CONTEXTUAL PERSON DETECTION IN MULTI-MODAL OUTDOOR SURVEILLANCE

Neil M. Robertson, Jonathan Letham

Heriot-Watt University, Edinburgh, UK

ABSTRACT

In this paper we present a new approach to person detection in outdoor surveillance tasks. A multi-modal segmentation (RGB, Polarimetric, thermal sensors) of the world into regions sky, road, bush, trees, grass etc. is used to learn the normal spatial context of people appearing in normal training data. The context feature is a novel application of the work of Wolf *et al.* [1] which enables the probability of a person appearing in a certain location to be computed. By using motion as a precursor to the deployment of a HOG person detector in conjunction with the spatial context likelihood we obtain significant improvement in person detection for challenging scenes. Comprehensive ROC analysis on 4 outdoor scenes is reported for normal activity detection. Anomaly detection is then achieved using learned context and we show that 72% of true positive anomalies are found for a false positive rate of 0.19% over all data in thermal and visual band data.

Index Terms— Surveillance, Sensor fusion, person detection.

1. INTRODUCTION AND RELATED WORK

In surveillance tasks, it is important to be able to detect objects that are at times hard to find in varying environments and also to detect unusual behaviour. In a dynamically changing scene, it is often difficult to detect objects due to occlusion, shade and clutter. Context is being used increasingly in computer vision to help perform scene recognition [2], region categorisation [3] and object detection [4]. Towards the goal of performing situation awareness tasks, context may be used to minimise false alarms and for optimisation so that workload is managed. The novel use of context is presented in this work and is used to improve the accuracy of scene segmentation and algorithm deployment. We develop algorithms to automatically segment a scene into basic region types (road, sky, bush, trees and grass) and these extracted regions are then used to provide spatial context to enhance HOG-based pedestrian detection [17].

In this brief review we focus on the use of spatial context in visual processing, noting that there has been extensive work on pedestrian detection, tracking and scene segmentation reported. Olivia *et al.* argue that a scene composed of contextually-related objects is more than than just the sum of constituent objects [5]. Bar and Ullman showed that objects are faster to localise and recognise when presented in a familiar context [6]. Torralba presented a method for object localisation given scene recognition [7]. Murphy uses a combination of local image features and global features (the *gist*) to perform object detection and localisation [8], achieving improved detection rates. Keller and Wang were early proponents of learning spatial relations in computer vision [9] using neural networs to generalise spatial relationships from simple examples. Singhal uses contextual spatial relations to improve region classification [10]. Conditional random fields (CRF) have been used for this task also by Wojek and Schele [11]. Fei-Fei tackles scene recognition[12] using a Bayesian hierarchical model to learn and recognise natural scene categories. Heitz and Koller directly exploit regions in an image as context to improve object detection [4], improving object detection by using the surrounding context rather than blindly applying a sliding window approach across the whole image. Wolf and Bileschi also use image regions to provide context to improve object detection [1].

(a) (b) (c)

Fig. 1. Improvements to calculations of Q: (a) Image of the scene; (b) Old method of absolute Q computation; (c) Our method of Q computation. Note false polarisation detection is massively reduced.

In this work we make the following research contributions, which to our knowledge have not been presented in a real multi-modal surveillance system: 1. Using two different forms of context, prior and temporal, region classification is improved; 2. Multiple sensory data is fused via contextual smoothing after classification; 3. Motion and object detection are improved using an object-centric context feature.

(a) (b) (c)

Fig. 2. Process of sky classification: (a) Input RGB image. (b) Graph-based segmented result on RGB image. (c) Extracted sky region from the segmented image.

