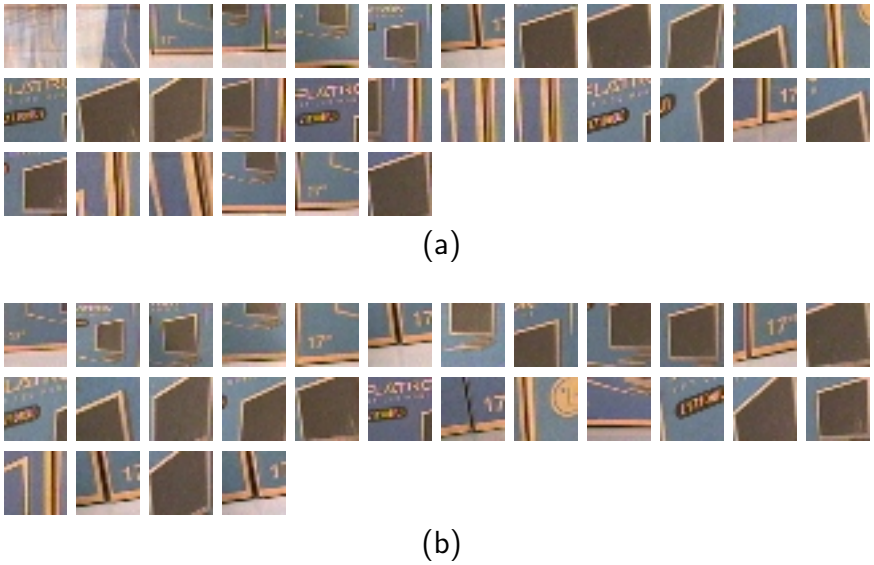


Figure 6.5: Image patches (fixed scale) acquired using incremental PCA: (a) interpolated saliency map; and (b) multi-scale Harris detector. Both models are similar in contents and in size.

## 6.2 Experiment 17: Automatic Scaling

As a result of robot motion around the environment, visual features are subject to several geometric transformations. The use of attention mechanisms provides robustness to translations by selecting salient characteristic locations within the image frame. Both attention mechanisms being investigated in this chapter rely on multi-scale pyramidal (also known as scale-space) representations, which provide stable salient locations regardless of changes in scale or translations in space.

Changes in scale are also relevant when the robot approaches objects. In our experiments so far, generalisation according to scale was achieved by acquiring multiple versions of salient visual features at different scales by the learning mechanism in use. If the image encoding stage is made invariant to changes in scale, this would improve the overall system generalisation ability and reduce the amount of acquired nodes or vectors in the model of normality of the environment.

**Characteristic scale.** Lindeberg has shown that the characteristic scale of a pixel within an image can be determined by locating the extremum of the Laplacian jet of that particular pixel (Lindeberg, 1998). The Laplacian jet of a given pixel is the function across the levels of the intensity Difference-of-Gaussian pyramid of an image at the coordinates of the pixel. The response of the Laplacian will be the highest at the scale in which the contrast between close neighbouring pixels is maximal, which by definition corresponds to the characteristic scale of that location.

Because both attention mechanisms used in this chapter already make use of Laplacian pyramids, we can use them to compute the characteristic scale of the selected interest points and use it to determine the approximate size of their corresponding region of interest, *i.e.* the size of the image patch to be cropped from the input frame. This strategy was used by Lowe (2004)

and by Crowley, Riff, and Piater (2002) to determine the region of interest surrounding visual features to be encoded.

Once the location of an interest point is defined, the Laplacian jet at that location needs to be searched for an extremum. The scale can then be determined precisely by interpolation using a second order Taylor expansion:

$$\hat{s} = -\frac{L_s}{L_{ss}} = \frac{L(s-1) - L(s+1)}{2[L(s+1) - 2L(s) + L(s-1)]}, \tag{6.4}$$

where $s$ is the level of the pyramid in which the extremum was found, $L_s$ and $L_{ss}$ are the first and second partial derivatives of the Laplacian function $L$ relative to the level $s$, respectively.

The offset $\hat{s}$ is added to the extremum level in order to determine scale with better accuracy. According to Crowley, Riff, and Piater (2002), the radius of the region of interest can be computed from the interpolated half-octave pyramid level by using the following equation:

$$r_{roi} = 1.18 \times b^{(s+\hat{s})}, \tag{6.5}$$

where the constant 1.18 is an empirical correction factor for the scale, which is given by a geometric progression with base $b = \sqrt{2}$. Two levels of the pyramid are necessary to change scale by a factor of two, hence the name "half-octave pyramid".

The procedure above can be performed directly in the case of the multi-scale Harris detector because in our implementation we use a scale-space with five octaves, *i.e.* a half-octave Laplacian pyramid with ten levels (Crowley et al., 2002), which provides sufficient scale resolution. However, in the case of the saliency map, the intensity Laplacian pyramid used has only five levels. Therefore, an additional half-octave Laplacian pyramid is built using the intensity channel with the sole purpose of computing the characteristic scale of salient points.

In our implementation of automatic scale selection, we selected regions

of interest with twice the radius computed with equation 6.5, in order to guarantee that edges would be present in the image patches. Also, the patch radius was limited to a minimum of 6 pixels and a maximum of 24 pixels:

$$r = \min\{\max\{6, 2 \times r_{roi}\}, 24\}, \tag{6.6}$$

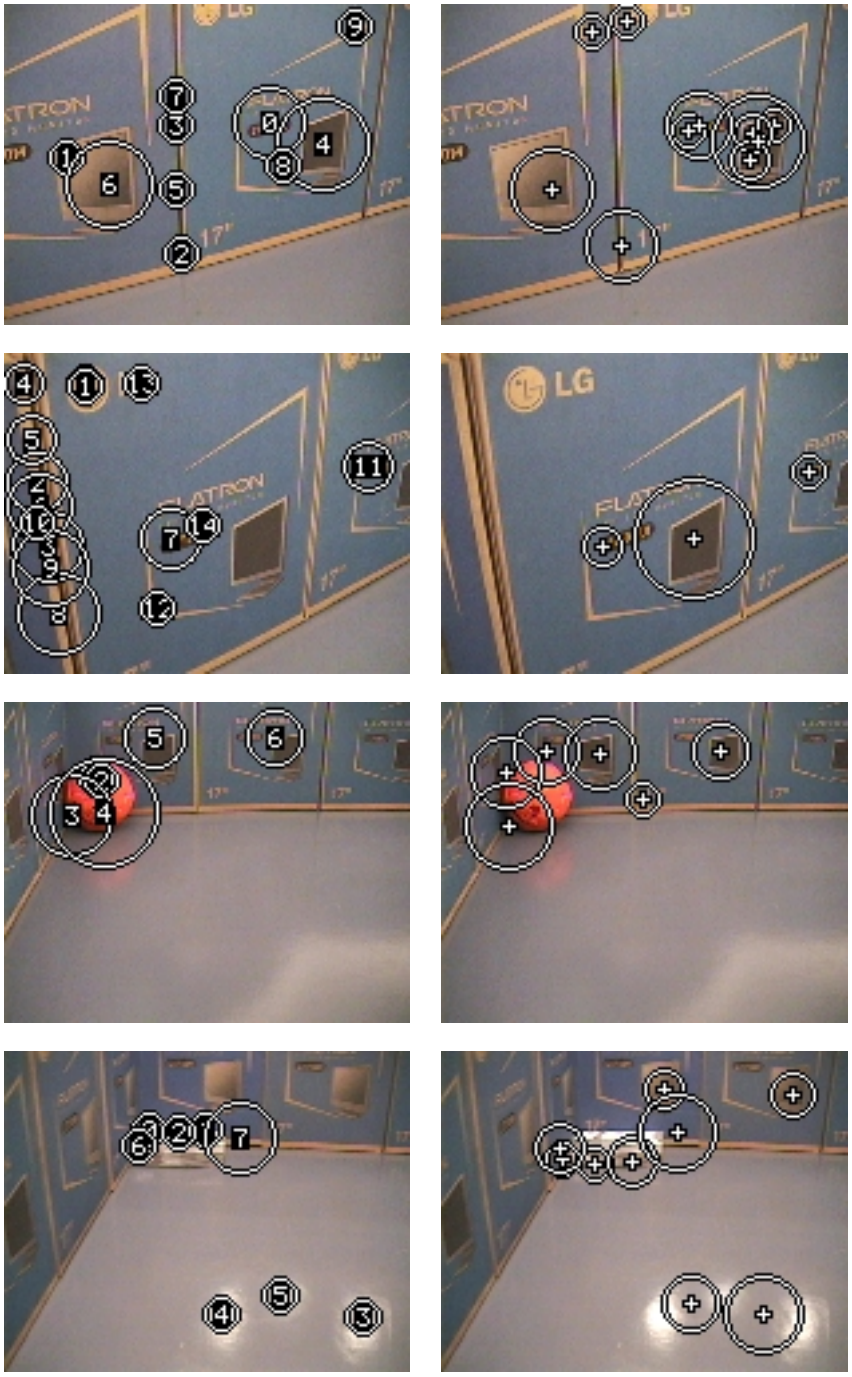
resulting in the selection of square image patches centred around the interest points ranging from $12 \times 12$ to $48 \times 48$ pixels in size.

Figure 6.6 shows examples of interest points selected by the interpolated saliency map and the multi-scale Harris detector, and their respective regions of interest, whose sizes were calculated according to equation 6.6.

The circles in Figure 6.6 designate the size of the regions of interest according to the automatic scale selection of the corresponding interest point. There was no novelty detection involved in the generation of these output images, just the use of the attention mechanisms with automatic scale selection to determine the size of the regions. It is important to notice that when it happens to both attention mechanisms to decide for interest points in similar locations, the size of the corresponding regions of interest are also similar. In these examples, like in Figure 6.4 on page 163, it is also possible to notice that the multi-scale Harris detector selects interest points on edges with high curvature, while the saliency map selects interest points on blobs and straight edges.

**Assessment of the automatic scale selection.** Using the available ground truth images for the objects of interest (the orange ball and the grey box), once more we computed the percentages of interest regions that contained at least 10% of pixels belonging to the objects ($P_A$ and $P_1$), this time using automatic scale selection. The results obtained for both attention mechanisms investigated are given in Table 6.4.

A comparison with the results using fixed scale (Table 6.1) shows that the percentage of selected regions that contain part of the objects of interest



(a) (b)

Figure 6.6: Output images with automatic scale selection: (a) interpolated saliency map; (b) multi-scale Harris detector. Interest points are indicated by numbers in (a) or crosses in (b) and the size of their respective regions of interest are indicated by white circles (there was no novelty detection involved in the generation of these images).

($P_A$) has increased, especially in the case of the multi-scale Harris detector. This means that there was an improvement in performance of both attention mechanisms in question. We therefore proceeded with experiments using

Table 6.4: Percentage of regions (auto scale) containing part of the object of interest ($P_A$) and the percentage of image frames in which at least one selected region had part of the object of interest ($P_1$).

| | Interpolated Saliency | Multiscale Harris Det. |
|---|---|---|
| Orange ball | $P_A = 58.5\%$<br>$P_1 = 100.0\%$ | $P_A = 38.1\%$<br>$P_1 = 95.0\%$ |
| Grey box | $P_A = 62.0\%$<br>$P_1 = 100.0\%$ | $P_A = 48.7\%$<br>$P_1 = 95.0\%$ |
| Overall | $P_A = 60.2\%$<br>$P_1 = 100.0\%$ | $P_A = 43.4\%$<br>$P_1 = 95.0\%$ |

the whole visual novelty detection framework to assess the impact caused in overall performance by using automatic scale selection.

To obtain the feature vectors with a fixed number of dimensions needed by the learning mechanisms, the image patches selected by the attention mechanisms were scaled to a fixed size of $24 \times 24$ pixels through bilinear interpolation. This allows changes in scale from 1:2 to 2:1 because the original patch sizes range from $12 \times 12$ to $48 \times 48$ pixels (see equation 6.6).

**Using the GWR network as novelty filter.** First, we trained a GWR network using image frames collected when the robot was exploring the empty arena, as happened in previous experiments (GWR network parameters were also the same). The acquired model of normality was then used to filter out abnormal visual features in image frames collected during inspection of the arena containing two different objects, the orange football and the grey box. Table 6.5 shows the quantitative results obtained.

The results in Table 6.5 show that only the use of the multi-scale Harris detector as attention mechanism resulted in statistically significant association between the GWR network response and ground truth data ($\chi^2$ test, $p \leq 0.01$) for both novel objects. Overall performance using both approaches was poor when compared to the results obtained with fixed scale (see Table 6.2). Also, it is important to bear in mind that the overall results

Table 6.5: Performance comparison between different interest point selection methods (automatic scale) using the GWR network. Only the multi-scale Harris detector contributed to statistically significant association between novelty filter response and actual novelty status for both novel objects.

| | Interpolated Saliency | Multiscale Harris Det. |
|---|---|---|
| Model Size | 4 nodes | 2 nodes |
| Orange ball | $V = 0.83$<br>$U = 0.69$<br>$\kappa = 0.88$ | $V = 0.47$<br>$U = 0.17$<br>$\kappa = 0.47$ |
| Grey Box | $V = 0.02$*<br>$U = 0.00$<br>$\kappa =$ -0.02 | $V = 0.25$<br>$U = 0.05$<br>$\kappa = 0.15$ |
| Overall | $V = 0.52$<br>$U = 0.20$<br>$\kappa = 0.50$ | $V = 0.31$<br>$U = 0.07$<br>$\kappa = 0.28$ |

*Ill-conditioned contingency table for the $\chi^2$ test

for the saliency map show statistical significance between ground truth data and system response exclusively due to the correct detection of the orange ball, as the grey box was completely missed. Although this does not correspond to the problem with adding contingency tables described at the end of Section 3.3, this result must be considered with care.

