Small-scale Anomaly Detection in Panoramic Imaging using Neural Models of Low-level Vision

Matthew C. Casey[*a], Duncan L. Hickman[b], Athanasios Pavlou[a], James R. E. Sadler[b]
[a]Department of Computing, University of Surrey, Guildford, Surrey, GU2 7XH, UK;
[b]Waterfall Solutions Ltd., 1 & 2 Saxton, Parklands, Guildford, Surrey GU2 9JX, UK

ABSTRACT

Our understanding of sensory processing in animals has reached the stage where we can exploit neurobiological principles in commercial systems. In human vision, one brain structure that offers insight into how we might detect anomalies in real-time imaging is the superior colliculus (SC). The SC is a small structure that rapidly orients our eyes to a movement, sound or touch that it detects, even when the stimulus may be on a small-scale; think of a camouflaged movement or the rustle of leaves. This automatic orientation allows us to prioritize the use of our eyes to raise awareness of a potential threat, such as a predator approaching stealthily. In this paper we describe the application of a neural network model of the SC to the detection of anomalies in panoramic imaging. The neural approach consists of a mosaic of topographic maps that are each trained using competitive Hebbian learning to rapidly detect image features of a pre-defined shape and scale. What makes this approach interesting is the ability of the competition between neurons to automatically filter noise, yet with the capability of generalizing the desired shape and scale. We will present the results of this technique applied to the real-time detection of obscured targets in visible-band panoramic CCTV images. Using background subtraction to highlight potential movement, the technique is able to correctly identify targets which span as little as 3 pixels wide while filtering small-scale noise.

Keywords: Panoramic imaging, neural networks, anomaly detection, low-level vision

1 INTRODUCTION

Our understanding of sensory processing in animals has reached the stage where we can exploit neurobiological principles in commercial systems. In human vision, one brain structure that offers insight into real-time imaging is the superior colliculus (SC)[1,2]. The SC is an evolutionary stable structure found in the midbrain of vertebrates, which is responsible for shifting gaze to focus the eye's fovea towards stimuli of interest[2]. This focusing of resources, leading ultimately to conscious perception, is a key survival mechanism. Imagine you are being stalked by a predator that is adept at camouflage and stealthy movement. The predator's approach will be masked by environmental clutter, such as foliage, while their footfalls will be quiet. The SC has developed to prioritize the detection of slight movement and quiet sounds emanating from the same location so that we automatically look towards these locations, and hence react to potential threats[1]. This specialization extends to detecting potential food sources as well, such as in frogs where the optic tectum (the equivalent of the SC in non-mammalian vertebrates) is sensitive to the movement of convex-shaped dark objects, which correspond to the location of flies[3].

As a specialized survival mechanism, the processing of the SC demonstrates interesting computational principles. First, the SC is closely connected to sensory input. Retinal input to the SC predominantly comes directly via the koniocellular pathway, which rapidly responds to motion and luminance[4], and hence fast connectivity and crude processing with a small lag time is key to indicating interesting visual activity. Second, the SC combines visual, auditory and touch information in order to localize[2]. This multisensory integration demonstrates the importance of fusing multiple sources of information from one event to provoke a response, even if the relative intensities of the individual visual, auditory or touch stimuli are low[5]. Third, the SC is closely connected to motor outputs so that it is able to rapidly react. For humans, this output controls the shift of our gaze towards the activity of interest via direct connection with the brain stem[6]. The output also feeds structures such as the amygdala, which primes our body to react whenever fearful stimuli

* m.casey@surrey.ac.uk; phone +44 1483 689635; fax +44 1483 686051; www.surrey.ac.uk/computing

are detected, such as running away from 'snakes in the grass'[7]. Computationally, the SC therefore forms an influential localization system which works on crude sensory input to provoke a rapid reaction.

Given its computational credentials, the SC is therefore an attractive structure to model in order to evaluate whether the same computational principles can be applied to commercial systems. In the first instance this means understanding how crude visual stimuli can be processed rapidly to detect areas of interest ignoring noise, with the longer term view of combining vision with other modalities and more complex, but rapid processing of threats. In this paper we concentrate on the visual properties of the SC, and we therefore describe the application of a neural network model[8] of the superficial layers of the SC to the detection of anomalies in imaging. Our application domain is that of panoramic CCTV images, which can be monitored by operators to detect threats from a wide area. With wide area input, operators may focus only on the central area of the image. In contrast, the SC shows how anomalies in peripheral vision can be detected, and hence we apply this to assist detection rates. Section 2 of this paper briefly describes the neural network model. Section 3 describes the application domain of panoramic imaging. In section 4 we present the results of our approach applied to the real-time detection of obscured targets in panoramic CCTV images. Our conclusion and discussion is provided in Section 5.