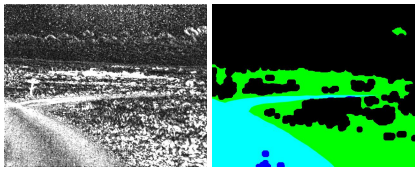
2. AUTOMATIC SCENE SEGMENTATION

Our trial vehicle has a top-mounted array of six visual cameras and six short-wave infra-red (SWIR) cameras. The six grey-scale visual cameras are sensitive to light of wavelength $450nm$, $500nm$, $550nm$, $650nm$, $700nm$ and $880nm$. The six grey-scale SWIR cameras are sensitive to light of wavelength $950nm$, $1050nm$, $1150nm$, $1250nm$, $1350nm$ and $1550nm$. We also use a Catherine MP thermal camera which operates in the $8-12\mu m$ range and is sensitive to polarised radiation. Radiation emitted or reflected from a surface can be described by four Stokes parameters I, Q, U and V. I is the total intensity of the radiation. P, highlights manmade objects, Q picks up horizontal and vertical polarised surfaces and U detects diagonal polarised surfaces. The Stokes parameters are defined as follows: $I = \frac{1}{2}(i_0 + i_{45} + i_{90} + i_{135})$, $Q = i_0 - i_{90}$, $U = i_{45} - i_{135}$, where the intensity, i_x, is measured on a pixel element with diffraction grating polariser oriented at x degrees to the horizontal. Using the Stokes parameters, the amount of polarisation, P, and the angle of polarisation, Φ, are calculated by: $P = \frac{\sqrt{Q^2+U^2}}{I}$, $\Phi = \frac{1}{2}tan^{-1}\frac{U}{Q}$

We briefly outline the classification step of the work (note this has been covered in detail in a previous article [13]). We aim to detect four regions: sky, road, foliage and the "other" class. Since we have strong prior knowledge about the scene, the sky is defined as the region above the horizon including clouds; road is any visible tarmac road; foliage is defined as green vegetation i.e. bush, tree or cut grass (BTG) in the image; and the other class is any pixels that do not fall into these regions. The four classifiers built to detect each of the described regions are outlined here.

Sky: The registered $450nm$, $550nm$ and $650nm$ images are used to create colour images. Felzenszwalb's graph-based image segmentation is then used to segment these RGB images [14]. The sky is then classified by taking the largest top region of the segmented image. It is acknowledged that this is heuristic assumption valid only for this data set. An example of this process is shown in Figure 2.

Road: The polarimetric data is exploited for road detection: as a smooth man-made object it has a strong polarisation signature. The Q Stokes parameter ($Q = i_0 - i_{90}$) is sensitive to polarised radiation in the horizontal direction [13]. Therefore, in our data, the Q data can be used as a road detector. An adaptive threshold is applied in order to binarise the image and then a median filter is applied to reduce false alarms as shown in Figure 3. Here the Q image is shown on the left and then the thresholded and median filtered road classification shown in green over the ground truth in blue on the right. The road detection method tends to be noisy with high false positives and is a good candidate for improvement which we achieve via temporal smoothing to reduce the false positives (vs ground truth) in road classification from 0.35 to 0.003, and the accuracy to 0.98 from 0.71.



(a) (b)

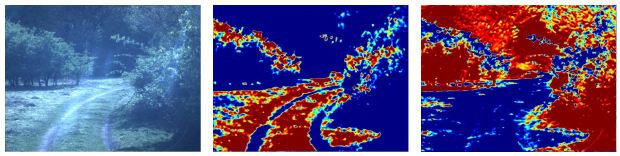
Fig. 3. Process of road classification: (a) Input Q image. (b) Median filtered and thresholded road classification shown in green, ground truth in blue. Correctly detected pixels are therefore cyan.

(a) (b) (c)

Fig. 4. Process of vegetation classification:(a) Input; (b) $NDVI$ image made using the $880nm$ (NIR) and $650nm$ (RED). (c) Classified vegetation.

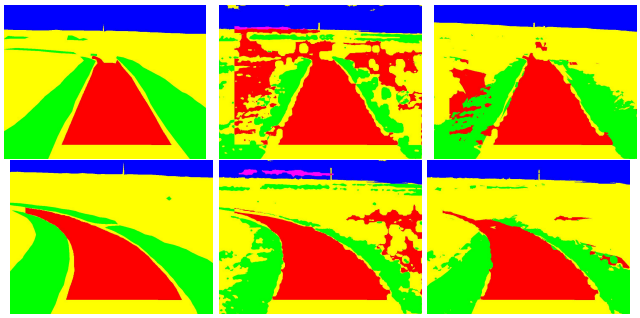
BTG: Live green plants absorb solar radiation in the photosynthetically active radiation (PAR) spectral region which includes the red waveband. They also scatter light in the near infra-red (NIR) region. Therefore, NIR and red wavebands can be combined in the Vegetation Index in order to highlight vegetation [15]. We use the Vegetation Index to classify bush, tree and cut grass. It is defined as $NDVI = \frac{NIR-RED}{NIR+RED}$, where NIR is the $880nm$ waveband of light (for these experiments), RED is the red band ($650nm$) and $NDVI$ is the normalised vegetation index. The output of the $NDVI$ has a median filter applied and is then thresholded to produce a binary image. The threshold used is an learned threshold found by testing over many images. An example of the vegetation classification is shown in Figure 4.

