Visual Attention and Novelty Detection: Experiments with Automatic Scale Selection

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

Abstract

We present experiments with an autonomous inspection robot, whose task was to highlight novel features in its environment using camera images.

Experiments were conducted with two different attention mechanisms — saliency map and multi-scale Harris detector — and two different novelty detection mechanisms — the Grow-When-Required neural network and incremental PCA. For both mechanisms we compared fixed-scale image encoding with automatically scaled image patches.

Results show that using 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.

1. Introduction

Novelty detection mechanisms and, more generally, attention mechanisms are 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 further 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. The ability to identify perceptions that were never experienced before — novelty detection — is therefore an essential component for mobile robots aiming at true autonomy, adaptability to new situations and continuous operation.

Obviously, the sensor modality used as perceptual input also plays an important role in the agent's performance for a given task or behaviour. If relevant features in the environment cannot be sensed and discriminated, it will be impossible for the agent to respond appropriately. Within the variety of sensors normally available to a mobile robot, vision allows measurement and estimation of several environmental features, such as texture, colour, shape, size and distance to a physical object. Therefore, vision is very versatile and also has the advantage of being able to provide high resolution readings in two dimensions, making the detection of small details of the environment more likely.

Novelty detection through vision. Our work investigates novelty detection mechanisms using vision as perceptual input, with potential application in automated inspection. Because novelty detection entails the identification of *any* unusual perceptions, which are unknown beforehand, it is not possible to construct and install models of abnormality. Therefore, we follow the approach of acquiring a model of *normality* from the operating environment and then filter out perceptions that fail to fit this model. This approach was successfully used in real robots using sonar sensing (Marsland et al., 2002a) and more recently using colour vision with unrestricted field of view (Vieira Neto and Nehmzow, 2004, Vieira Neto and Nehmzow, 2005) in inspection tasks.

In order to deal with the large amount of data provided by a camera, we used an attention model to select raw image patches from the input image frame in previous work (Vieira Neto and Nehmzow, 2005). These image patches were normalised to unit-length vectors and fed to a novelty filter that indicated the presence or not of novelty. Here we follow the same approach, shown in Figure 1.

The attention model plays an important role in the overall performance of our visual novelty detection framework. It not only allows the localisation of *where* novel visual features are within an input image frame, but also contributes to reduce the dimensionality of input vectors through the use of local image patches rather

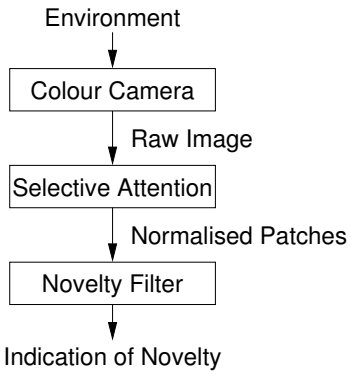


Figure 1: Visual novelty detection using raw image patches: an attention model selects candidate regions, which are then normalised and fed to a novelty filter.

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 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.

Image scaling. The motivation to investigate automatic scale selection (image patch size) stems from the fact that obtaining an image encoding method that is robust to changes in scale would improve the novelty filter's ability to generalise, reducing the number of acquired concepts by the learning mechanism in use. In previous work (Vieira Neto and Nehmzow, 2005), generalisation according to scale is achieved through the acquisition of multiple image patches, in different scales, for the salient visual features found in the environment. The use of raw image data is advantageous in the sense that it allows patch reconstruction from the acquired model of normality, providing visual feedback of which aspects of the environment were actually learnt.

In this work, we exploit the characteristic scale property (Lindeberg, 1998) 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 (Itti et al., 1998) and the multi-scale Harris detector (Mikolajczyk and Schmid, 2001) as attention mechanisms. The Grow-When-Required (GWR) neural network (Marsland et al., 2002b) and the incremental PCA algorithm (Artač et al., 2002) are used as novelty filters.

