Visual Novelty Detection with Automatic Scale Selection

Hugo Vieira Neto [a,*] and Ulrich Nehmzow [b]

[a] *Department of Electronics, Federal University of Technology - Paraná*
Avenida Sete de Setembro 3165, Curitiba-PR 80230-901, Brazil
[b] *Department of Computer Science, University of Essex*
Wivenhoe Park, Colchester CO4 3SQ, United Kingdom

Abstract

This paper presents experiments with an autonomous inspection robot, whose task was to highlight novel features in its environment from camera images.

The experiments used two different attention mechanisms — saliency map and multi-scale Harris detector — and two different novelty detection mechanisms — Grow-When-Required (GWR) neural network and an incremental Principal Component Analysis (PCA). For all mechanisms we compared fixed-scale image encoding with automatically scaled image patches.

Results show that automatic scale selection provides a more efficient representation of the visual input space, but that performance is generally better using a fixed-scale image encoding.

Key words:
automated inspection, scale invariance, on-line novelty detection

1. Introduction

The ability to identify perceptions that were never experienced before — novelty detection — is an attractive component of "intelligent" robot control for a number of reasons. First, the ability to differentiate between common and rare perceptions is essential component for mobile robots aiming at true autonomy, adaptability to new situations and continuous operation. Second, from an operational point of view, the robot's limited computational resources can be used more efficiently by selecting the aspects of the surroundings which are relevant to the task in hand or uncommon aspects which deserve further analysis. Third, by using a novelty detection mechanism, previously unknown aspects of the environment can be incrementally learnt by the robot without supervision. Finally, novelty detection is a core competence in industrial applications of autonomous mobile robots, for instance in inspection, surveillance or fault detection tasks. The research presented here — while currently focusing on research aspects — will ultimately lead to autonomous inspection robots that can relieve human operators from tedious and therefore fatiguing aspects of inspection work (scrutinising repetitive, normal data) and allow them to concentrate on unusual signals, highlighted by the inspection robot.

Obviously, relevant features in the environment need to be sensed and discriminated, otherwise it would be impossible for the agent to respond appropriately. Therefore, the sensor modality used to generate the perceptual input plays an important role in the agent's performance for a given task or behaviour. Among the various sensors commonly available to a mobile robot, vision allows measurement and estimation of several environmental features and provides high resolution readings in two dimensions, making the detection of small details of the environment more likely.

Our work investigates novelty detection mechanisms using vision as perceptual input, with potential applications in automated inspection. In order to deal with the large amount of data provided by a camera, we use an attention model to select raw image patches from the input image frame. These image patches are then normalised to unit-length vectors and fed to a novelty filter that indicates the presence or absence of novelty [1,2]. This approach is summarised in Fig. 1.

The role of the attention model The attention model plays an important role in the overall performance of our visual novelty detection framework, allowing the localisation of

[*] Corresponding author.
Email addresses: hvieir@utfpr.edu.br (Hugo Vieira Neto), udfn@essex.ac.uk (Ulrich Nehmzow).

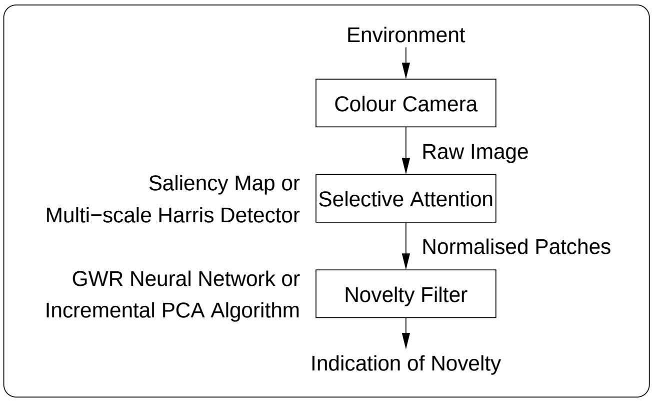


Fig. 1. Visual novelty detection using raw image patches: an attention model selects candidate regions, which are then normalised and fed to a novelty filter.

where novel visual features are within an input image frame and reducing the dimensionality of input vectors through the use of *local* image patches rather than *the whole* image frame. Moreover, because visual input is acquired from a moving platform and therefore subject to several geometrical transformations, the use of visual attention intrinsically deals with translations of features within the image frame by centring image patches on stable salient locations. However, other alterations in appearance such as changes in scale, rotations or affine-transformations need more elaborated image encoding to be dealt with efficiently.

The role of automatic scale selection Our hypothesis is that an image encoding method that is robust to changes in scale would improve the novelty filter's ability to generalise and reduce the number of acquired concepts by the learning mechanism in use — we therefore investigate here how image patch size can be determined automatically. In previous work we achieved generalisation according to scale by acquiring multiple image patches in different scales for the salient visual features found in the environment [1]. We used raw image data to allow image patch reconstruction and provide visual feedback of which aspects of the environment were actually learnt by the robot.