We may make a further refinement to the detection of vegetation to distinguish between and detect bush and grass. To discriminate between bush and grass in the highlighted region, a RGB image is converted to L*a*b*. When used in combination with the difference, NIR and red band image,



(a) (b) (c)

Fig. 5. Separation of the BTG class: (a) input RGB image; (b) examples bush classification; (c) example grass classification (red corresponds to likely and blue unlikely).



ground truth raw classes smoothed

Fig. 6. Temporally-smoothed segmentation, Context improves the segmentation vs. ground truth. Blue is sky, red is road, green is BTG and yellow is "other".

we may produce a probability per pixel for the bush and grass areas, illustrated in Figure 5.

Smoothing of the image segementation. The contextual framework used by Matzka is adapted to fit temporally smooth the raw classifier results [16]. (NB rather than using the Bayesian framework to calculate the probability that a detection of a vehicle is true given a particular road type, the probability that a region exists given a detection at a pixel is computed from prior and temporal knowledge.)

For evaluation, a training set of images are ground-truthed by hand so that the accuracy of the classifiers can be calculated. The classifiers were tested on these images and the true positive rates (TPR), false positive rates (FPR) and accuracies (Acc) of the detectors were computed over each image. These metrics are based on pixel measurements. For a particular classifier, the total number of true detected pixels are counted (true positive - TP), the number of missed pixels (false negative - FN), the number of wrongly detected pixels (false positive - FP) and the correct non-detections (true negative - TN). The metrics are defined as follows: $TPR = \frac{TP}{TP+FN}$; $FPR = \frac{FP}{FP+TN}$; $Acc = \frac{TP+TN}{TP+FN+TN+FP} = \frac{TP+TN}{Totalpixels}$

The medians of the results for all of the temporally-smoothed test data are shown in Table 1. Some classification image results are shown in Figure 6. These show the colour image, the ground truth and the detection results for two illustrative frames.

	Sky	Road	BTG	Other
Raw	0.96	0.90	0.71	0.76
Smoothed	0.99	0.93	0.87	0.89

Table 1. Accuracy of the individual classifiers, temporally smoothed, vs manually-acquired ground truth.

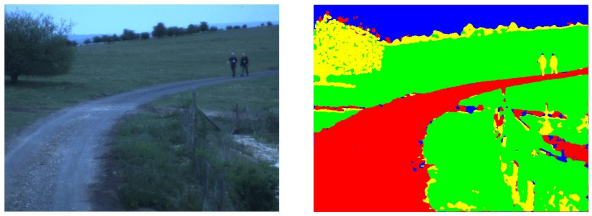


Fig. 7. Colour image (left) and corresponding segmented image (right). Red = road, blue = sky, green = sky, yellow = bush.

3. IMPROVING PERSON DETECTION USING SPATIAL CONTEXT

Having achieved a stable and accurate segmentation into basic background regions, we now show how extracted regions of an image can improve object detection. The goal is to enable context to improve the detection rate and false alarm rate of a detector. A further goal is to show context can be used to detect unusual behaviour. A spatial object-centric approach is taken: we look for how the regions relate to the object and learn likelihoods of appearance in regions. The regions extracted from the segmentation serve as spatial context, an example input is shown in Figure 7.

The context feature we propose is inspired by Wolf [1]. In order to gather information about the context surrounding an object of concern, we sample points at a pseudo logarithmic distance from the object. We sample in four directions (above, below, left, right) and at 4 pixel distances along those directions (1, 40, 100, 200). Given a bounding box of an object, the measure of pixel distance from the object begins from the edge of the box.

When an object is detected in an image, the surrounding regions (which have been detected in the scene segmentation) are sampled using this context feature and the region at each sample location recorded. Each sample location, l_k, has a learnt prior probability of expected region R_c at that location: $P(R_c|l_k)$, where c is one of the four region classes and k is one of the 16 sample locations. The prior probabilities $P(R_c|l_k)$ at each sample location are learnt normalised histograms for the four region classes. For example, the sample location 200 pixels above the object has learnt probabilities of expected regions (sky, road, grass or bush). If the segmented region sampled at this sample location is grass, then the probability learnt for grass is taken.

Probabilities at all other sample locations are similarly

recorded depending on the segmented region sampled. We now combine this information to find the probability for the object detected to exist given the surrounding context i.e what is the likelihood that an object would have the surrounding regions located at the sample points. This is simply the average of all of the recorded probabilities: the total probability of an object, O, to exist given the surrounding context, C, is given by, $P(O|C) = \frac{1}{n}\sum(P(R_c|l_k))$, where n is the number of sampled locations. The priors are learnt by using the manual ground truth data for labelled objects.