2 MODEL SPECIFICATION

The SC is divided into superficial and deep layers[1]. The superficial layer of the SC processes visual input which comes directly from the retina[6], predominantly via the koniocellular pathway[4]. The deep layers of the SC combine visual sensitivity with auditory and somatosensory input to perform multisensory localization[2,5]. Each layer within the SC is formed from topographic maps of the visual, auditory or somatosensory space. These layers are all aligned into an eye-centered representation so that different sensory modalities can be combined[9,10].

Focusing just on the visual localization that the SC performs in its superficial layer, this provides motion[11-13] and contrast[14,15] sensitivity on a binocular representation of visual space. For example, a study on the frog optic tectum by Lettvin et al[3] demonstrated that the frog's equivalent to the SC was formed from four topographic layers. The first layer was sensitive to spatial contrast at sharp edges. The second layer was sensitive to convex-shaped dark objects, which perhaps coincides with the broad shape of the frog's preferred diet of flies. The third layer responds to temporal contrast of moving edges. The fourth layer responds to a sudden reduction in illumination, as might occur if a predator's shadow moves towards the frog.

These examples show how the superficial layer of the SC has developed to process simple spatial contrast, temporal contrast and illumination information in order to localize threats or sources of food across the entire visual field. Of course, because the processing is simple, this provides only a rapid response which can then be moderated by later, more complex processing, and it is therefore subject to false alarms. However, what is important computationally is that the processing, albeit crude, is sufficiently specialized to filter out noise and to concentrate on important stimuli. This approach of having specialist layers of processing is used throughout the visual system. Of particular relevance is the amygdala[7], which is the structure responsible for responding to crude, fearful stimuli so that the body is prepared to react once the stimuli has been consciously perceived and processed. Here, the amygdala goes one stage further by allowing the types of stimuli being detected to be learnt[16], such as through classical conditioning[17]. The question we address in this paper is whether the principles of detecting simple stimuli to rapidly react, such as observed in the SC or other structures like the amygdala, can be applied to imaging systems to improve anomaly detection? To understand this, we focus on how the SC can be modeled computationally before applying this to imaging.

2.1 Modeling the Superior Colliculus

Focusing just on the visual processing conducted in the superficial layers of the SC, we can see from the work of Lettvin et al[3] that this processing is achieved by using layers of topographic maps, where each layer is specialized to one type of visual feature. To achieve this, the retina provides spatial, temporal and illumination information. The core of the processing conducted in the SC is therefore achieved with specialized topographic maps. Each map responds only to a specific type of input, such as contrast at a particular scale, and can automatically localize to the area which has the most salient (or strongest) input.

Our previous work on modeling the SC provides us with a topographic map algorithm that can achieve this localization[8,18]. The map consists of a two-dimensional array of rate-coded neurons, where each neuron corresponds to a

location in the input space. The output y from the neuron at location (i, j) in the map given an m-dimensional input x is calculated as[8]:

$$u_{ij} = \sum_{k=1}^{m} x_k w_{kij}(t) \tag{1}$$

$$y_{ij} = \begin{cases} f(u_{ij}) & \text{if } \left\| c_{ij} - c_{win} \right\| < h(t) \\ f(u_{ij} - y_{win}) & \end{cases} \tag{2}$$

$$y_{ij} = \begin{cases} 1 & u \geq 1 \\ u & 0 < u < 1 \\ 0 & u \leq 0 \end{cases} \tag{3}$$

where $w_{kij}(t)$ is the weight at time step t during training which moderates the input k to neuron (i, j). The key element of this calculation is that some neurons are considered to be the winners in responding to the input (equation 2), and hence all other neurons outside the defined neighborhood of the winners, represented by the function $h(t)$, have their output suppressed. The location of each neuron in the map is defined by the 2-dimensional co-ordinate vector c_{ij}, where the neuron with the maximum activation is located at c_{win}. This improves upon the scheme defined by Armony et al[19] for a single row of neurons to ensure that the map localizes to the strongest part of the input. It is therefore important that the weights are defined appropriately in order for the map to localize to specific patterns of input, such as the signal with the greatest spatial contrast, or particular type of motion

In order to set the weights for each neuron, the map undergoes a period of training using a competitive learning algorithm in which target patterns are presented to the network, the output from each neuron is calculated by selecting a neighborhood of winners, and then each neuron's weight within the neighborhood is adjusted so that the neuron responds more strongly to a similar input in future time steps. During training the neighborhood and the amount of change to each weight is reduced to produce a stable map. Details of this training algorithm are defined by Pavlou & Casey[8]. This training regimen ensures that the map is able to localize to the required patterns, whatever they may be.