In this work, we exploit the characteristic scale property [3] present in multi-scale attention mechanisms to determine the size of interest regions (image patches) automatically. We compare results obtained during a novelty detection task using the saliency map [4] and the multi-scale Harris detector [5] as attention mechanisms; the GWR neural network [6] and the incremental PCA algorithm [7] were compared as novelty filters.

2. Novelty Detection

The objective of novelty detection is to highlight any *previously unknown* feature. This differs from pattern recognition tasks, in which the features of interest are already known beforehand.

Therefore, we acquire a model of *normality* from the environment using robot learning, and then use this model to filter out abnormal perceptions. In this work we focus on on-line unsupervised learning mechanisms based either on neural networks or statistical approaches.

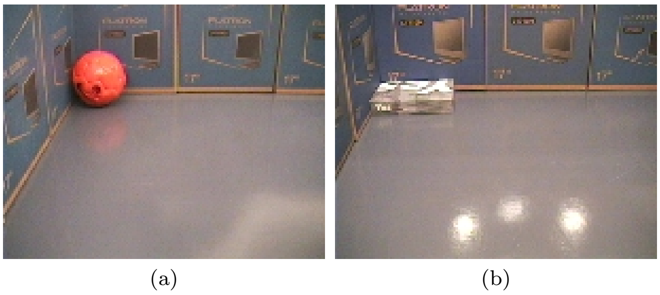
Initially we use the GWR network, which was originally designed for the purpose of on-line novelty detection, as novelty filter. This neural network combines a self-organising clustering mechanism and a model of habituation to decide if a given input vector is novel or not [6].

In a second experiment we use an alternative method for novelty detection based on the incremental PCA algorithm introduced by Artac *et al* [7]. In this approach, the magnitude of the residual vector (the RMS error between the original input data and its reconstruction from the current eigenspace projection) is used as a means to determine novelty. Incremental PCA offers the advantage of intrinsically reducing input dimensionality, allowing optimal reconstruction (minimal squared error). Implementation details of both novelty detection methods used in this work are given in [1].

2.1. *Experimental Setup*

Our experiments were carried out using a *Magellan Pro* robot operating in a square arena (2.56 × 2.56m) made from cardboard boxes. The robot used a simple force-field obstacle avoidance behaviour to navigate slowly around the arena while acquiring images at one frame per second. During an initial learning phase the robot acquired a model of normality from its environment. The learning phase consisted of five loops around the empty arena, resulting in the acquisition of 225 image frames.

Once the model of normality was obtained, a novel object (either an orange football or a grey box) was placed in the arena and the trained robot was used to inspect the environment again. The inspection phase also comprised five loops around the arena containing a novel object, resulting again in 225 acquired image frames for each case. Examples of acquired images containing the novel objects inside the arena are given in Fig. 2.



(a) (b)

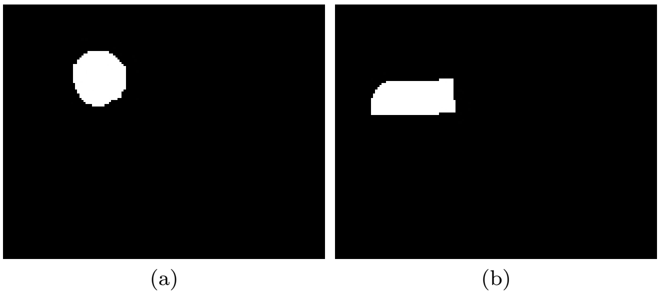
Fig. 2. Input images containing novel objects inside the arena: (a) orange football; (b) grey box.

The expected outcome of the experiments was that the novelty filters would highlight the location of novel stimuli during inspection and ignore visual features that were previously learnt.

2.2. *Quantitative Assessment of Results*

To ensure a fair comparison of all algorithms used, all images obtained in the arena were stored and algorithms were tested off-line, using identical images. The novelty filters used here, however, have the ability to process data on-line in real time, and in actual robot applications would be used to identify novelty while operating in the environment.

After acquiring the images from the environment, ground truth data was generated in the form of a binary image for each image frame where novel objects were present. The pixels corresponding to novel features were manually highlighted in these ground truth templates, as Fig. 3 shows.



(a) (b)

Fig. 3. Ground truth templates corresponding to the input images in Fig. 2: (a) orange football; (b) grey box.

Using the ground truth information, contingency tables were built relating the system response to the actual novelty status, as shown in Table 1. For the novelty status of a given region of the input image to be considered as "novelty present", it had to have at least 10% of highlighted pixels in the corresponding region of the respective ground truth template.