The resulting poor performance is attributed to the use of bilinear interpolation scaling, which causes image patch smoothing (a low-pass filtering effect). Smoothing makes discrimination of image patches using the Euclidean metric, used by the GWR network algorithm, more difficult and this is reflected in the small number of acquired nodes.

**Using incremental PCA as novelty filter.** The experiments were repeated using incremental PCA, which we expected to be less sensitive to bilinear interpolation smoothing than the GWR network.

The expected outcome of using automatic scale selection was that smaller PCA models of normality would be acquired as a consequence of the ability to generalise scale, rather than acquiring multiple scaled versions of the same features by the learning mechanism.

A quantitative comparison of the results obtained is given in Table 6.6, which also includes the size of the acquired models.

Table 6.6: Performance comparison between different interest point selection methods (automatic scale) using incremental PCA. All experiments resulted in statistically significant association between novelty filter response and actual novelty status ($\chi^2$ test, $p \leq 0.01$ except otherwise noted).

| | Interpolated Saliency | Multiscale Harris Det. |
|---|---|---|
| Model Size | 20 vectors (19 dim.) | 11 vectors (10 dim.) |
| Orange ball | $V = 0.94$<br>$U = 0.80$<br>$\kappa = 0.94$ | $V = 0.51$<br>$U = 0.20$<br>$\kappa = 0.50$ |
| Grey Box | $V = 0.56$<br>$U = 0.28$<br>$\kappa = 0.50$ | $V = 0.17^*$<br>$U = 0.02$<br>$\kappa = 0.10$ |
| Overall | $V = 0.76$<br>$U = 0.49$<br>$\kappa = 0.75$ | $V = 0.29$<br>$U = 0.06$<br>$\kappa = 0.27$ |

$^*p \leq 0.05$

Despite revealing statistically significant association between system response and ground truth data ($\chi^2$ test, $p \leq 0.01$), the results in Table 6.6 are poorer than the results obtained using fixed scale (see Table 6.3). Nevertheless, both systems were able to correctly highlight the novel objects, as shown in Figure 6.7. The acquired models of normality are smaller than the ones acquired using image patches with fixed size, as expected. Detection of the grey box was more difficult when using the multi-scale Harris detector as attention mechanism and incremental PCA as novelty filter. This is indicated by a lower level of statistical significance between system response and actual novelty status for the grey box ($\chi^2$ test, $p \leq 0.05$) and small values for $V$, $U$ and $\kappa$ (comparable to random guessing, see Section 3.3).

The interpolated saliency map shows better performance (strong agreement between novelty filter response and actual novelty status) than the multi-scale Harris detector (weak agreement) in this context. The reconstructed images from the acquired incremental PCA models using automatic

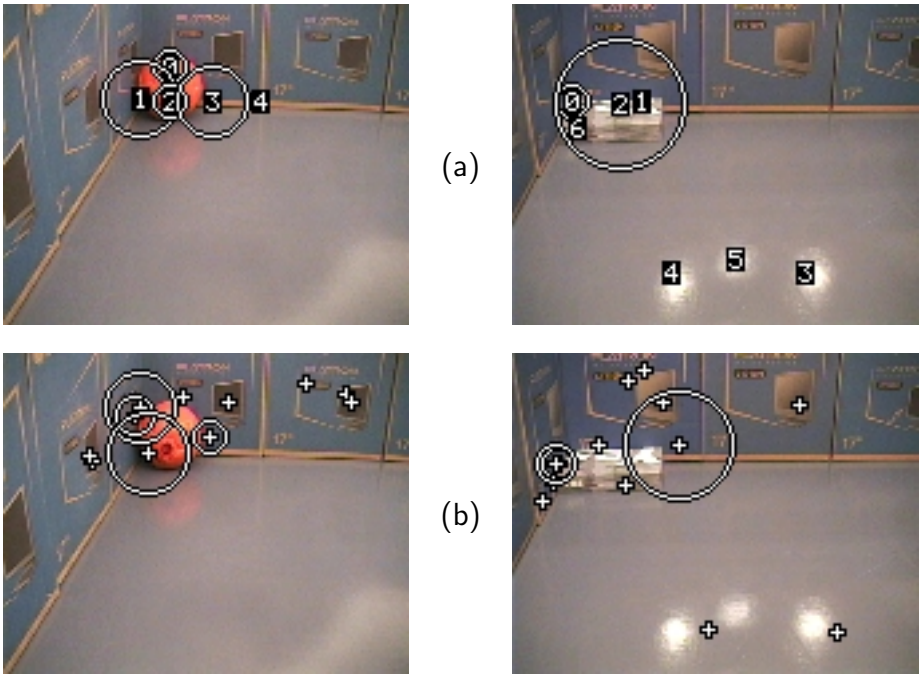



Figure 6.7: Output images (auto scale, incremental PCA): (a) interpolated saliency map; (b) multi-scale Harris detector. Most of the image regions containing part of novel objects are correctly identified with white circles.

scale selection are shown in Figure 6.8, where the similarity between the acquired models using either interest point detector can be seen. The fact that the acquired models using automatic scale selection are smaller can also be confirmed by comparisons with Figure 6.5 on page 165.

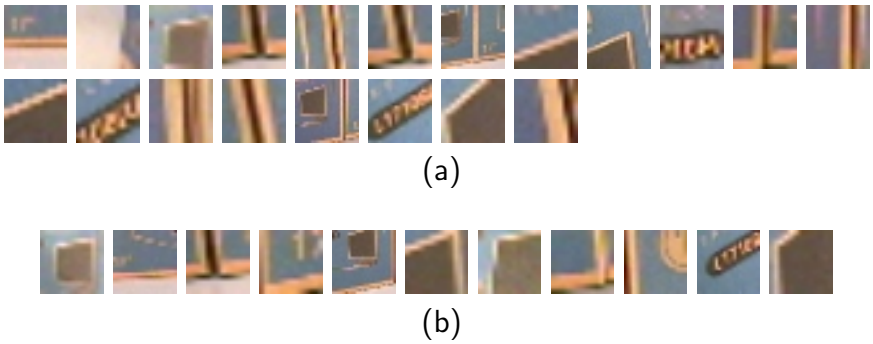



Figure 6.8: Image patches (auto scale) acquired using incremental PCA: (a) interpolated saliency map; and (b) multi-scale Harris detector. Both models are smaller than the ones acquired using fixed scale (multiple scaled versions of equivalent features have disappeared).