2.2 Anomaly Detection

This simple topographic map therefore provides neurons that can respond strongly to particular patterns at different locations within the input. For example, if the input to the map is an image, where the image has been filtered to represent a specific feature through intensity values, then the map can respond to the location and strength of each feature. Here the location is represented by the winning neurons which have the highest output y. These neurons have the highest output because their weights are selective for patterns occurring at specific locations within the input. Within the context of anomaly detection, this means that a suitably trained map can take as input an image which represents, say, movement recorded between successive video frames, with the map then localizing a particular shape or type of movement. The important aspect to this is the ability of the map to generalize so that neurons provide a graded response to partial shapes. In this way, simple video processing techniques can be coupled with a trained neural topographic map in order to detect anomalies which may be obscured.

In the examples we explore in this paper, we train maps to respond to different crude shapes. When presented with filtered images the map can then detect the location in the input which has a signal that most represents the trained shape. For example, if trained on a general shape of a walking person (section 4.1), then the map can detect the location within the image of the signal that most closely matches to the walking shape. However, in order to achieve this we need to provide appropriate filters. Looking back at the role of the SC, one feature that is particularly important is motion. Through temporal information supplied by the retina, the SC can detect high contrast motion[11-13], and this is where our anomaly detection is focused. To simplify our experiments, we use a single filtered input to the map which highlights motion (or more accurately change).

We achieve motion detection through the use of background subtraction on grayscale images (Figure 1). Background subtraction relies upon the definition of an appropriate background (assuming the camera is static), which can result in issues with noise and problems with variation in lighting. Noise may occur, for example, with normal motion in the environment, such as from moving leaves. While variations on background subtraction have been described[20] we use simple thresholding and then rely upon the ability of the map to filter noise in the image when detecting shapes. One of the computational principles important in modeling the SC is its ability to rapidly detect areas of interest. Additional

filtering techniques may be applied to provide better motion detection, however these each add a computational overhead. Instead we follow the principle of performing only a small number of simple filtering stages in order to determine if the map can operate correctly only on crude input. The stages involved in our anomaly detection are shown in Figure 1.