Table 1
Example contingency table for the quantitative assessment of novelty filters.

	Novelty Detected	Novelty Not Detected
Novelty Present	A	B
Novelty Not Present	C	D

Statistical significance of the association between actual novelty status (ground truth) and the novelty filter response was established using a χ^2 analysis of the contingency table [8,9]. The strength of this association was then quantified through Cramer's V $(0 \leq V \leq 1)$ and the uncertainty coefficient U $(0 \leq U \leq 1)$. Smaller values for these statistics indicate weaker associations [9,10].

A further statistic used in this paper is the κ index of agreement, which is computed as follows [11]:

$$\kappa = \frac{2(AD - BC)}{(A+C)(C+D) + (A+B)(B+D)}, \qquad (1)$$

where A, B, C and D are the entries in the contingency table (see Table 1).

This statistic is used to assess the agreement between ground truth and novelty filter response, in a similar fashion to V and U. However, κ has the advantage of having an established semantic meaning associated with some intervals [11], as Table 2 shows.

Table 2
κ 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 κ statistic may yield negative values $(-1 \leq \kappa \leq 1)$. If κ is negative, the level of *disagreement* between system response and manually generated ground truth can be assessed.

3. Models of Visual Attention

3.1. *Saliency Map*

In previous work [1,12] we reported experiments using the saliency map [4] as a mechanism of visual attention using a fixed number of salient points. The saliency map is inspired by the early primate visual system and consists of computing multi-scale feature maps that allow the detection of local changes in intensity, colour and orientation in different scales.

The feature maps are obtained from image pyramids computed from the original input image. In our implementation, Gaussian and oriented Gabor pyramids with eight scales were built, as described in [4]. Across-scale differences were computed between finer and coarser scales from the pyramids to yield the feature maps, which were combined in conspicuity maps for intensity, opponent colours and orientation. A normalisation operator was used in order to combine these conspicuity maps with different dynamic ranges into a single saliency map, giving more weight to unusual features in the input image frame. Full implementation details are available in [10].

The highest value within the saliency map needs to be found 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. Salient locations were determined by a search for local maxima whose values were above the average saliency value of the map. The determined coordinates and their neighbours were then used to interpolate the salient location with sub-pixel accuracy using a Taylor expansion up to the second derivative:

$$\hat{x} = -\frac{S_x}{S_{xx}}, \tag{2}$$

$$\hat{y} = -\frac{S_y}{S_{yy}}, \tag{3}$$

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 2 and 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 because the saliency function is reasonably smooth.

3.2. *Multi-scale Harris Detector*

We also implemented the multi-scale Harris detector [5] as an alternative interest point selection strategy to the saliency map. This algorithm basically consists of building an intensity Laplacian pyramid from the input image and then searching it for extrema. Interest points correspond to extrema due to their stability in both space and scale [13]. We used the fast and efficient algorithm proposed in [14] to build Difference-of-Gaussian (Laplacian) image pyramids by successive Gaussian filtering, sub-sampling and subtraction. In our implementation we used Laplacian pyramids with 12 scale levels.

After the Laplacian pyramid is built for an input image, search for extrema in scale-space is performed. Each pixel in the pyramid is compared to its eight neighbours in the same level and its 18 neighbours in the levels above and below. The location of extrema is interpolated using equations 2 and 3 for better accuracy. Extrema corresponding to locations with low contrast (less than 2.5% of the maximum value in the image) were rejected. Also, because the Difference-of-Gaussian function has strong responses along edges even if localisation is poorly defined and unstable due to noise, we rejected locations with a principal curvature ratio $r < 4$ in order to achieve better stability (poorly defined extrema have a large principal curvature across the edge but a small curvature in its perpendicular direction) [13]. Further details are given in [10].

4. Experiments with Fixed Scale

To compare performances of different strategies to select interest points, we conducted experiments using unit-length normalised raw image patches in the image encoding stage, the same approach followed in [1]. The main reason to use raw image patches was to allow reconstruction from the acquired model of normality. However, by using this image encoding approach, the overall performance of the visual novelty detection system is sensitive to patch misalignment, which obviously depends on the accuracy and stability of the attention mechanism being used. Therefore, an attention mechanism that provides better interest point stability and accuracy is expected to also provide better overall performance when using raw image patches.

First set of experiments For the first experiments a fixed scale size of 25 × 25 pixels was used for the image patches, which were centred at the locations selected either by the saliency map or by the multi-scale Harris detector. Both of these approaches automatically decide the number of salient points to be selected within the input image according to the threshold parameters mentioned in section 3.