## 6.3 Summary and Discussion

In this chapter we have made an assessment of the influence of the attention mechanism within our visual novelty detection framework, particularly with respect to the accuracy in the localisation of interest points and the use of automatic scale selection.

We investigated two distinct interest point detection techniques: the saliency map (Itti et al., 1998), which selects interest points on blob-like features and straight edges, and the multi-scale Harris detector (Mikolajczyk and Schmid, 2001), which selects interest points on edges with high curvature. Both approaches had their localisation accuracy improved through function interpolation using a second order Taylor expansion as suggested by Lowe (2004).

Performance of the attention mechanisms was evaluated by computing the percentage of selected regions that contained part of the objects of interest for which novelty ground truth was available (an orange football and a grey box). We also computed the percentage of image frames in which at least one selected region contains the novel object. These measurements are related to the probability of the attention mechanism to select regions containing part of the objects of interest as candidate regions to be classified by the novelty filter. As expected, our results show that the use of attention criteria to select interest points yields superior performance than random selection of regions (Table 6.1).

Because there are advantages in using raw image data in order to allow image reconstruction, accuracy in interest point selection also became an important issue. Accurate localisation reduces errors due to misalignment of image patches during comparisons, having an impact in the overall performance of the visual novelty filter and also reducing the size of the model of normality that is learnt from the environment.

Another issue of concern is the robustness to changes in scale of vi-

sual features as a result of robot navigation around the environment. In the experiments in previous chapters, generalisation with respect to scale happened through the acquisition of many scaled versions of the same visual features by the learning mechanism. We tested the hypothesis that some degree of scale invariance incorporated in the image encoding stage would reduce the size of the learnt models and improve overall robustness to changes in scale, through experiments using the automatic scale selection method originally proposed by Lindeberg (1998) and efficiently implemented by Crowley, Riff, and Piater (2002). The results in Table 6.6 and Figure 6.8 corroborate the hypothesis that the use of automatic scale selection reduces the size of the PCA model of normality. However, overall performances of the novelty filters were worse than when using fixed scale image patches (compare Tables 6.3 and 6.6).

Among the experimented models of attention, the interpolated saliency map is the one that offers the most consistent results, particularly when using incremental PCA as novelty filter. Concerning automatic scale selection, the implementation reported here is not the most efficient because it uses an additional Laplacian pyramid to this effect. An implementation of the saliency map built from half-octave pyramids as in (Crowley et al., 2002) instead of the pyramidal structure suggested in (Greenspan et al., 1994; Itti et al., 1998) constitutes a better scenario for future investigations in automatic scale selection.

The saliency map also offers additional advantages, such as the possibility of top-down biasing when *a priori* knowledge about important visual features are known. For instance, if bright spots constitute important visual features for some application, the combination of features (see Chapter 2) can be biased towards the intensity channel of the saliency map, rendering intensity more salient than colour or orientation. Furthermore, the saliency map is open for the inclusion of more visual features in addition to the ones

included in its original implementation (Itti et al., 1998). Examples of additional visual features of interest are flicker and motion (Dhavale and Itti, 2003).

# Chapter 7

# Conclusion

An approach to perform visual novelty detection with applications in environment inspection using mobile robots is presented in this thesis. The general framework uses a mechanism of visual attention that selects candidate image patches in the input frame, which are then encoded and classified as novel or non-novel by a novelty-detecting clustering mechanism. This ability to differentiate between common and uncommon stimuli is essential to robots operating in dynamic environments and is at the core of applications involving automated exploration, inspection and surveillance.

For real world applications like sewer inspection, vision is the sensor modality of choice because of the rich range of information about the environment in high resolution that it can provide. Moreover, vision does not only provide information restricted to the visual domain, such as colour and texture, but can also be used to estimate shape, size and distance to objects. All of this information is useful for robots operating in complex environments, but comes at the cost of processing large amounts of data with limited computational resources and therefore poses challenges to real-time operation.

We demonstrated that the use of attention mechanisms such as the saliency map (Itti et al., 1998) or the multi-scale Harris detector (Mikolajczyk and Schmid, 2001) minimise the amount of data to be processed and

at the same time localise *where* the novel features are in the image frame. Robustness to geometric transformations (*e.g.* translations and changes in scale) due to robot motion is also improved by the use of stable interest point detectors, avoiding explicit segmentation of the input image.

Because novelty is of contextual nature and therefore can not be easily modelled, the approach that we follow is to first acquire a model of *normality* through robot learning and then use it as a means to highlight any abnormal features that are introduced in the environment. For this purpose, we have used unsupervised clustering mechanisms such as the GWR neural network (Marsland et al., 2002b) and the incremental PCA algorithm (Artač et al., 2002), which are able to learn aspects of the environment incrementally.

We proposed an experimental setup to evaluate performance and functionality of visual novelty filters in Chapter 3. The experimental procedure was divided into two stages: an exploration phase, in which the learning mechanism was enabled to allow the robot to build a model of normality while experiencing the environment; and an inspection phase, in which the acquired model of normality is used as a novelty filter. Novel objects were inserted in the robot's environment during the inspection phase of experiments with the expected outcome that the visual novelty filter would produce indications of novelty and localise these new objects in the input image frame.

**Performance assessment.** As the precise location and nature of the novelty introduced during the inspection phase is known by the experimenter, it is possible to generate ground truth data to be compared with the responses given by the novelty filter. In order to assess the performance of a novelty filter objectively, we used $2 \times 2$ contingency tables relating actual novelty status (ground truth) to system response, followed by the computation of statistical tests to quantify the association or agreement between them. Here we used the $\chi^2$ test in order to check the statistical significance

of the association between ground truth and novelty filter response, followed by the computation of Cramer's $V$, the uncertainty coefficient $U$ and the $\kappa$ index of agreement (Sachs, 2004).

Extensive experimental data was logged to evaluate and compare the efficiency of the different components of the visual novelty filter, namely the visual attention mechanism, the image encoding and the unsupervised clustering algorithm. The $\chi^2$ analysis of the generated contingency tables revealed statistical significance in the associations between system response and actual novelty status in most of the reported experiments. The best configurations of the visual novelty filter presented strong agreement with the ground truth data. Typical quantitative analyses resulted in the following values: $V = [0.65, 0.85]$, $U = [0.35, 0.65]$ and $\kappa = [0.65, 0.85]$.

Qualitative assessment of the learning procedure during exploration, as well as consistent identification of novel features during inspection was made through the use of novelty bar graphs. In these graphs, a measure of the degree of novelty in each image frame is plotted against time/position. Novelty graphs are particularly useful to identify novelty indications in unexpected locations of the environment and investigate their reasons, leading to improvements in overall system robustness and ability to generalise.