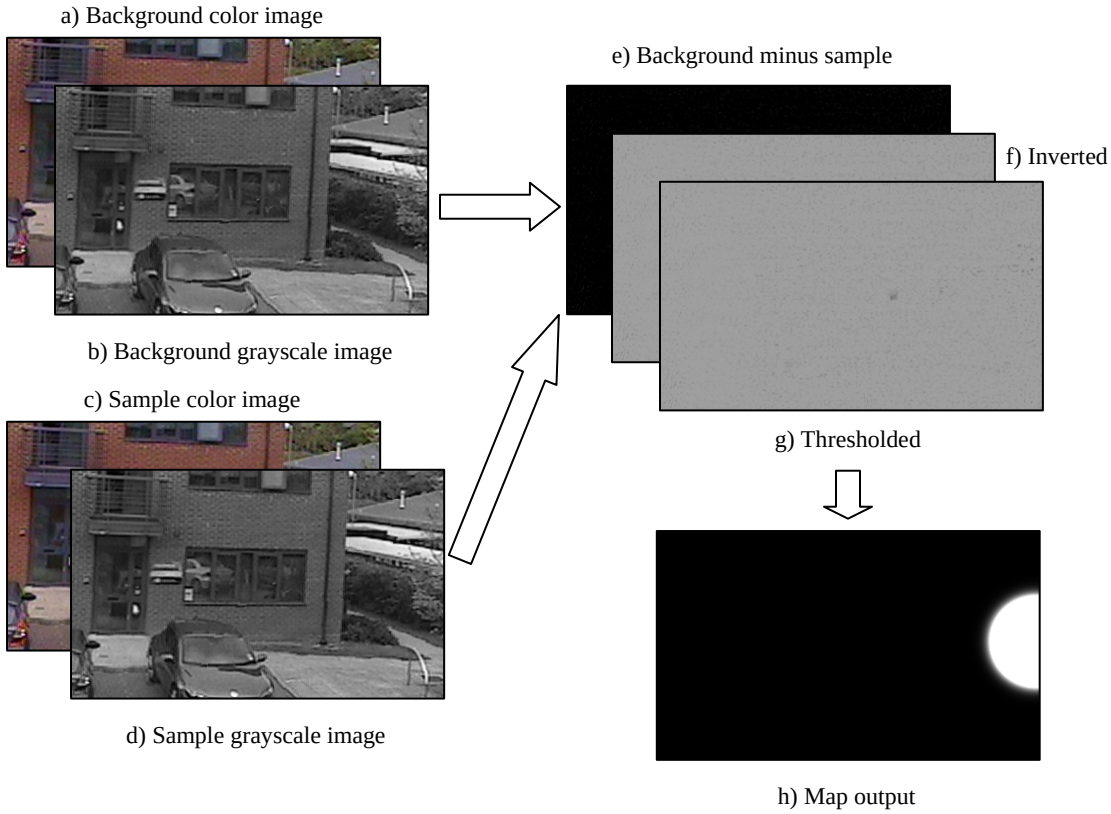


Figure 1. Anomaly detection using a topographic map that is motivated by visual processing in the superior colliculus. Images show the complete processing on a sample image used during the experiments (section 4.2). a) selected background image which is b) converted to grayscale, c) sample image which is converted to d) grayscale. The background and sample are then e) subtracted and the resulting image inverted to show change as f) a high value between 0 and 1. The resulting input is then thresholded so that only values above 0.5 are input to the neural map which then h) localizes the area with the closest matching input to its training patterns.

3 PANORAMIC IMAGING

Human binocular vision has a wide field of view in excess of 100° in the horizontal direction. Although human eyes can only foveate on a small section of this image, typically around 1% to 2%, the system has evolved to detect motion or other stimuli in the periphery of the field of view using structures such as the SC. The SC then provides input to later processing stages, such as the amygdala[21] and successive areas of the visual cortex. This natural processing chain is now being exploited in a variety of applications that centre on the use of panoramic imaging, or wide area surveillance, to provide a method of increasing the situational awareness of the human observer whilst minimizing the footprint of the imaging system.

The use of a panoramic imaging system for the work detailed in this paper enables the neurobiological response to be tested using imagery that closely matches the human visual field. The panoramic imaging system used has previously demonstrated the benefits of wide area situational awareness in a number of user lead trials[22].

Within the context of anomaly detection, it is necessary to detect the small changes in the scene's content. These small changes, or anomalies, can originate from a bag being placed in a secluded corner or people loitering and behaving

against the general pattern of life. A number of detection and tracking methodologies can either be run concurrently or combined to give a single integrated solution. This has been done automatically within the context of processing the received imagery in order to form the panoramic image. The panoramic image itself is formed from any number of partially overlapping images. Typically we have constructed panoramic images from two or three camera inputs up to 360° full panoramic images from twelve cameras. Detection algorithms are then applied to the resulting imagery enabling objects to be tracked across a full field of view. This provides the user with a significant enhancement in situational awareness whilst minimizing the additional workload of a user.

One of the hardest problems to overcome is the issue of obscuration of the target. This is of particular interest for the protection of installations and wider security and surveillance applications. In many real life cases there is often additional clutter, or cover, provided in the scene which can result in reducing the probability of detection. The use of a detection algorithm that can identify small amounts of motion in the scene, and through the intelligent merging of the resulting detections, these disparate features can be linked to provide a heightened situational awareness.

3.1 Generating Panoramic Images

The panoramic images used for the experiments were generated using proprietary image warping and panorama formation algorithms[22]. The resulting sequences were based upon panoramic images formed from three visible-band cameras arranged to give a 100° horizontal field of view (Figure 2).

Figure 2. Example office area panoramic image combined from three visible band camera images. Note the use of a cylindrical warp in order to correctly align the images.

The process of generating a panoramic image involves a number of steps that correct for distortions in the imagery. As with any imaging system, there are a number of distortions that arise from the optical assemblies used. At their simplest these can be approximated by a simple barrel or pincushion distortion, however there are more complicated compound distortions that are characteristic of some lens assemblies.

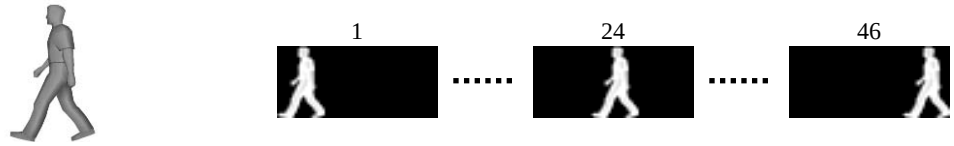
The process of imaging a wide area using a number of sensors with smaller fields of view also leads to the necessity to apply a cylindrical warp. This corrects discrepancies in the overlapping regions between neighboring sensors. The result is a seamless image where corresponding features are co-aligned enabling objects of interest to be tracked across the common field of view between neighboring sensors.

The imagery used for the experiments was captured in a built-up office area (Figure 2) and an industrial area (Figure 7). Both areas offered plenty of natural cover, provided by buildings and parked cars, for the target to hide behind or move amongst. A number of color sequences were recorded with the target popping their head up from behind a car or around the side of a building. This generally equated to a few pixels in the 100° horizontal field of view. The office area was captured during day time and the industrial area at dusk using low-light level street lighting. There was no significant variation in lighting within each scene, given the use of simple background subtraction.