In order to assess the impact of the attention mechanism on the overall visual novelty detection performance, a GWR network was trained with the normalised raw image patches selected from the empty arena (the activation threshold for the GWR network was $a_T = 0.85$, see [1]). The acquired model of normality of the empty arena was then used to filter out any abnormal perceptions during inspection, which was conducted with the presence of novel objects (an orange football or a grey box) in the arena. The results obtained with each attention mechanism are given in Table 3, including the sizes of the acquired models.

Table 3
Visual novelty detection performance comparison using different interest point selection methods (fixed scale) and the GWR network. The larger the V, U and κ values, the stronger the agreement between novelty postulated by the filter and manually determined ground truth.

	Model Size	Orange Ball	Grey Box
Saliency Map	23 nodes	$V = 0.70$ $U = 0.41$ $\kappa = 0.66$	$V = 0.71$ $U = 0.42$ $\kappa = 0.67$
Multi-scale Harris Detector	31 nodes	$V = 0.98$ $U = 0.92$ $\kappa = 0.98$	$V = 0.80$ $U = 0.58$ $\kappa = 0.78$

All experiments resulted in statistically significant correlation between novelty ground truth and the classification made by the GWR network (χ^2 analysis, $p \leq 0.01$). It can be noticed in Table 3 that the performance of the saliency map is consistent for both novel objects and corresponds to a "strong agreement" between system response and actual novelty status. The multi-scale Harris detector performs better, resulting in "almost complete agreement" between the novelty filter response and ground truth for the orange ball. It should be noted that for the parameters chosen the multi-scale Harris detector selects a larger number of interest points than the saliency map and therefore increases the chances of finding novelties.

Figure 4 shows the reconstructed images from the trained GWR networks when using either of the interest point detectors. From these reconstructions one can notice that the GWR network learned the visual features corresponding mostly to drawings and inscriptions on the cardboard boxes, edges between adjacent boxes and edges between the boxes and the floor. The number of interest points selected by each attention mechanism is reflected in the number of acquired concepts — the use of the saliency map resulted in a trained GWR network with 23 nodes, while the use of the multi-scale Harris detector resulted in a GWR network with 31 nodes. Notice that in both models of normality there are several concepts corresponding to scaled versions of the same visual features.

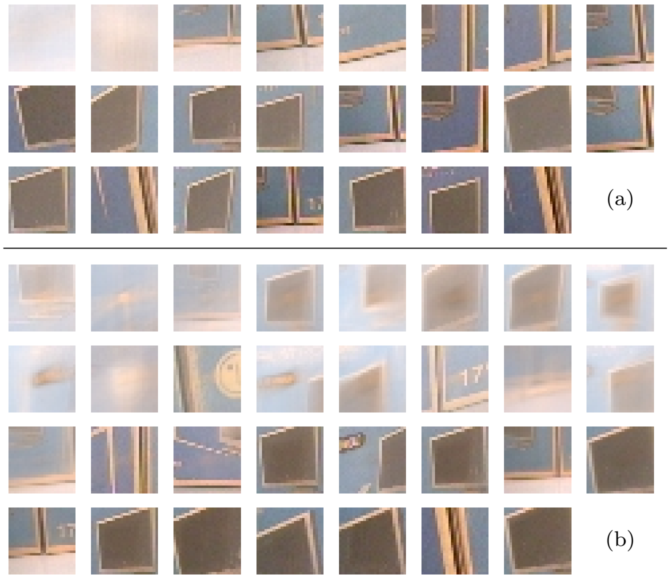


Fig. 4. Image patches (fixed scale) acquired using the GWR network: (a) interpolated saliency map; and (b) multi-scale Harris detector.

The same experiments were repeated using the incremental PCA algorithm as novelty filter (the residual error threshold for the incremental PCA algorithm was $r_T = 0.25$, see [1]). Table 4 shows the results obtained.

Again, all results showed statistically significant association between system response and actual novelty status (χ^2 analysis, $p \leq 0.01$) when using incremental PCA as novelty filter. Table 4 shows that the strength of the association between system response and ground truth was in the same order of magnitude as when using the GWR network as novelty filter (see Table. 3).

The reconstructed images from the acquired PCA models using the saliency map or the multi-scale Harris detector are shown in Fig. 5, where it can be noticed that the concepts acquired by both PCA models roughly correspond to the same visual features acquired by the GWR network. Once again, multiple image patches corresponding to scaled versions of the same visual features are present in both models of normality.

Table 4
Visual novelty detection performance comparison using different interest point selection methods (fixed scale) and incremental PCA. The larger the V, U and κ values, the stronger the agreement between novelty postulated by the filter and manually determined ground truth.