2. Novelty Detection

In novelty detection tasks one commonly desires to detect any *previously unknown* feature, as opposed to recognition tasks in which features of interest are already known. Therefore, the feasible approach to be followed is to learn a model of normality from the environment using an unsupervised learning mechanism, and then to use this acquired model to filter out abnormal perceptions. In our work we concentrate on on-line unsupervised learning based either on neural networks or statistical approaches.

Here we use the GWR network, which was especially designed for the task of on-line novelty detection. This neural network combines a clustering mechanism and a model of habituation to decide if a given input vector is novel and therefore needs to be incorporated to the current model (Marsland et al., 2002b). Another alternative is the incremental PCA algorithm introduced in (Artač et al., 2002). In this case, we used the magnitude of the residual vector (the RMS error between original data and its reconstruction from the eigenspace projection) as a means to determine novelty. Incremental PCA offers the advantage of intrinsically reducing input dimensionality, allowing optimal reconstruction (minimal squared error). Details of both novelty detection methods used here are given in (Vieira Neto and Nehmzow, 2005).

2.1 Experimental Setup

Our experiments were conducted with a Magellan Pro robot (*Radix*) 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.

Initially a learning phase was carried out so that the robot could acquire a model of normality from the environment. The learning phase comprised five loops around the empty arena, resulting in the acquisition of 225 image frames. After the acquisition of the model of normality, a novel object (either an orange football or a grey box) was deliberately placed inside the arena and the trained robot was used to inspect the environment. Examples of acquired images containing the novel objects inside the arena are given in Figure 2.

(a) (b)

Figure 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 filter would highlight the location of novel stimuli. The inspection phase also comprised five loops around the arena containing each of the novel objects, resulting in 450 acquired image frames.

2.2 Quantitative Assessment of Results

After acquiring images from the arena, we manually generated ground truth data in the form of a binary image for each input image where novel objects were present (the pixels corresponding to novelty were highlighted in these binary images). Using this ground truth information, contingency tables were built relating system response to 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

Cramer's V and uncertainty coefficient U. We established statistical significance of the association between actual novelty status (ground truth) and the novelty filter response using a χ^2 analysis of the contingency tables (Nehmzow, 2003). The strength of this association was quantified through Cramer's V ($0 \leq V \leq 1$, with smaller values indicating a weaker association) and the uncertainty coefficient U ($0 \leq U \leq 1$, again with smaller values indicating a weaker association). Full details are given in (Vieira Neto and Nehmzow, 2005).

Index of agreement κ. A further statistic used in this paper is the κ index of agreement, which is computed as follows (Sachs, 2004):

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

where A, B, C and D are the entries in Table 1.

This statistic is used to assess the agreement between ground truth and novelty filter response, in a similar fashion to the V and U statistics. However, κ has the advantage of having an established semantic meaning associated with some intervals, as shown in Table 2 (Sachs, 2004).

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.

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

3. Models of Visual Attention

In previous work (Vieira Neto and Nehmzow, 2004, Vieira Neto and Nehmzow, 2005) we reported experiments using the saliency map (Itti et al., 1998) as a mechanism of visual attention using a fixed number of salient points. This model is inspired by the early primate visual system and consists of multi-scale feature maps that allow the detection of local changes in intensity, colour and orientation in different scales.

The feature maps are computed from image pyramids obtained from the original input image. In our implementation, Gaussian and oriented Gabor pyramids with five scales were built, as described in (Itti et al., 1998). Across-scale differences were then 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. Full implementation details are available in (Vieira Neto, 2006).

A normalisation operator is used in order to combine intensity, opponent 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 (Itti et al., 1998). The final saliency map was computed in pyramid level 2, meaning that there was a 1:4 reduction in scale of the saliency map with respect to the original image size.

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. In previous work (Vieira Neto and Nehmzow, 2004, Vieira Neto and Nehmzow, 2005), 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 a serious shortcoming: the resulting resolution for the location of salient points in this case is four pixels. A solution to this problem is to interpolate the location of local maxima in the saliency map to sub-pixel accuracy using a Taylor expansion up to the second order term.