4 EXPERIMENTS AND EVALUATION

In the examples we explore in this paper, we train maps to respond to different human-type shapes at a specific small-scale. The first experiment demonstrates how a walking person can be detected when they are in full view of the camera

or partially obscured by a tree. To achieve this, we trained a map for 400 epochs on 46 images of 64 by 32 pixels each of which had a representation of a human walking pose at different locations within the image (Figure 3). The walking pose was generated from a 3-dimensional model, but was static across the training images.



a) 3-dimensional template

b) Sequence of training images used

Figure 3. Training images used for the first experiment, a) template of human walking, b) three examples from the resulting 46 training images. Note that the training data are grayscale images with the item of interest depicted by a high (white) intensity value and the background as zero (black).

For the subsequent experiments we used a generic Gaussian blob shape to capture any change of a particular scale. In particular this allows us to detect a person even at a scale of 3 pixels wide but filters out small scale noise. This was then applied to various scenarios, including a person popping their head out from behind a wall, to a person crouching behind a car or walking behind a tree. To achieve this, we trained a map for 400 epochs on 32 images of 64 by 32 pixels each of which had the Gaussian blob at different locations within the image (Figure 4).

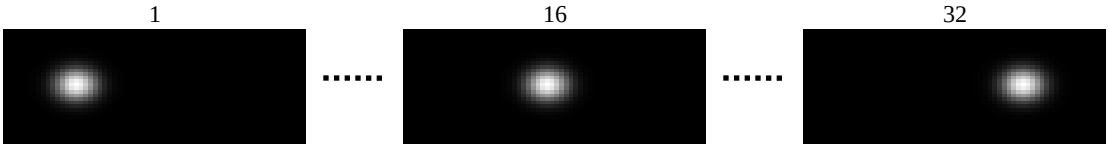


Figure 4. Training images used for detecting obscured small scale motion. Note that the training data are grayscale images with the item of interest depicted by a high (white) intensity value and the background as zero (black).

The Gaussian blob was defined to have a radius $\sigma=3$, such that the input x at pixel (i,j) is defined as:

$$x_{ij} = \lambda e^{-\left(\frac{(i-c)^2+(j-d)^2}{2\sigma^2}\right)} \tag{4}$$

where the center of the blob is at (c,d) and the amplitude $\lambda=1$.

The generated panoramic images were 1344 pixel wide by 580 pixels high. Each topographic map was trained on images 64 by 32 pixel images. Therefore, in order to apply the map to the panoramic images, each image used a mosaic of 21 by 18 topographic maps. Each map consisted of 2 by 64 neurons so that a neuron represented 1 pixel horizontally and 16 pixels vertically. The results from each map were then combined to provide a single localization output.

4.1 Pilot Experiment: Detecting People using Shape Profiles

The first experiment was used to test the approach on a single (non-panoramic) image in order to evaluate whether the neural algorithm could be applied to visible band images to detect crude shapes. To evaluate the approach, we recorded scenes which included people walking across the foreground and background of the image (hence different scales), which had both natural features, such as trees, and man-made features, such as buildings and walls. From the candidate scenes recorded, we chose 114 frames from one sequence of 744 frames (recorded at 25 frames per second). Each of the 114 frames included a person depicted at a scale that the map should be able to detect, as well as people at larger scales. The person was either shown clearly, partially or was totally obscured. The scene was taken during the rain and therefore included additional environmental noise (Figure 5).

Figure 5. Example image from the pilot experiment. This was tested against maps trained on a walking person silhouette of approximately the same scale as the person highlighted in the red circle.

Since the image was smaller than the panoramic images tested in later experiments (768 by 576 pixels), a mosaic of 12 by 18 duplicate topographic maps were used. Instead of using background subtraction, in this pilot we used the grayscale pixel intensity as the input to the map. Since the objects of interest were darker than the background (black clothes against a light colored wall), the image was inverted such that dark pixel values were given a high input to the map. The source map was trained on the patterns as shown in Figure 3 so that the person in the selected frames should be detected. This experiment was therefore designed only to determine if the map could detect crude shapes of the relevant size that were obscured to various degrees and with environmental noise.

The mosaic of maps correctly localized the person when they were in clear view in 20 out of 85 frames (24%), and when they were obscured by trees or other artifacts in 2 out of 20 frames (10%), giving a total detection rate of 21%. When the person was completely obscured (9 frames), and throughout all 114 frames tested, there were no false alarms. This therefore demonstrated that the map could detect the required shapes, albeit with at a low rate.