We compared several instances of our framework for visual novelty detection, using both qualitative and quantitative tools. In Chapter 4, colour statistics were used as image encoding method in order to provide feature vectors to a GWR network. We experimented with colour histograms and colour angles (Finlayson et al., 1996) to encode colour distributions, both in global and local fashion. Local encoding through the use of the saliency map as attention mechanism provided better results than global encoding in most cases. Furthermore, local encoding also enabled localisation of novel features in the input frame.

**Colour statistics.** However, colour histograms and colour angles cannot represent colour distributions containing shades of grey. In order to solve this problem, we added information about the intensity distribution to the colour angular encoding. This enhanced colour angular encoding provided the best overall results (strong agreement with ground truth data) when used in a local fashion. Also, the compact representation provided by the enhanced colour angular indexing resulted in a visual processing speed of up to eight frames per second when using the robot's on-board computer.

**Raw image data.** Despite being very discriminative, colour statistics obviously do not encode other important visual features such as texture and structural information. Therefore, in Chapter 5 we conducted more experiments using normalised raw image patches as input vectors to the clustering mechanism. The use of raw visual data was only possible with the use of local encoding through the selection of candidate regions using the saliency map as attention mechanism in order to deal with image transformations.

**GWR network versus incremental PCA.** Two novelty-detecting clustering mechanisms, both capable of on-line unsupervised learning, were used in these experiments with raw image patches, the GWR neural network and the incremental PCA algorithm. Both mechanisms provided very good results (almost perfect agreement with ground truth data). The use of raw image patches also added a useful functionality to the framework: the visual reconstruction of the acquired models. In general, the models acquired by the GWR-based novelty filter were smaller than the ones acquired using incremental PCA. On the other hand, incremental PCA provided better image patch reconstruction.

**Reliance on repeatable trajectories.** Some of the experiments using raw image patches also investigated the sensitivity to changes in the robot's

trajectory, because the navigation behaviour was always based on a very simple force-field strategy for obstacle avoidance, which was completely independent of the visual input. Results of these experiments revealed that the performance of our approach degrades when the robot's trajectory during inspection is not the same as the trajectory during exploration of the environment. Nevertheless, quantitative analysis showed that performance in all instances was still very good (strong agreement with ground truth data) even with variations in the robot's trajectory in the environment.

**Saliency map versus multi-scale Harris detector.** Concerning the attention mechanism, experiments in Chapter 6 were conducted to compare the performances of the saliency map and the multi-scale Harris detector. We also conducted experiments using a random interest point selection strategy to serve as a baseline for comparisons. Both saliency map and multi-scale Harris detector produced better results than random selection of interest regions, with the saliency map performing noticeably better than the multi-scale Harris detector for the purposes of novelty detection. We attribute this to the fact that the saliency map uses a normalisation operator, which enhances features that are less common and suppresses more common features within the input frame.

**Automatic scaling.** Using the fact that both attention mechanisms — saliency map and multi-scale Harris detector — use image pyramids to form a scale-space representation, we also conducted experiments using Lindeberg's automatic scale selection method (Lindeberg, 1998). The aim of these experiments was to determine the patch size surrounding interest points automatically and to make raw image patches more robust to changes in scale. The acquired models of normality had their sizes reduced when using automatic scale selection, especially when the incremental PCA algorithm was used, confirming an improvement in the ability to generalise scale. However,

overall performance in visual novelty detection was worse than when using raw image patches with fixed scale, particularly when using the multi-scale Harris detector. This effect is attributed to the use of bilinear interpolation to rescale image patches, which removes the high frequency contents of edges and therefore makes image matching more difficult. In general, the saliency map performed better than the multi-scale Harris detector within our framework of visual novelty detection.

We consider the results obtained to be very good and likely to succeed in real world applications that involve exploration and inspection of environments using vision. An example of such an application is the automated inspection of sewer pipes or air-conditioning ducts, as demonstrated in the real-world experiments reported in Subsection 5.4.1. However, more elaborate processing to obtain robustness to general affine transformations is necessary for applications in which the environment is not as structured as the arenas, corridors and ducts that were used as operating environments in this work (see Section 7.1).

**Contributions.** One of our main contributions was to implement and experiment with visual novelty detection mechanisms for applications in automated inspection using autonomous mobile robots. Previous work done in novelty detection used only low resolution sonar readings (Crook et al., 2002; Marsland et al., 2002a) or very restricted monochrome visual input (Crook and Hayes, 2001; Marsland et al., 2001). In contrast to this, here we used colour visual stimuli with unrestricted field of view. The selection of algorithms had emphasis on bottom-up and unsupervised learning approaches to allow exploitation of relevant characteristics of the acquired data from the ground-up.

Quantitative performance assessment tools based on contingency table analysis and statistical tests were developed in order to support objective comparisons between different visual novelty filters. For comparison pur-

poses, novelty ground truth maps were generated in the form of binary images, in which novel visual features are highlighted manually. Because vision is a sensor modality shared between robots and humans, generation of novelty ground truth maps occurs in a natural and relatively easy way (although it demands time because of the volume of images involved).

Our experiments demonstrated that visual models of normality can be readily acquired by on-line unsupervised clustering mechanisms and used later to detect novelties introduced in the operating environment correctly. Other approaches to visual novelty detection mentioned in the literature require off-line supervised training (Diehl and Hampshire II, 2002; Singh and Markou, 2004). The novelty filters studied in this work are able to learn on-line and our general framework presents real-time processing capabilities, which are important characteristics for the autonomous operation of mobile robots with limited computational resources.

Another main contribution was the demonstration that attention mechanisms extend the functionality of visual novelty filters, enabling them to localise *where* the novel regions are in the input frame, improving image encoding robustness to geometric transformations due to robot motion. Also, the use of an interest point detector as attention mechanism avoids explicit segmentation of the input image, unlike in (Singh and Markou, 2004).

We also showed that normalised raw image data can be fed directly to the clustering mechanism, adding the extra functionality of image patch reconstruction from the acquired model of normality. This enables the user to perform a visual assessment of which aspects of the environment were actually learnt after training. The need to reduce dimensionality of raw image patches led us to a learning mechanism based on incremental PCA, which was adapted to function as a novelty filter using the magnitude of the residual reconstruction error as a means to highlight novel stimuli.

There were also other contributions to improve performance to the al-

gorithms used throughout this thesis. For example, we implemented a new strategy for the adaptation and habituation of the GWR network's nodes during learning. We also used local maxima interpolation to improve interest point localisation accuracy in the saliency map, as well as a method for feature map normalisation that is more efficient than the original. In the case of the multi-scale Harris detector, we used a very efficient algorithm — the half-octave pyramid (Crowley et al., 2002) — to construct the scale-space structure, combined with bilinear interpolation to search for extrema. Finally, we enhanced colour angular encoding by appending information about intensity spread, making it possible to discriminate shades of grey.

## 7.1 Future Research

The results and conclusions drawn from the experiments in visual novelty detection reported in this thesis open a series of avenues for future investigations and improvements.