	Model Size	Orange Ball	Grey Box
Saliency Map	28 vectors (26 dim.)	$V = 0.76$ $U = 0.50$ $\kappa = 0.74$	$V = 0.78$ $U = 0.53$ $\kappa = 0.76$
Multi-scale Harris Detector	37 vectors (36 dim.)	$V = 0.97$ $U = 0.89$ $\kappa = 0.97$	$V = 0.78$ $U = 0.58$ $\kappa = 0.76$

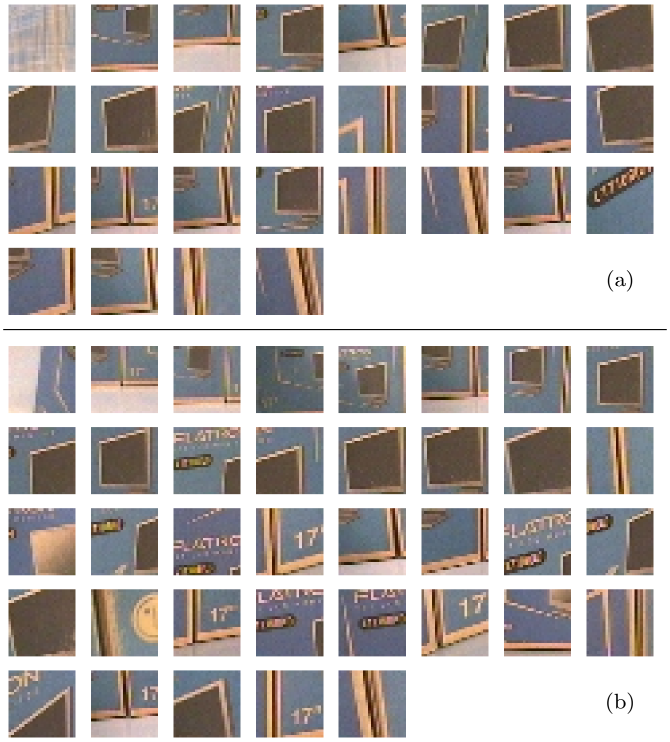


Fig. 5. Image patches (fixed scale) acquired using incremental PCA: (a) interpolated saliency map; and (b) multi-scale Harris detector.

5. Experiments with Automatic Scale

On a moving mobile robot, visual features are subject to several geometric transformations as a result of robot motion. The use of attention mechanisms provides robustness to translations of visual features by selecting salient characteristic locations within the image frame. Both attention mechanisms being investigated in this paper rely on a multi-scale pyramidal (also known as scale-space) representation, which provides them with a good degree of stability when selecting salient locations, regardless of translations or changes in scale.

Changes in scale are evident when the robot approaches objects. In our experiments using image patches with *fixed* size, we achieved generalisation according to scale by learning multiple versions of salient visual features in different scales. If the image encoding stage is made invariant to changes in scale, this would obviously improve the overall system ability to generalise and reduce the amount of acquired concepts in the model of normality of the environment.

Determining the 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 [3]. The Laplacian jet of a given pixel is the function across the levels of a Difference-of-Gaussian image pyramid at the coordinates of the given 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 paper already make use of Laplacian (Difference-of-Gaussian) 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 successfully used in [13] and [14] to determine the region of interest surrounding visual features.

Once the location of an interest point is found, the Laplacian jet profile at that location needs to be searched for an extremum. A more precise location in scale is also determined by interpolation using a second order Taylor expansion:

$$\hat{s} = -\frac{L_s}{L_{ss}}, \tag{4}$$

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 then added to the extremum level in order to determine scale with better accuracy. According to [14], the radius r_{roi} of the region of interest can be computed from the interpolated pyramid level by using the equation:

$$r_{roi} = k_s \times b^{(s+\hat{s})}, \tag{5}$$

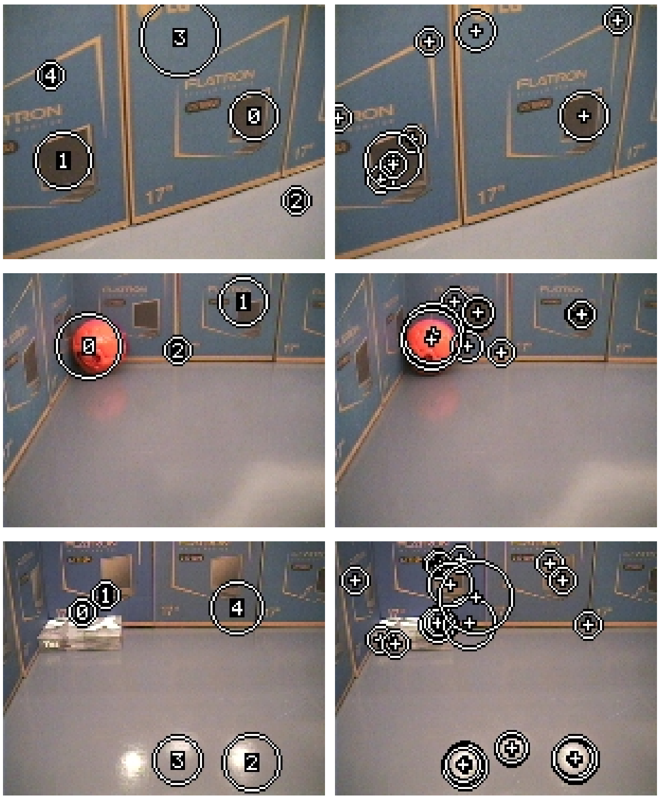
where the constant $k_s = 1.6$ is an empirical correction factor for the scale, which is given by a geometric progression with base $b = \sqrt{2}$.