We also devised and implemented a method to determine the number of salient locations automatically. The average saliency value ($\bar{S}$) and the maximum saliency value ($\hat{S}$), which corresponds to the first location to be attended by the attention mechanism, are used to determine a saliency threshold (S_T) for the selection of salient points:

$$S_T = \bar{S} + k(\hat{S} - \bar{S}), 0 \leqslant k \leqslant 1, \quad (2)$$

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

Salient locations are then determined by a search for local maxima whose value is above the saliency threshold S_T. The determined coordinates and their neighbours are then used to interpolate the location of maxima with sub-pixel accuracy using a Taylor expansion up to the second derivative:

$$\hat{x} = -\frac{S_x}{S_{xx}} = \frac{S(x-1,y) - S(x+1,y)}{S(x+1,y) - 2S(x,y) + S(x-1,y)}, \quad (3)$$

$$\hat{y} = -\frac{S_y}{S_{yy}} = \frac{S(x,y-1) - S(x,y+1)}{S(x,y+1) - 2S(x,y) + S(x,y-1)}, \quad (4)$$

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 3 and 4 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.

We also implemented the multi-scale Harris detector (Mikolajczyk and Schmid, 2001) 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 because they are stable in both space and scale (Lowe, 2004). A fast and efficient algorithm to build Laplacian image pyramids proposed in (Crowley et al., 2002) was used in our implementation. In this algorithm, half-octave pyramids are constructed by successive Gaussian filtering, subsampling and subtraction.

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 in a Gaussian pyramid with a scale factor of $\sqrt{2}$ (Crowley et al., 2002). Our implementation used a half-octave Laplacian pyramid with ten levels (scales).

After the Laplacian pyramid is built, 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 eighteen neighbours in the levels above and below. The location of extrema is interpolated using equations 3 and 4 for better accuracy. Extrema corresponding to locations with low contrast ($|L| < 0.02$, assuming normalised values in the range [0,1]) were rejected. 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 poorly defined 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. Therefore, we also rejected locations with a principal curvature ratio $r < 4$. For full implementation details the reader is referred again to (Vieira Neto, 2006).

4. Experiments With Fixed Scale

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 as in (Vieira Neto and Nehmzow, 2005). Raw image patches were used to allow reconstruction from the acquired model of normality. 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. An attention mechanism that provides better interest point stability and accuracy is expected to also provide better overall performance when using raw image patches.

Initially, a fixed scale size of 24×24 pixels was used for the image patches. As attention mechanisms we used the interpolated saliency map, as described before, and 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 parameter 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 acquired model of normality of the empty arena was then used to filter out any abnormal perceptions during inspection of the arena. Inspection was conducted with the presence of novel objects (an orange football and a grey box) in the arena and the results obtained with each attention mechanism are given in Table 3, including the sizes of the acquired models.

Results. All experiments resulted in statistically significant correlation between novelty ground truth and

Table 3: Visual novelty detection performance comparison using different interest point selection methods (fixed scale) and the GWR network.

	Interpolated Saliency	Multiscale Harris Det.
Model Size	5 nodes	4 nodes
Orange ball	$V = 0.93$ $U = 0.81$ $\kappa = 0.92$	$V = 0.89$ $U = 0.73$ $\kappa = 0.89$
Grey box	$V = 0.76$ $U = 0.53$ $\kappa = 0.73$	$V = 0.59$ $U = 0.30$ $\kappa = 0.51$

the classification made by the GWR network (χ^2 analysis, $p \leq 0.01$). The same experiments were repeated using the incremental PCA algorithm as novelty filter (the residual error threshold for the incremental PCA algorithm was $r_T = 0.25$). The results obtained are shown in Table 4.

Table 4: Visual novelty detection performance comparison using different interest point selection methods (fixed scale) and incremental PCA.

	Interpolated Saliency	Multiscale Harris Det.
Model Size	30 vectors (28 dim.)	28 vectors (27 dim.)
Orange ball	$V = 0.84$ $U = 0.61$ $\kappa = 0.84$	$V = 0.94$ $U = 0.83$ $\kappa = 0.94$
Grey box	$V = 0.75$ $U = 0.50$ $\kappa = 0.73$	$V = 0.63$ $U = 0.31$ $\kappa = 0.62$

Once 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. The reconstructed images from the acquired incremental PCA models using the interpolated saliency map and the multi-scale Harris detector are shown in Figure 3, where one can notice that the acquired models using either interest point detector are quite similar.