It would be interesting to conduct more experiments using alternative interest point detectors, especially those which can determine affine transformation parameters for the selected regions of interest. Possible options are the Harris-affine detector (Mikolajczyk and Schmid, 2002, 2004) and the interest point detector developed by Shi and Tomasi (1994). The use of such algorithms are expected to result in image encoding with extra robustness to affine transformations, improving the ability to generalise and reducing the number of stored vectors or nodes by the novelty filter. Experiments are needed to compare performances with the attention mechanisms already studied here to confirm or reject this hypothesis.

There are also some alternative methods of interest for the image encoding stage, which are likely to improve robustness to changes in scale and orientation of visual features. One possibility is the use of space-variant (log-polar) foveation (Balasuriya and Siebert, 2004; Bernardino et al., 2002;

Bolduc and Levine, 1998) and the Fourier-Mellin transform (Bonmassar and Schwartz, 1997; Casasent and Psaltis, 1976; Cavanagh, 1978, 1985; Derrode, 2001; Reddy and Chatterji, 1996) in order to encode visual features.

In case the application of sewer or pipe inspection is to be developed for real, it might be worthwhile building a specific robot for the task. Of particular interest would be the use of an omnidirectional camera similar to the one described in (Sandini et al., 2002), mounted longitudinally to the robot's motion axis in such a way that panoramic images from the entire cross-section of the pipe could be taken. This camera configuration would minimise the need of affine-invariant image representations because the resulting images from the pipe walls would be practically planar. However, the use of an interest point detector and automatic scale selection for the regions of interest would still be of utmost importance to localise relevant visual features within the image frame.

For applications that demand a systematic exploration of complex large-scale environments, such as a whole floor in a building, the integration of the proposed visual novelty detection framework with the environment exploration scheme developed in (Prestes e Silva Jr. et al., 2002, 2004) is of particular interest. This approach uses potential fields to generate a dynamic exploration path that systematically covers the entire free area of the robot's environment, while generating a grid map of the obstacles that are present. Later on, the generated grid map can be used to produce arbitrary inspection paths or even paths towards specific goals.

If a novelty detection algorithm is used to learn and associate the local visual appearance of the environment to the grids of the environmental map, it is possible to determine novelty not only in terms of uncommon features that may appear in the environment, but also to establish if known features appear in unusual locations. A potential application of such an ability is the automated organisation of a room, in which an autonomous mobile robot

would be able to identify which objects are not in the places they should be and take actions to correct the situation, "tidying up" the environment.

Concerning the unsupervised clustering mechanisms that can be used for novelty detection, it would be interesting to combine the ability of the GWR neural network to build topological maps of the input space with the learning mechanism of the incremental PCA algorithm. This would result in embedded dimensionality reduction within the GWR network with the use of hyper-ellipsoids (Mahalanobis distance) rather than hyper-spheres (Euclidean distance) as clusters in input space, possibly improving the network's learning and reconstruction capabilities. Furthermore, the use of several local PCA clusters in input space is likely to give better overall results than a single global PCA model.

Also of research interest is the use of robust incremental PCA in order to provide the visual learning system with tolerance to partial occlusions, as suggested by Skočaj and Leonardis (2003). Merging and splitting PCA clusters (Hall et al., 2000) in order to improve the performance of the novelty filter is also an interesting research avenue. Finally, unsupervised learning using Independent Component Analysis (ICA) (Bell and Sejnowski, 1996; Karhunen, 1996; van Hateren and Ruderman, 1998; Vicente et al., 2004) and a recently developed mechanism called on-line subtractive clustering (Angelov, 2004; Angelov and Filev, 2004) are very attractive for further investigations in novelty detection and incremental learning using vision.

# Bibliography


P. Angelov. An approach for fuzzy rule-base adaptation using on-line clustering. *International Journal of Approximate Reasoning*, 35:275–289, 2004.

P. P. Angelov and D. P. Filev. An approach to online identification of Takagi-Sugeno fuzzy models. *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics*, 34(1):484–498, 2004.

M. Artač, M. Jogan, and A. Leonardis. Incremental PCA for on-line visual learning and recognition. In *Proceedings of the 16th International Conference on Pattern Recognition (ICPR'2002)*, volume 3, pages 781–784, Quebec, Canada, 2002.

L. S. Balasuriya and J. P. Siebert. Saccade generation for a space-variant artificial retina. In *Early Cognitive Vision Workshop*, Isle of Skye, UK, 2004.

A. J. Bell and T. J. Sejnowski. Edges are the 'independent components' of natural scenes. In *Advances in Neural Information Processing Systems (NIPS)*, volume 9, pages 831–837, Denver, CO, 1996.

A. Bernardino, J. Santos-Victor, and G. Sandini. Model-based attention fixation using log-polar images. In V. Cantoni, A. Petrosino, and M. Marinaro, editors, *Visual Attention Mechanisms*. Plenum Press, New York, NY, 2002.

M. Bolduc and M. D. Levine. A review of biologically motivated space-variant data reduction models for robotic vision. *Computer Vision and Image Understanding*, 69(2):170–184, 1998.

G. Bonmassar and E. L. Schwartz. Space-variant fourier analysis: The exponential chirp transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(10):1080–1089, 1997.

N. Boujemaa, J. Fauqueur, M. Ferecatu, F. Fleuret, V. Gouet, B. Le Saux, and H. Sahbi. IKONA: Interactive specific and generic image retrieval. In *Proceedings of the International Workshop on Multimedia Content-Based Indexing and Retrieval (MMCBIR'2001)*, Rocquencourt, France, 2001.

P. J. Burt and E. H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, April 1983.

D. Casasent and D. Psaltis. Position, rotation, and scale invariant optical correlation. *Applied Optics*, 15(7):1795–1799, 1976.

P. Cavanagh. Size and position invariance in the visual system. *Perception*, 7:167–177, 1978.

P. Cavanagh. Local log polar frequency analysis in the striate cortex as a basis for size and orientation invariance. In D. Rose and V. G. Dobson, editors, *Models of the Visual Cortex*, pages 85–95. Wiley, New York, NY, 1985.

S. Chandrasekaran, B. S. Manjunath, Y. F. Wang, J. Winkler, and H. Zhang. An eigenspace update algorithm for image analysis. *Graphical Models and Image Processing*, 59(5):321–332, 1997.

P. Chang and J. Krumm. Object recognition with color cooccurrence histograms. In *Proceedings of the 1999 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'99)*, Fort Collins, CO, 1999.

P. A. Crook and G. Hayes. A robot implementation of a biologically inspired method for novelty detection. In *Towards Intelligent Mobile Robots (TIMR'01)*, 2001.

P. A. Crook, S. Marsland, G. Hayes, and U. Nehmzow. A tale of two filters — on-line novelty detection. In *Proceedings of the 2002 IEEE International Conference on Robotics and Automation (ICRA'2002)*, pages 3894–3899, Washington, DC, 2002.