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 (Laplacian pyramid) with 12 levels, which provides sufficient scale resolution. However, the intensity Laplacian pyramid of the saliency map is not built using the same algorithm. Therefore an additional Laplacian pyramid was 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 1.5 times the radius computed with Equation 5, in order to guarantee that edges would be present in the image patches. The final radius was limited to a minimum of six pixels and a maximum of 24 pixels, and the size of the patch was computed as follows:

$$p = 2 \times \min\{\max\{6, 1.5 \times r_{roi}\}, 24\} + 1. \tag{6}$$

This results in the selection of square image patches centred around the interest points ranging from 13 × 13 to 49 × 49 pixels in size. Figure 6 shows examples of points selected by the saliency map and the multi-scale Harris detector, and their regions of interest, whose sizes were calculated according to Equation 6.



(a) (b)

Fig. 6. Output images with automatic scale selection: saliency map (left column, a) and multi-scale Harris detector (right column, b). 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.

The circles in Fig. 6 designate the size of the regions of interest according to the automatic scale selection of the corresponding interest points at their centres. There was no novelty detection involved in the generation of these output images, just the use of the attention models with automatic scale selection to determine the size of the regions of interest. In these examples it is possible to notice the preference of both algorithms for interest points on blobs and edges with high curvature, although the saliency map also selects interest points on straight edges.

Second set of experiments In a second set of experiments we investigated how automatic scale selection affects our visual novelty detection mechanism. We expected that smaller models of normality would be acquired than in the first set of experiments, because generalisation according to scale improves through the image encoding mechanism itself, rather than the acquisition of multiple scaled versions of the same features by the learning mechanisms. To obtain input vectors with fixed size for the novelty filters, the image patches selected by the attention models were scaled to a fixed image patch size of 25 × 25 pixels (the original size of image patches when fixed scale was used) through bilinear interpolation, allowing changes in scale from 1:2 to 2:1.

First, we trained a GWR network using images acquired when the robot was exploring the empty arena, as in the previous experiments. The acquired model of normality was then used to filter out abnormal visual features in images acquired during inspection of the arena containing either of two novel objects (again, the orange football or the grey box). Table 5 shows the quantitative results obtained.

Table 5
Performance comparison between different interest point selection methods (automatic scale) using the GWR network. The larger the V, U and κ values, the stronger the agreement between novelty postulated by the filter and manually determined ground truth.

	Model Size	Orange Ball	Grey Box
Saliency Map	17 nodes	$V = 0.74$ $U = 0.46$ $\kappa = 0.71$	$V = 0.54$ $U = 0.28$ $\kappa = 0.54$
Multi-scale Harris Detector	23 nodes	$V = 0.95$ $U = 0.83$ $\kappa = 0.94$	$V = 0.58$ $U = 0.34$ $\kappa = 0.54$

The use of both attention mechanisms resulted in statistically significant association between the GWR network response and ground truth data (χ^2 test, $p \leq 0.01$). The results in Table 5 show that performance of both interest point detectors was worse than that obtained with fixed scale for the case of the grey box ("clear agreement" between novelty filter response and actual novelty status), but was kept at the same level for the orange ball (compare with Table 3). We surmise that this is because the grey box stands out less well from the grey floor than the orange ball, and furthermore has smaller details that attract interest points. Also, the use of bilinear interpolation for scaling causes image patch smoothing, i.e. a low-pass filtering effect, which makes differentiation of image patches using the Euclidean metric (used by the GWR network) more difficult. Figure 7 depicts the image patches reconstructed from the acquired models of normality when using either of the interest point detectors.

The hypothesis that the use of automatic scale selection results in smaller models of normality is confirmed in Fig. 7.