Person detection: We aim to improve a standard person detection algorithm by using the proposed context feature. Person detection is achieved using the HOG detector. The HOG detector used is the optimised detector proposed by Maji *et al.* [17]. We take their pre-trained classifier and test it on our data. The classifier is trained on colour images and we test it on visual and thermal data. The performance of the detector is scene dependent, depending on depth of scene, occlusion etc. Hence we split our analysis into the four scenes evident in our dataset, quoting aggregated results for the performance in the visual and thermal data for each of the four.

Spatial context and motion detection: To increase the detection rate of the detector, and also decrease the false positives, motion and context are used together. The motion detector and context feature are used in combination to give an indication of an area where an object is *expected* to be. The HOG detector is then applied to this region. To do this, if motion is detected with a context feature above the context threshold of $P(O|C) = 0.6$ (which is found to be optimum in training) the region surrounding this motion is considered as a candidate pedestrian area. If the motion is at distance from the camera, the area is magnified by 3x and the HOG detector is then run on the magnified area in order to detect pedestrians. Otherwise, the HOG detector is run on the inference area with no magnification.

To give an indication of whether the motion detected is far or near from the camera, additional range data is used derived from the hand-labelled pedestrian ground truth for a given scene. Given a vertical location in the image, the expected height of a pedestrian is known from this data. Motion is considered to be far from the camera when the height of the hand-labelled at the y-coordinate of the motion is less than 150 pixels and is considered near otherwise. Examples of this method working successfully to improve the detection performance in both modalities are shown in Figures 8 and 9. Improved results using this method compared to using the HOG detector alone are shown in the ROC plots in Figure 10.

Improvement in performance can be seen in the plots when the '$\circ$' plots for motion plus context are upwards and to the left of the '$*$' plots for no motion or context for the same scene (same scenes are plotted in the same colour). Again, it can be seen from the ROC plots that the results are scene dependant. It has been shown here that when the

Fig. 8. the HOG detector fails to detect one of the pedestrians in a visual image. Using motion plus context, area is zoomed into and HOG now detects pedestrian.

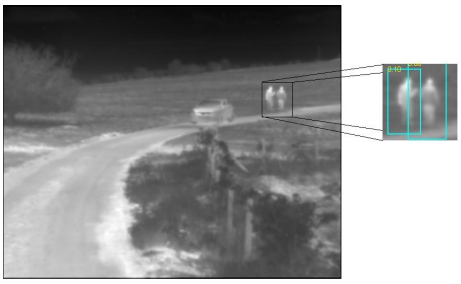


Fig. 9. The HOG detector fails to detect both of the pedestrians in a thermal image. Using motion plus context, the area is zoomed into and HOG now detects both pedestrians.

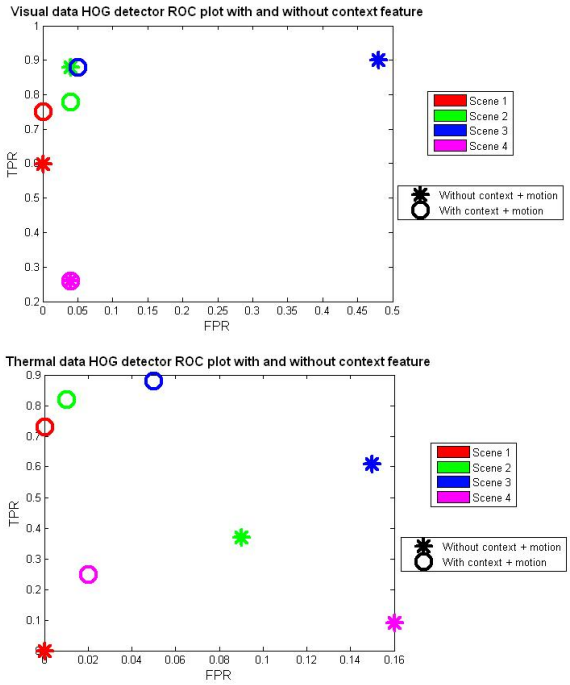


Fig. 10. Complete ROC plot for HOG detector on visual data (*top*) and thermal (*bottom*) with no context and with motion and context. Median results for all four scenes are shown independently.

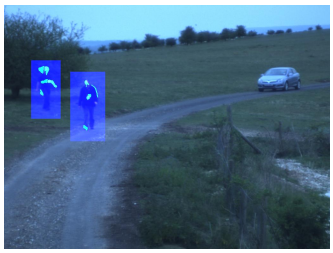


Fig. 11. Two person detections are found in this image, however, the contextual feature for each differs: the pedestrian on the left is considered to behaving unusually $P(O|C) = 0.52$ and the pedestrian on the right usually $P(O|C) = 0.67$.