J. L. Crowley, O. Riff, and J. Piater. Fast computation of characteristic scale using a half octave pyramid. In *Proceedings of the International Workshop on Cognitive Computing (CogVis'2002)*, Zurich, Switzerland, 2002.

R. B. Davies. The Newmat C++ Matrix Library. World Wide Web: http://www.robertnz.net/nm10.htm, 2002. Retrieved December 30, 2004.

S. Derrode. Robust and efficient Fourier-Mellin transform approximations for gray-level image reconstruction and complete invariant description. *Computer Vision and Image Understanding*, 83(1):57–78, 2001.

N. Dhavale and L. Itti. Saliency-based multi-foveated MPEG compression. In *Proceedings of the 7th IEEE International Symposium on Signal Processing and its Applications*, Paris, France, 2003.

C. P. Diehl and J. B. Hampshire II. Real-time object classification and novelty detection for collaborative video surveillance. In *Proceedings of the 2002 International Joint Conference on Neural Networks (IJCNN'02)*, pages 2620–2625, 2002.

B. Eckel. *Thinking in C++*. Prentice-Hall, New Jersey, NJ, 2000.

W. D. Ferreira and D. L. Borges. Detecting and ranking saliency for scene description. In *Proceedings of the 9th Iberoamerican Congress on Pattern Recognition*, pages 76–83, 2004.

G. D. Finlayson, S. S. Chatterjee, and B. V. Funt. Color angular indexing. In *Proceedings of the 4th European Conference in Computer Vision (ECCV'96)*, pages 16–27, Cambridge, UK, 1996.

G. D. Finlayson, B. Schiele, and J. L. Crowley. Comprehensive colour image normalization. In *Proceedings of the 5th European Conference on Computer Vision (ECCV'98)*, pages 475–490, Freiburg, Germany, 1998.

B. V. Funt and G. D. Finlayson. Color constant color indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):522–529, 1995.

T. Gevers and A. W. M. Smeulders. Object recognition based on photometric color invariants. In *Proceedings of the 10th Scandinavian Conference on Image Analysis (SCIA'97)*, Lappeenranta, Finland, 1997.

J. Gonzales-Barbosa and S. Lacroix. Rover localization in natural environments by indexing panoramic images. In *Proceedings of the 2002 IEEE International Conference on Robotics and Automation (ICRA'2002)*, pages 1365–1370, Washington, DC, 2002.

H. Greenspan, S. Belongie, R. Goodman, P. Perona, S. Rakshit, and C. H. Anderson. Overcomplete steerable pyramid filters and rotation invariance. In *Proceedings of the 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'94)*, pages 222–228, 1994.

P. M. Hall, D. Marshall, and R. R. Martin. Incremental eigenanalysis for classification. In *Proceedings of the 9th British Machine Vision Conference (BMVC'98)*, pages 286–295, 1998.

P. M. Hall, D. Marshall, and R. R. Martin. Merging and splitting eigenspace models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):1042–1049, 2000.

C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.

G. Healey and D. Slater. Global color constancy: Recognition of objects by use of illumination-invariant properties of color distributions. *Journal of the Optical Society of America A*, 11(11):3003–3010, November 1994.

G. Healey and L. Wang. Illumination-invariant recognition of texture in color images. *Journal of the Optical Society of America A*, 12(9):1877–1883, September 1995.

J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, April 1982.

C. S. Horstmann. *Mastering C++: An Introduction to C++ and Object-Oriented Programming for C and Pascal Programmers, 2nd ed.* Wiley, New York, NY, 1996.

T. Hosoya, S. A. Baccus, and M. Meister. Dynamic predictive coding by the retina. *Nature*, 436(7):71–77, July 2005.

iRobot Corporation. *Magellan Pro Compact Mobile Robot User's Guide, Rev. 4.* iRobot Corporation, Jaffrey, NH, 2001.

iRobot Corporation. *Mobility Robot Integration Software User's Guide, Rev. 5.* iRobot Corporation, Jaffrey, NH, 2002.

L. Itti and C. Koch. A comparison of feature combination strategies for saliency-based visual attention systems. In *Proceedings of SPIE: Human Vision and Electronic Imaging IV (HVEI'99)*, pages 473–482, San Jose, CA, 1999.

L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

T. Kadir and M. Brady. Scale saliency: A novel approach to salient feature and scale selection. In *Proceedings of the International Conference on Visual Information Engineering (VIE'2003)*, pages 25–28, 2003.

T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *Computer Vision - ECCV 2004: 8th European Conference on Computer Vision (ECCV'2004)*, volume 1, pages 228–241, Prague, Czech Republic, 2004.

J. Karhunen. Neural approaches to independent component analysis and source separation. In *Proceedings of the 4th European Symposium on Artificial Neural Networks (ESANN96)*, pages 249–266, Bruges, Belgium, 1996.

T. Kohonen. *Self-organization and Associative Memory.* Springer-Verlag, New York, NY, 1984.

F. Linåker and L. Niklasson. Time series segmentation using an adaptive resource allocating vector quantization network based on change detection. In *Proceedings of the 2000 International Joint Conference on Neural Networks (IJCNN'2000)*, pages 323–328, 2000.

F. Linåker and L. Niklasson. Environment identification by alignment of abstract sensory flow representations. In N. Mastorakis, V. Mladenov, B. Suter, and L. Wang, editors, *Advances in Neural Networks and Applications*, pages 229–234. WSES Press, 2001.

T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):194–203, 1998.

E. Loupias, N. Sebe, S. Bres, and J. M. Jolion. Wavelet-based salient points for image retrieval. In *Proceedings of the 2000 IEEE International Conference on Image Processing*, volume 2, pages 518–521, 2000.

D. Lowe. Radial basis function networks. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge, MA, 1998.

D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV'99)*, pages 1150–1157, 1999.

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

M. Markou and S. Singh. Novelty detection: A review - part 1: Statistical approaches. *Signal Processing*, 83:2481–2497, 2003a.

M. Markou and S. Singh. Novelty detection: A review - part 2: Neural network based approaches. *Signal Processing*, 83:2499–2521, 2003b.

S. Marsland. Novelty detection in learning systems. *Neural Computing Surveys*, 3:157–195, 2003.

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

S. Marsland, U. Nehmzow, and J. Shapiro. Vision-based environmental novelty detection on a mobile robot. In *Proceedings of the International Conference on Neural Information Processing (ICONIP'01)*, Shanghai, China, 2001.

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

S. Marsland, J. Shapiro, and U. Nehmzow. A self-organising network that grows when required. *Neural Networks*, 15(8-9):1041–1058, 2002b.

B. W. Mel. SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation*, 9(4):777–804, 1997.

K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *International Conference on Computer Vision*, volume 1, pages 525–531, 2001.