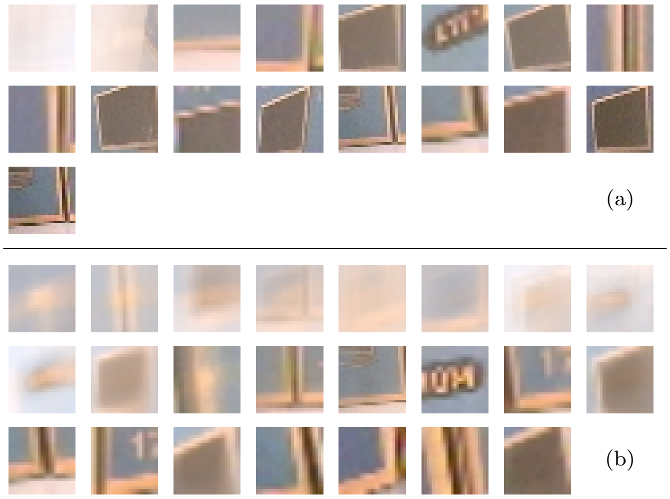


Fig. 7. Image patches (auto scale) acquired using the GWR network: (a) interpolated saliency map; and (b) multi-scale Harris detector. Both models are smaller than the ones acquired using fixed scale.

When using the saliency map with automatic scale selection the number of concepts was reduced from 23 to 17 and when using the multi-scale Harris detector from 31 to 23, a reduction of approximately 26% in both cases. One can notice that in these models there are fewer image patches corresponding to scaled versions of the same visual features (see Fig. 4). The ability to generalise scale results in a reduction in the number of acquired concepts, as predicted.

As in the first set of experiments, the experiments were then repeated using the incremental PCA algorithm as learning mechanism, which is expected to be less sensitive to bilinear interpolation smoothing. A quantitative comparison of the results obtained is in Table 6.

Table 6
Performance comparison between different interest point selection methods (automatic scale) using incremental PCA. The larger the V, U and κ values, the stronger the agreement between novelty postulated by the filter and manually determined ground truth.

	Model Size	Orange Ball	Grey Box
Saliency Map	17 vectors (16 dim.)	$V = 0.91$ $U = 0.76$ $\kappa = 0.91$	$V = 0.37$ $U = 0.18$ $\kappa = 0.35$
Multi-scale Harris Detector	22 vectors (21 dim.)	$V = 0.98$ $U = 0.93$ $\kappa = 0.98$	$V = 0.34$ $U = 0.20$ $\kappa = 0.24$

Despite revealing statistically significant association between system response and ground truth data (χ^2 test, $p \leq 0.01$), the results in Table 6 are much poorer than the results obtained using fixed scale for the case of the grey box ("weak agreement" between system response and ground truth data). For the orange ball, performance of the saliency map was improved and resulted in "almost complete agreement" between novelty filter response and actual novelty status (compare with Table 4). The explanation for this fact is that the details of the grey box correspond to interest points in small scales and with relatively low contrast, making their discrimination from features in larger scales more difficult as a result of bilinear interpolation smoothing.

The reconstructed images from the acquired incremental PCA models using automatic scale selection are shown in Fig. 8, where the fact that the acquired models using automatic scale selection are smaller can be confirmed by comparisons with Fig. 5.

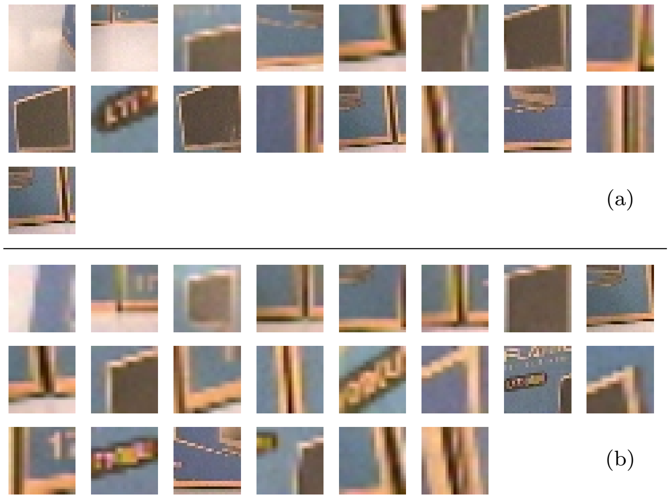


Fig. 8. Image patches (auto scale) acquired using incremental PCA: (a) interpolated saliency map; and (b) multi-scale Harris detector. Both models are smaller than the ones acquired using fixed scale.

Figure 8 shows that the use of automatic scale selection reduced the number of acquired vectors from 28 to 17 when using the saliency map and from 37 to 22 vectors when using the multi-scale Harris detector. This corresponds to a reduction of approximately 40% in acquired concepts. Once again, it can be noticed that there are fewer image patches corresponding to scaled versions of the same visual features in these models (see Fig. 5).