There were three clear limitations with this approach. First, input to the map relied upon pixel intensity which was inverted to provide detection of dark shapes. The map could therefore not detect people walking, for example, in light colored clothing. Second, the pattern on which the map was trained has a specific silhouette where the person is depicted with their legs apart (Figure 3). For a significant portion of the time when a person is walking, their legs are together. The pattern that the map was trained upon was therefore too specific and detection only occurred when the pattern matched closely. Third, because we use a mosaic of maps, when a relevant shape crosses a map boundary, it is not detected because each map only receives input for part of the shape. We address the first two of these issues in subsequent experiments.

4.2 Evaluating Detection on Panoramic Images

The limitations of the pilot experiment demonstrated the need to make both the input and the pattern being detected more generic. For example, by directly using pixel values we are constraining the detection to intensity, while using a very specific pattern to train the map means that there is less generalization to similar patterns. We overcome this by looking for changes in the input using background subtraction, and use a more generic training pattern, recognizing that may significantly increase the number of false alarms.

Background subtraction is used to record change between an initial, background image and the current frame. This provides a high-valued input for any pixel values that differ to the background (Figure 1). Instead of using a walking pose to train the map, we use a Gaussian blob with radius of 3 pixels and therefore the map will detect the most significant change that is at least 3 pixels or larger, but this will filter out multiple responses for co-located change. For example, only the most significant change resulting from a person moving will be detected, not the movement of the whole person.

This setup was applied to panoramic images captured during daylight, an example of which is shown in Figure 2. Three video sequences were recorded and frames selected from these to test detection under different scenarios. Details of each testing sequence and the detection results are given in Table 1 (sequences 1 to 4).

Sequences 1 to 3 were designed to test the ability of the map to detect targets that were obscured or in clear view. They did not contain any significant distracters and hence had a low false alarm rate. The total detection rate for sequence 1 was 60% (100% when the target was in clear view, 56% when obscured) with 0 false alarms. For sequence 2 this was

63% with 0 false alarms, and sequence 3 achieved 84% detection but with 3 false alarms. An example output from sequence 3 is shown in Figure 6. None of the target patterns crossed a map boundary. False alarms were generated by changes in the reflection of sunlight in the office windows. Therefore, by having the input based upon change and by making the training pattern more generic, a greater detection rate was achieved.

Table 1. Panoramic testing sequences selected to test the model. Sequences 1 to 4 were of the office area (Figure 2), sequence 5 was of the industrial area at dusk (Figure 7). Both consisted of a mixture of buildings, pavements, parked cars, trees and other foliage.

Sequence	**Description**	**Frames**	**Detection**		**False Alarms**
			Obscured (total)	**In clear view (total)**	
1	Detection of obscured person walking Single movement from a person walking behind a tree and parked motorbike. In full view the person was 75 by 25 pixels. No other significant movement. Some environmental noise (leaves and reflections).	70	35 (63)	7 (7)	0
2	Detection of person crouching behind a car Single movement of person who starts crouched behind a car, stands up to be half visible, moves along the car then crouches down. Visible area of person varies from 7 to 28 pixels. No other significant movement. Some environmental noise (leaves and reflections).	56	35 (56)	0 (0)	0
3	Detection of person's head popping out from behind a wall Single movement from a person standing out of view behind a corner who then pops their head out. Visible portion of the head is 3 pixels wide. No other significant movement. Some environmental noise (leaves and reflections).	32	21 (25)	0 (0)	3
4	Detection of person's head popping out from behind a wall with other movement Movement from a person standing out of view behind a corner who then pops their head out. Other significant movement from a car passing out of frame and a person walking through the center of the image with size 55 by 20 pixels.	61	48 (48)	0 (0)	28
5	Detection of obscured person walking and running Various types of movement from a person starting behind a parked car (head only visible as 1 pixel) walking towards the camera, passing behind signs, hiding in a bush before running towards and then away from the camera. In full view the person was 68 by 34 pixels. Other significant movement from a fluttering flag. Environmental noise from moving clouds as well as noise from the low light conditions.	821	Person: 66 (243)	Person: 155 (379) Flag: 375 (821)	859

Sequence 4 was designed to test detection in a more realistic scenario where there was other movement and environmental noise. This repeated the same scenario as sequence 3 but had another person walking through the center of the image and a car just leaving the left side of the frame. Sunlight also varied in a number of reflections from parked cars and windows. The total detection rate was 100% for the target but with 28 false alarms, assuming we treat other large-scale movement as a false alarm. Here, the car was detected in 2 out of the 4 frames it was present, leaving the remaining 26 false alarms occurring because of reflections and background noise. The second person walking through the scene was not detected because they did not move far enough from their initial (background) position.