K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Computer Vision - ECCV 2002: 7th European Conference on Computer Vision (ECCV'2002)*, volume 1, pages 128–142, Copenhagen, Denmark, 2002.

K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

H. Murakami and B. V. K. Vijaya Kumar. Efficient calculation of primary images from a set of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4(5):511–515, 1982.

V. Navalpakkam and L. Itti. Sharing resources: Buy attention, get recognition. In *Proceedings of the International Workshop on Attention and Performance in Computer Vision (WAPCV'03)*, Graz, Austria, 2003.

U. Nehmzow. *Mobile Robotics: A Practical Introduction, 2nd ed.* Springer-Verlag, London, UK, 2003.

U. Nehmzow and H. Vieira Neto. Locating objects visually using opposing-colour-channel coding. In *Towards Intelligent Mobile Robots (TIMR'03)*, Bristol, UK, 2003.

U. Nehmzow and H. Vieira Neto. Novelty-based visual inspection using mobile robots. In *Towards Autonomous Robotic Systems (TAROS 2004)*, Colchester, UK, 2004.

E. Prestes e Silva Jr., P. M. Engel, M. Trevisan, and M. A. P. Idiart. Exploration method using harmonic functions. *Robotics and Autonomous Systems*, 40(1):25–42, 2002.

E. Prestes e Silva Jr., M. A. P. Idiart, M. Trevisan, and P. M. Engel. Autonomous learning architecture for environmental mapping. *Journal of Intelligent and Robotic Systems*, 39:243–263, 2004.

B. S. Reddy and B. N. Chatterji. An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, 5(8):1266–1271, 1996.

L. Sachs. *Angewandte Statistik: Anwendung statistischer Methoden.* Springer Verlag, Berlin, Germany, 2004.

G. Sandini, J. Santos-Victor, T. Pajdla, and F. Berton. OMNIVIEWS: Direct omnidirectional imaging based on a retina-like sensor. In *Proceedings of the First IEEE International Conference on Sensors (IEEE Sensors 2002)*, 2002.

B. Schiele and J. L. Crowley. Probabilistic object recognition using multidimensional receptive field histograms. In *Proceedings of the 13th International Conference on Pattern Recognition (ICPR'96)*, volume B, pages 50–54, Vienna, Austria, 1996.

B. Schiele and J. L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, 2000.

C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5): 530–535, 1997.

J. Shi and C. Tomasi. Good features to track. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, pages 593–600, Seattle, WA, 1994.

E. P. Simoncelli and M. Werman. The steerable pyramid: a flexible architecture for multi-scale derivative computation. In *Proceedings of the 2nd Annual IEEE International Conference on Image Processing*, Washington, DC, 1995.

S. Singh and M. Markou. An approach to novelty detection applied to the classification of image regions. *IEEE Transactions on Knowledge and Data Engineering*, 16(4):396–407, 2004.

D. Skočaj and A. Leonardis. Weighted and robust incremental method for subspace learning. In *Proceedings of the 9th International Conference on Computer Vision (ICCV'03)*, pages 1494–1501, 2003.

M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(11):11–32, 1991.

L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady. Novelty detection for the identification of masses in mammograms. In *Proceedings of the 4th IEE International Conference on Artificial Neural Networks (ICANN'95)*, pages 442–447, 1995.

O. Taylor and J. McIntyre. Adaptive local fusion systems for novelty detection and diagnostics in condition monitoring. In *Proceedings of the SPIE International Symposium on Aerospace/Defense Sensing*, 1998.

A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

J. H. van Hateren and D. L. Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London A*, 265:2315–2320, 1998.

M. A. Vicente, F. C., O. Reinoso, and C. Pérez. Robust object detection in complex backgrounds using ICA compression. In *Proceedings of the 12th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG'2004)*, 2004.

H. Vieira Neto and U. Nehmzow. Object localisation and tracking through subsymbolic classification. In *Proceedings of the AISB'03 Symposium on*

*Biologically-Inspired Machine Vision, Theory and Application*, pages 85–90, Aberystwyth, UK, 2003.

H. Vieira Neto and U. Nehmzow. Visual novelty detection for inspection tasks using mobile robots. In *Proceedings of the 8th Brazilian Symposium on Neural Networks (SBRN 2004)*, São Luís, Brazil, 2004.

H. Vieira Neto and U. Nehmzow. Incremental PCA: An alternative approach for novelty detection. In *Towards Autonomous Robotic Systems (TAROS 2005)*, pages 227–233, London, UK, 2005a.

H. Vieira Neto and U. Nehmzow. Automated exploration and inspection: Comparing two visual novelty detectors. *International Journal of Advanced Robotic Systems*, 2(4):355–362, 2005b.

I. T. Young, L. J. van Vliet, and M. van Ginkel. Recursive Gabor filtering. *IEEE Transactions on Signal Processing*, 50(11):2798–2805, 2000.

A. Ypma and R. P. W. Duin. Novelty detection using self-organizing maps. In *Proceedings of the International Conference on Neural Information Processing and Intelligent Information Systems*, Dunedin, New Zealand, 1997.

# Publications

The following publications resulted from the work reported in this thesis:

### Conference papers

1. Hugo Vieira Neto and Ulrich Nehmzow. Object localisation and tracking through subsymbolic classification. In *Proceedings of the AISB'03 Symposium on Biologically-Inspired Machine Vision, Theory and Application*, pages 85–90, Aberystwyth, UK, 2003.

2. Ulrich Nehmzow and Hugo Vieira Neto. Locating objects visually using opposing-colour-channel coding. In *Towards Intelligent Mobile Robots (TIMR'03)*, Bristol, UK, 2003.

3. Hugo Vieira Neto and Ulrich Nehmzow. Visual novelty detection for inspection tasks using mobile robots. In *Proceedings of the 8th Brazilian Symposium on Neural Networks (SBRN 2004)*, São Luís, Brazil, 2004.

4. Ulrich Nehmzow and Hugo Vieira Neto. Novelty-based visual inspection using mobile robots. In *Towards Autonomous Robotic Systems (TAROS 2004)*, Colchester, UK, 2004.

5. Hugo Vieira Neto and Ulrich Nehmzow. Incremental PCA: An alternative approach for novelty detection. In *Towards Autonomous Robotic Systems (TAROS 2005)*, pages 227–233, London, UK, 2005.

6. Ulrich Nehmzow and Hugo Vieira Neto. Visual attention and novelty detection: Experiments with automatic scale selection. Submitted to *Towards Autonomous Robotic Systems (TAROS 2006)*.

### Journal papers

1. Hugo Vieira Neto and Ulrich Nehmzow. Automated exploration and inspection: Comparing two visual novelty detectors. *International Journal of Autonomous and Robotic Systems*, 2(4):355–362, 2005.

2. Hugo Vieira Neto and Ulrich Nehmzow. Real-time automated visual inspection using mobile robots. Submitted to *Journal of Intelligent and Robotic Systems*.