6. Conclusion

This paper addressed the question of how an attention mechanism influences a robot's performance to detect novelty in its environment. In particular, we were interested to find out if automatic scale detection improves performance. We therefore investigated two distinct interest point detection schemes: the saliency map [4] and the multi-scale Harris detector [5]. Both approaches had their localisation accuracy improved through function interpolation using a second order Taylor expansion, as suggested in [13].

Interest point stability The accuracy and stability in interest point selection becomes an important issue when using raw data for image encoding. Accurate localisation reduces errors due to misalignment of image patches during matching, having an impact in the overall performance of the visual novelty filter and also contributing to reduce the size of the model of normality that is learnt from the environment. The use of raw image data allows the reconstruction of visual information from the acquired model of normality, which is essential to understand which aspects of the environment were actually learnt.

Another issue of concern is the robustness to changes in scale of visual features as a result of robot navigation around the environment. In experiments involving image patches with fixed size, generalisation with respect to scale happened through the acquisition of several scaled versions of the same visual features by the learning mechanisms. 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 in [3].

Results The results in Figs. 4, 5, 7 and 8 corroborate our hypothesis, showing that the use of automatic scale selection reduced the size of the acquired models of normality (26% in the case of the GWR network and 40% in the case of incremental PCA). However, performance of the novelty filters for inconspicuous features (the grey box) was generally worse than when using fixed scale image patches, while performance for conspicuous features (the orange ball) was better or at least the same. The conclusion drawn is that when the model size matters one should choose automatic scale selection, otherwise performance is better when using fixed scale. We attribute the deterioration in performance for inconspicuous features to the difficulty in discriminating scaled image patches due to the smoothing effects that arise from bilinear interpolation. Alternative techniques to perform comparisons between image patches with different sizes are currently being investigated.

Quantitative assessment We performed quantitative performance comparisons through contingency table analysis and computation of Cramer's V, uncertainty coefficient U and the κ index of agreement [8,9,11]. The multi-scale Harris detector gave the best results, particularly when using a fixed scale strategy and the GWR network as novelty filter.

Future work Our future research aims at improving performance through the use of affine-invariant interest point detectors [15,16]. Concerning automatic scale selection, the implementation of the saliency map reported here is not the most efficient, because it uses an *additional* Laplacian pyramid. To implement the saliency map directly from Laplacian pyramids as in [14], instead of the pyramidal structure originally used in [4], improves efficiency and is currently under investigation.

Acknowledgements

This paper describes work conducted at the University of Essex and was sponsored by the CAPES Foundation (Brazil). Their support is gratefully acknowledged.

References

[1] H. Vieira Neto, U. Nehmzow, Automated exploration and inspection: Comparing two visual novelty detectors, International Journal of Advanced Robotic Systems 2 (4) (2005) 355–362.

[2] U. Nehmzow, H. Vieira Neto, Visual attention and novelty detection: Experiments with automatic scale selection, in: Proceedings of TAROS 2006: Towards Autonomous Robotic Systems, Guildford, UK, 2006, pp. 139–146.

[3] T. Lindeberg, Feature detection with automatic scale selection, International Journal of Computer Vision 30 (2) (1998) 194–203.

[4] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11) (1998) 1254–1259.

[5] K. Mikolajczyk, C. Schmid, Indexing based on scale invariant interest points, in: International Conference on Computer Vision, Vol. 1, 2001, pp. 525–531.

[6] S. Marsland, J. Shapiro, U. Nehmzow, A self-organising network that grows when required, Neural Networks 15 (8-9) (2002) 1041–1058.

[7] M. Artač, M. Jogan, A. Leonardis, Incremental PCA for online visual learning and recognition, in: Proceedings of the 16th International Conference on Pattern Recognition (ICPR 2002), Vol. 3, Quebec, Canada, 2002, pp. 781–784.

[8] R. Levin, D. Rubin, Applied Elementary Statistics, Prentice-Hall, London, UK, 1980.

[9] U. Nehmzow, Mobile Robotics: A Practical Introduction, 2nd ed., Springer-Verlag, London, UK, 2003.

[10] H. Vieira Neto, Visual novelty detection for autonomous inspection robots, Ph.D. thesis, University of Essex, Colchester, UK (2006).

[11] L. Sachs, Angewandte Statistik: Anwendung statistischer Methoden, Springer Verlag, Berlin, Germany, 2004.

[12] H. Vieira Neto, 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.

[13] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[14] J. L. Crowley, O. Riff, J. Piater, Fast computation of characteristic scale using a half octave pyramid, in: Proceedings of the International Workshop on Cognitive Vision (CogVis 2002), Zurich, Switzerland, 2002.

[15] K. Mikolajczyk, C. Schmid, Scale and affine invariant interest point detectors, International Journal of Computer Vision 60 (1) (2004) 63–86.

[16] J. Shi, C. Tomasi, Good features to track, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94), Seattle, WA, 1994, pp. 593–600.