Figure 6. Example output from the model for sequence 3. The top frame shows the input panoramic image with the target highlighted by the red box and inset. The target is a person's head popping out from behind a wall with a width of 3 pixels. The bottom frame shows the model output superimposed over the input.

These sequences therefore demonstrated that the technique could be applied to panoramic images to detect small-scale movement. However, by using simple background subtraction and a generic training pattern we have increased the false alarm rate. The advantage of the technique is that it can detect particular shapes rapidly and, despite the false alarms, it can filter out some small-scale noise.

Having tested the capability of the technique, we then went on to evaluate it on a longer video sequence which included a target that we wish to detect that was either hidden, obscured or in full view, as well as other targets and environmental noise (sequence 5). This scenario was also captured at dusk so that there was a mixture of street lighting and changing patterns of light from the sun (Figure 7).

Figure 7. Example industrial area panoramic image combined from three visible band camera images.

In the 821 frames, 622 showed the target person moving towards and then away from the camera. The visible portion of the target started off at 1 pixel. Clear throughout the sequence was a flag fluttering in the wind which should be detected when it moves from its initial position. There was also environmental movement, such as clouds, which deviated incrementally from the background image to have the biggest change by the last frame.

The total detection rate of the person was 36% (27% obscured and 41% in clear view). This is low in comparison with the previous sequences, but demonstrates the difficulty of detecting small-scale movement in the order of a few pixels. The technique detected the person when their visible portion was as little as 3 pixels, right up until the maximum extent at 68 pixels. This is most evident when the person was hiding behind a bush and moving around so that their head popped out for only one or two frames at a time. However, the person was not detected at the boundary between maps. The detection rate of the flag was 46%, although this does not take into account when the flag was close to its initial, background position (no wind) where only slight movement would not be detected because it was less than 3 pixels. The false alarm rate was significant at 859 for the whole sequence, with multiple alarms being raised in many images (maximum 4 in any given frame). These were mostly caused by the small movement of clouds in the scene which increased from the initial background over time, and which was particular prominent towards the end of the sequence in the top right hand area of the image where the sun was strongest.

5 CONCLUSION

In this paper we have evaluated the use of a biologically inspired model of the SC for anomaly detection in panoramic CCTV imaging. We have demonstrated how a rate-coded neural network model of the visual layers of the SC can be used to rapidly localize small-scale movement in an image based upon a pre-defined shape. This is different to other types of image detection algorithm in that the technique goes beyond the use of convolutions by using competition between neurons in a topographic map to provide the most salient localization. We have evaluated this against the use of a specific shape (a walking person) and a generic shape (a Gaussian blob). For either, small-scale change in an image can be detected in as little as 3 pixels allowing detection of partially obscured movement. However, by using a specific shape we restrict detection and ignore other anomalies which do not match the target. When using a generic shape the false alarm rate increases considerably but this allows us to detect any small-scale change. The output from the maps may therefore benefit from further processing, and could perhaps be used as an early warning input to more sophisticated analysis and tracking algorithms which could not detect small-scale anomalies.

There are three limitations with this work, notably focused on the balance between detection rates and false alarms. First, to detect anomalies, the topographic map relies on input that has been sufficiently processed to provide contrast (spatial or temporal). Limited pre-processing is rapid but also contributes to the false alarm rate because of naïve techniques such as background subtraction. To reduce the reliance on pre-processing, the topographic maps may be trained to automatically detect, for example, temporal contrast (movement) by introducing lateral inhibition between neurons across frames. This would also improve the detection rate while keeping the ability to pre-define shapes. Second, we have kept the map sizes small and used a mosaic of maps for detection. The reason for this is the lengthy process of training which takes longer for larger map sizes. A simple one-shot training scheme can be used to overcome this and hence allow us to scale up the maps to avoid having map boundaries. This will increase the detection rate by removing map boundaries. Third, when using a generic training pattern the number of false alarms is increased. These can be decreased by either using improved pre-processing techniques, rather than simple background subtraction, or by training the weights within the map to detect movement through lateral inhibition. Such an approach will remove the need for pre-processing, as described above. Alternatively, the output from additional layers of maps trained on different scale shapes could be compared to highlight the most salient signal at a range of scales.

A further opportunity for the work is to explore the combination of different modality inputs, such as combining images with sound. The SC integrates visual, auditory and somatosensory information for localization. Initial work combining images and sound has already been attempted[23] and this may prove beneficial to the detection of multimodal anomalies.

ACKNOWLEDGEMENTS

This work was sponsored by the EPSRC through the University of Surrey Knowledge Transfer Account project (EP/H500189/1).

REFERENCES

[1] King, A. J., "The Superior Colliculus," Current Biology, 14(9) R335-R338 (2004).