5. Experiments With Automatic Scale

As discussed previously, as a result of the robot navigation 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 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,

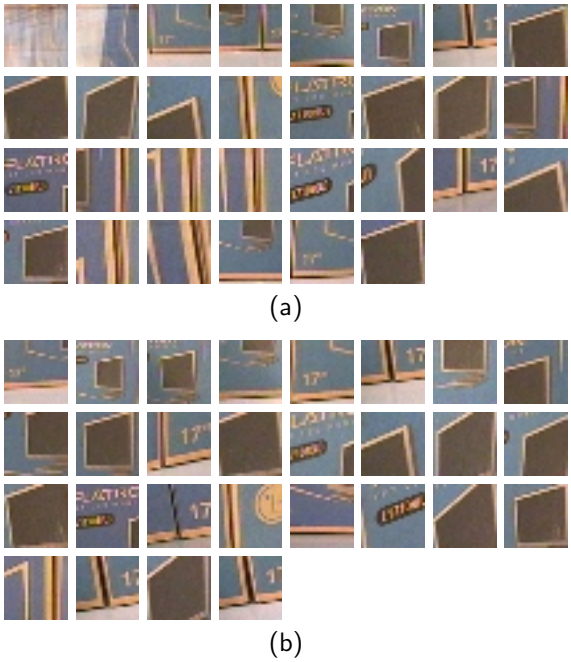


Figure 3: Image patches (fixed scale) acquired using incremental PCA: (a) interpolated saliency map; and (b) multi-scale Harris detector. Both models are similar in contents and in size.

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, generalisation according to scale was achieved by the acquisition of multiple versions of salient visual features in different scales by the learning mechanism. 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 concepts in the model of normality of the environment.

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 a Difference-of-Gaussian image pyramid at the coordinates of that 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 (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 used in (Lowe, 2004, Crowley et al., 2002) to determine the region of interest surrounding visual features.

Once the location of an interest point is found, the Laplacian jet 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:

$$\hat{s} = -\frac{L_s}{L_{ss}} = \frac{L(s-1) - L(s+1)}{L(s+1) - 2L(s) + L(s-1)}, \qquad (5)$$

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 (Crowley et al., 2002), the radius of the region of interest can be computed from the interpolated half-octave pyramid level by using the following equation:

$$r_{roi} = 1.18 \times b^{(s+\hat{s})}, \qquad (6)$$

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".

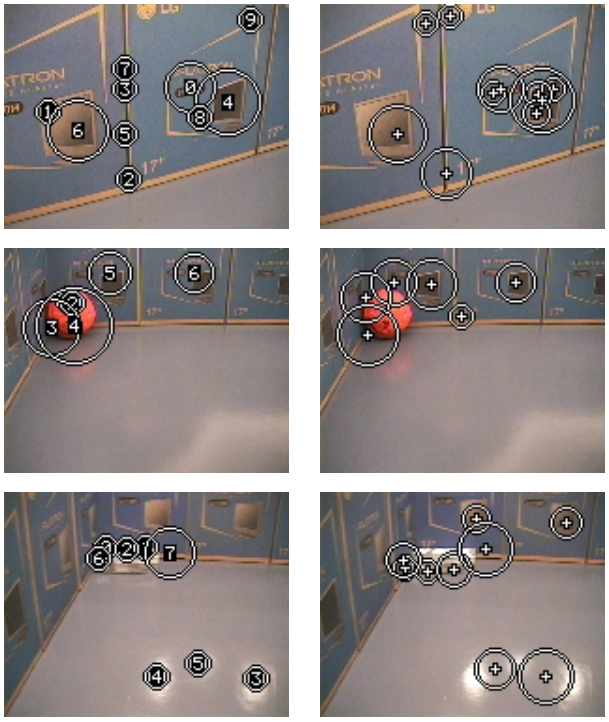
The procedure above can be performed directly in the case of the multi-scale Harris detector because in our implementation we use a scale-space with five octaves, *i.e.* a half-octave Laplacian pyramid with ten levels, which provides sufficient scale resolution. However, in the case of the saliency map, the intensity Laplacian pyramid used has only five levels. Therefore, an additional half-octave Laplacian pyramid was built for the saliency map (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, 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, as shown in the following equation:

$$r = \min\{\max\{6, 2 \times r_{roi}\}, 24\}. \qquad (7)$$

This results in the selection of square image patches centred around the interest points ranging from 12×12 to 48×48 pixels in size. Figure 4 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 7.

The circles in Figure 4 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 models with automatic scale selection to determine the size of the regions. It is important to notice that when both attention mechanisms happen to decide on interest points in similar locations, the size of the corresponding regions of interest are also similar. In these examples it is also possible to notice the preference of the multi-scale Harris detector for the selection of interest points on edges with high curvature, while the saliency map prefers to locate interest points on blobs and straight edges.



(a) (b)

Figure 4: 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.

Experiments using the whole visual novelty detection framework were conducted to assess the impact caused in overall performance by the use of automatic scale selection. In order to obtain input vectors with fixed size for the learning mechanisms, the image patches selected by the attention models were scaled to a fixed image patch size of 24×24 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 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 (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. Only the multi-scale Harris detector contributed to statistically significant association between novelty filter response and actual novelty status at all times.

	Interpolated Saliency	Multiscale Harris Det.
Model Size	4 nodes	2 nodes
Orange ball	$V = 0.83$ $U = 0.69$ $\kappa = 0.88$	$V = 0.47$ $U = 0.17$ $\kappa = 0.47$
Grey box	$V = 0.02^*$ $U = 0.00$ κ = -0.02	$V = 0.25$ $U = 0.05$ $\kappa = 0.15$

*No statistical significance according to the χ^2 test

Table 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 (χ^2 test, $p \leq 0.01$).

	Interpolated Saliency	Multiscale Harris Det.
Model Size	20 vectors (19 dim.)	11 vectors (10 dim.)
Orange ball	$V = 0.94$ $U = 0.80$ $\kappa = 0.94$	$V = 0.51$ $U = 0.20$ $\kappa = 0.50$
Grey box	$V = 0.56$ $U = 0.28$ $\kappa = 0.50$	$V = 0.17^*$ $U = 0.02$ $\kappa = 0.10$

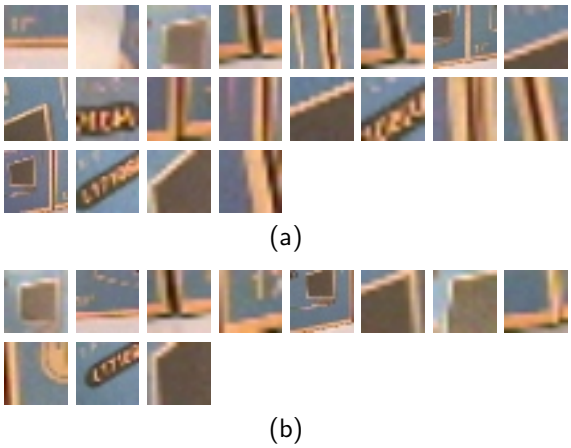
*$p \leq 0.05$

Results. The results in Table 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 (χ^2 test, $p \leq 0.01$) at all times. Overall performances of both learning approaches were worse than the obtained with fixed scale (see Table 3). This fact is attributed to the use of bilinear interpolation scaling, which causes image patch smoothing (a low-pass filtering effect). Smoothing makes differentiation 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. A solution to this problem is to use a higher activation threshold a_T for the GWR network.

The experiments were repeated using the incremental PCA algorithm as learning mechanism, which was expected to be less sensitive to bilinear interpolation smoothing. The expected outcome of using automatic scale selection was that smaller models of normality would be acquired because generalisation according to scale would be improved by the image encoding mechanism itself, rather than the acquisition of multiple scaled versions of the same features by the learning mechanism. A quantitative comparison of the results obtained is given in Table 6.