	TP	FP
Visual	0.73	0.21
Thermal	0.70	0.18

Table 2. Combined results of anomaly detection in visual and thermal data over all scenes.

context feature is used with a motion cue, performance of the detector generally improves and in situation awareness tasks, the objects of interest will normally be moving.

Contextual anomaly detection: The contextual feature can also be used to detect unusual behaviour. Conversely to the normal detection, where the HOG detections are not considered to be true detections if the context feature computed falls below the context threshold, detection in an area where the object *is not expected* can be considered as anomalous behaviour. The threshold of $P(O|C) = 0.6$ is used, as before, and the prior probabilities are learnt from the expected behaviour of the pedestrian. Rather than considering motion in an area with a context feature above $P(O|C) = 0.6$, motion detected in an area *below* the threshold only is considered. The region is magnified if considered to be far form the camera. An example image of anomalous behaviour detection is shown in Figure 11. Overall results are given in Table 2.

4. CONCLUSION

We have presented a complete system for improving person detection using learned spatial object context which relies only on simple image classifiers. The approach is generic and not limited to HOG detection or scenes of the type shown here. Object/Person detector accuracy tends to be scene dependent and our technique may be used to mobilise prior knowledge of the scene to reduce false alarms and improve accuracy, as demonstrated, in such cases. Anomaly detection is a natural result of the learned spatial context which is very useful in visual situation awareness tasks. Future work is focussing on developing probabilistic scene segmentation, a real-time implementation using heterogeneous hardware (FPGA, GPU etc.) and integrating other types of object into the system (vehicles, for example).

5. REFERENCES

[1] Lior Wolf and Stanley Bileschi, "A critical view of context," *IJCV*, vol. 69, pp. 251–261, 2006.

[2] Aude Oliva and Antonio Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress in brain research*, vol. 155, pp. 23–26, 2006.

[3] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan, "Matching words and pictures," *Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.

[4] Geremy Heitz and Daphne Koller, "Learning spatial context: Using stuff to find things," *ECCV*, vol. 10, pp. 30–43, 2008.

[5] Aude Oliva and Antonio Torralba, "The role of context in object recognition," *Trends in cognitive sciences*, vol. 11, pp. 520–527, 2007.

[6] Moshe Bar and Shimon Ullman, "Spatial context in recognition," *Perception*, vol. 23, pp. 343–352, 1996.

[7] Antonio Torralba, Kevin P. Murphy, William T. Freeman, and Mark A. Rubin, "Context-based vision system for place and object recognition," *ICCV*, 2003.

[8] Daniel Eaton Kevin Murphy, Antonio Torralba and William Freeman, "Object detection and localization using local and global features," *Towards Category-Level Object Recognition*, vol. 1, 2005.

[9] James M. Keller and Xiaomei Xang, "Learning spatial relationships in computer vision," *Fuzzy Systems*, vol. 1, pp. 118–124, 1996.

[10] Amit Singhal, Jiebo Luo, and Weiyo Zhu, "Probabilistic spatial context models for scene content understanding," *Computer Vision and Pattern Recognition*, vol. 1, pp. 235–241, 2003.

[11] Christian Wojek and Bernt Schiele, "A dynamic conditional random field model for joint labeling of object scene and classes," *Lecture Notes in Computer Science*, vol. 5305, pp. 733–747, 2008.

[12] Li Fei-Fei and Pietro Perona, "A bayesian hierarchial model for learning natural scene categories," *Computer Vision and Pattern Recognition,*, vol. 2, pp. 524–531, 2005.

[13] Jonathan Letham, Neil Robertson, and Barry Connor, "Contextual smoothing of image segmentation," *Workshop on the use of context in video processing, CVPR*, 2010.

[14] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vision*, vol. 59, pp. 167–181, 2004.

[15] C. J. Tucker, "Red and photographic infrared linear combinations for monitoring vegetation," *Remote Sensing of Environment*, vol. 8, pp. 127–150, 1979.

[16] Yvan R. Petillot Stephan Matzka and Andrew M. Wallace, "Effcient resource allocation using a multiobjective utility optimisation method," p. 12, 2008.

[17] Subhransu Maji, Alexander C. Berg, and Jitendra Malik, "Classification using intersection kernel support vector machines is efficient," *Computer Vision and Pattern Recognition*, 2008.