[2] Stein, B. E. and Meredith, M. A., [The Merging of the Senses], A Bradford Book, MIT Press, Cambridge, MA., (1993).

[3] Lettvin, J. Y., Maturana, H. R., Mcculloch, W. S. and Pitts, W. H., "What the Frog's Eye Tells the Frog's Brain," Proceedings of the IRE, 47(11) 1940-1951 (1959).

[4] Isbell, L. A., "Snakes as Agents of Evolutionary Change in Primate Brains," Journal of Human Evolution, 51(1) 1-35 (2006).

[5] Stein, B. E. and Stanford, T. R., "Multisensory Integration: Current Issues from the Perspective of the Single Neuron," Nature Reviews Neuroscience, 9(4) 255-266 (2008).

[6] May, P. J., [The Mammalian Superior Colliculus: Laminar Structure and Connections], Buttner-Ennever, J.A. (Ed), Progress in Brain Research: Neuroanatomy of the Oculomotor System, Volume 151, Elsevier, 321-378 (2006).

[7] Öhman, A., Flykt, A. and Esteves, F., "Emotion Drives Attention: Detecting the Snake in the Grass," Journal of Experimental Psychology: General, 130(3) 466-478 (2001).

[8] Pavlou, A. and Casey, M. C., "Simulating the Effects of Cortical Feedback in the Superior Colliculus with Topographic Maps," Proceedings of the International Joint Conference on Neural Networks (IJCNN) 2010, (2010).

[9] Zella, J. C., Brugge, J. F. and Schnupp, J. W. H., "Passive Eye Displacement Alters Auditory Spatial Receptive Fields of Cat Superior Colliculus Neurons," Nature Neuroscience, 4(12) 1167-1169 (2001).

[10] King, A. J., Schnupp, W. H. and Thompson, I. D., "Signals from the Superficial Layers of the Superior Colliculus Enable the Development of the Auditory Space Map in the Deeper Layers," Journal of Neuroscience, 18 9394-9408 (1998).

[11] Sterling, P. and Wickelgren, B. G., "Visual Receptive Fields in the Superior Colliculus of the Cat," Journal of Neurophysiology, 32(1) 1-15 (1969).

[12] Rauschecker, J. P. and Harris, L. R., "Auditory and Visual Neurons in the Cat's Superior Colliculus Selective for the Direction of Apparent Motion Stimuli," Brain Research, 490(1) 56-63 (1989).

[13] Wallace, M. T., McHaffie, J. G. and Stein, B. E., "Visual Response Properties and Visuotopic Representation in the Newborn Monkey Superior Colliculus," Journal of Neurophysiology, 78(5) 2732-2741 (1997).

[14] Prévost, F., Lepore, F. and Guillemot, J. P., "Spatio-temporal Receptive Field Properties of Cells in the Rat Superior Colliculus," Brain Research, 1142 80-91 (2007).

[15] Schneider, K. A. and Kastner, S., "Visual Responses of the Human Superior Colliculus: A High-Resolution Functional Magnetic Resonance Imaging Study," Journal of Neurophysiology, 94(4) 2491-2503 (2005).

[16] Morris, J. S., Öhman, A. and Dolan, R. J., "Conscious and Unconscious Emotional Learning in the Human Amygdala," Nature, 393(6684) 467-470 (1998).

[17] Pavlov, I. P., [Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex], Oxford University Press, London, (1927).

[18] Pavlou, A. and Casey, M. C., "A Computational Platform for Visual Fear Conditioning," Proceedings of the International Joint Conference on Neural Networks (IJCNN) 2009, (2009).

[19] Armony, J. L., Servan-Schreiber, D., Cohen, J. D. and LeDoux, J. E., "Computational Modeling of Emotion: Explorations Through the Anatomy and Physiology of Fear Conditioning," Trends in Cognitive Sciences, 1(1) 28-34 (1997).

[20] Piccardi, M., "Background Subtraction Techniques: A Review," Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics, 4 3099-3104 (2004).

[21] Shi, C. and Davis, M., "Visual Pathways Involved in Fear Conditioning Measured With Fear-Potentiated Startle: Behavioral and Anatomic Studies," The Journal of Neuroscience, 21(24) 9844-9855 (2001).

[22] Sadler, J. R. E., Davis, J. and Hickman, D. L., "A Compact Wide-area Surveillance System for Defence and Security Applications," Proceedings of SPIE Defense, Security, and Sensing Conference 2011 on Enhanced and Synthetic Vision, 8042B (2011).

[23] Casey, M. C., Pavlou, A. and Timotheou, A., "Mind the (Computational) Gap," Proceedings of the UK Workshop on Computational Intelligence (UKCI 2010), (2010).