Despite revealing statistically significant association between system response and ground truth data (χ^2 test, $p \leq 0.01$ except otherwise noted), the results in Table 6 are poorer than the results obtained using fixed scale, except for the case of the orange ball when using the saliency map as attention mechanism (see Table 4). Nevertheless, both visual novelty detectors were still able to highlight novel objects correctly. The acquired models of normality are smaller than the ones acquired using image patches with fixed size, as expected.

The interpolated saliency map results in 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 scale selection are shown in Figure 5, where the fact that the acquired models using automatic scale selection are smaller can also be confirmed by comparisons with Figure 3.



(a) (b)

Figure 5: 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.

6. Conclusion

In this paper we have made an assessment of the influence of the attention mechanism within our visual novelty detection framework, particularly with respect to the use of automatic scale selection. Two distinct interest point detection schemes were investigated, the saliency map (Itti et al., 1998), which presents a preference to locate interest points on blob-like features and straight edges, and the multi-scale Harris detector (Mikolajczyk and Schmid, 2001), which prefers

high curvature edges to locate interest points. Both approaches had their localisation accuracy improved through function interpolation using a second order Taylor expansion as suggested in (Lowe, 2004).

Because there are advantages in using encoding techniques that allow image reconstruction, the accuracy and stability in interest point selection became an important issue. Accurate localisation reduces errors due to misalignment of image patches during matching, having an impact in the overall performance of the visual novelty filter, also contributing to reduce 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 visual features as a result of robot navigation around the environment. In previous work (Vieira Neto and Nehmzow, 2005), 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 in (Lindeberg, 1998).

The results in Figure 5 and Table 6 corroborate our hypothesis because the use of automatic scale selection reduced the size of the acquired PCA model of normality. However, overall performance of the novelty filters was generally worse than when using fixed scale image patches, requiring further research in this topic. Performance comparisons were made quantitatively through contingency table analysis and computation of Cramer's V, uncertainty coefficient U and the κ index of agreement (Sachs, 2004).

Among the investigated 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 of the saliency map reported here is not the most efficient because it uses an additional Laplacian pyramid. An implementation built from half-octave pyramids as in (Crowley et al., 2002) instead of the pyramidal structure originally used in (Itti et al., 1998) constitutes a better scenario for further investigations in automatic scaling.

Future research aims at improving performance results and adding robustness to general affine transformations through the use of affine-invariant interest point detectors (Mikolajczyk and Schmid, 2004, Shi and Tomasi, 1994).

Acknowledgements

Hugo Vieira Neto is sponsored by the Brazilian Government through CAPES Foundation and UTFPR, whose support is gratefully acknowledged.

References

Artač, M., Jogan, M., and Leonardis, A. (2002). 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.

Crowley, J. L., Riff, O., and Piater, J. (2002). Fast computation of characteristic scale using a half octave pyramid. In *Proceedings of the International Workshop on Cognitive Vision (CogVis 2002)*, Zurich, Switzerland.

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

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

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

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

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

Mikolajczyk, K. and Schmid, C. (2001). Indexing based on scale invariant interest points. In *International Conference on Computer Vision*, volume 1, pages 525–531.

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

Nehmzow, U. (2003). *Mobile Robotics: A Practical Introduction, 2nd ed.* Springer-Verlag, London, UK.

Sachs, L. (2004). *Angewandte Statistik: Anwendung statistischer Methoden*. Springer Verlag, Berlin, Germany.

Shi, J. and Tomasi, C. (1994). Good features to track. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, pages 593–600, Seattle, WA.

Vieira Neto, H. (2006). *Visual Novelty Detection for Autonomous Inspection Robots*. PhD thesis, University of Essex, Colchester, UK.

Vieira Neto, H. and Nehmzow, U. (2004). Visual novelty detection for inspection tasks using mobile robots. In *Proceedings of the 8th Brazilian Symposium on Neural Networks (SBRN 2004)*, São Luís, Brazil.

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