Consistency in Multi-modal Automated Target Detection using Temporally Filtered Reporting

Toby P. Breckon[a] and Ji W. Han[a] and Julia Richardson[b]

[a]School of Engineering, Cranfield University, Bedfordshire, UK;
[b]Stellar Research Services, Southampton, UK

ABSTRACT

Autonomous target detection is an important goal in the wide-scale deployment of unattended sensor networks. Current approaches are often sample-centric with an emphasis on achieving maximal detection on any given isolated target signature received. This can often lead to both high false alarm rates and the frequent re-reporting of detected targets, given the required trade-off between detection sensitivity and false positive target detection. Here, by assuming that the number of samples on a true target will both be high and temporally consistent we can treat our given detection approach as a ensemble classifier distributed over time with classification from each sample, at each time-step, contributing to an overall detection threshold. Following this approach, we develop a mechanism whereby the temporal consistency of a given target must be statistically strong, over a given temporal window, for an onward detection to be reported.

If the sensor sample frequency and throughput is high, relative to target motion through the field of view (e.g. 25fps camera) then we can validly set such a temporal window to a value above the occurrence level of spurious false positive detections. This approach is illustrated using the example of automated real-time vehicle and people detection, in multi-modal visible (EO) and thermal (IR) imagery, deployed on an unattended dual-sensor pod. A sensitive target detection approach, based on a codebook mapping of visual features, classifies target regions initially extracted from the scene using an adaptive background model. The use of temporal filtering provides a consistent, fused onward information feed of targets detected from either or both sensors whilst minimizing the onward transmission of false positive detections and facilitating the use of an otherwise sensitive detection approaches within the robust target reporting context of a deployed sensor network.

Keywords: object recognition, temporal filtering, intelligent target reporting, thermal imaging, multi-modal object detection, cross-spectral surveillance, sensor networks, temporal fusion, optical thermal detection

1. INTRODUCTION

Autonomous target detection is an important goal in the wide-scale deployment of unattended sensors and sensor networks.[1] Although current approaches offer a significant reduction in the required manual processing or review of sensor information they can equally suffer from frequent false positive reporting and the subsequent trade-off required in detection sensitivity.[10] Furthermore without the explicit use of tracking multiple detection reports are often generated for a single target as it transits the sensor field-of-view. Both of these issues impact on the effective use of such systems by human operators and similarly on wide area multi-sensor target tracking approaches deployed across such a sensor network. Given the assumption of a sensitive yet noisy target detector – with high detection rates but significant false positive detections - we look here at how to can improve the consistency of the resulting global target/object reporting based on applying temporal filtering.

Whilst widely considered in sensor processing for image stabilization, de-noising and compression, temporal filtering has received little attention within autonomous target detection and more generally object detection and classification outside of explicit object tracking approaches.[7,10,12,22,24] Most studies in object detection are primarily focused on detection within a single image[10] and focus primarily on various approaches to identify the spatial consistency of the object feature signature within the image.[7,12] By contrast video based studies most often consider explicit spatio-temporal features,[23] commonly aimed at activity recognition tasks,[15] or the

Corresponding author: toby.breckon@cranfield.ac.uk

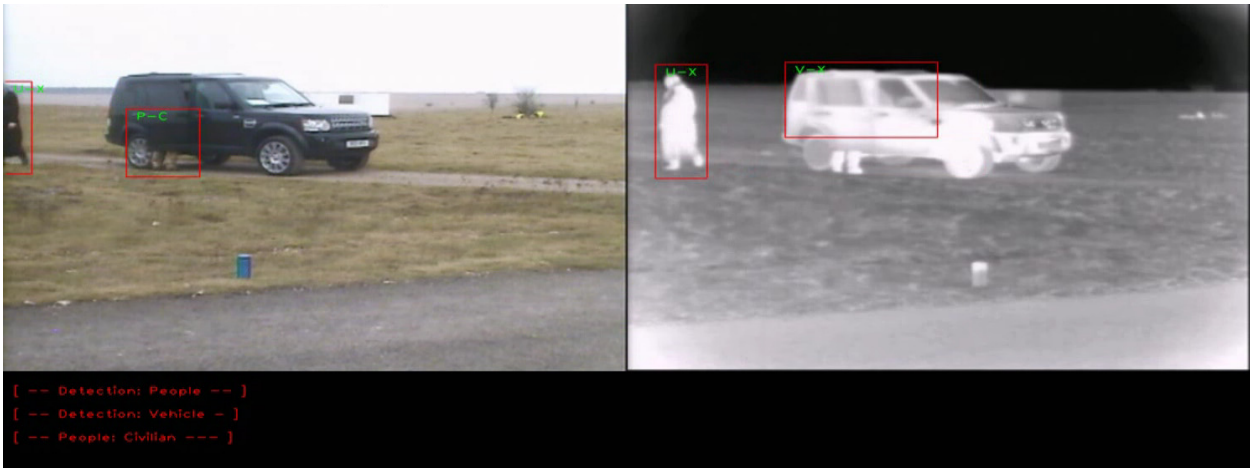

Figure 1. Example of target detection in multi-modal (visible/thermal) imagery

explicit task of object tracking.[17] The use of temporal filtering strategies to consider either the concatenation or fusion of multiple image samples is generally limited within the literature.[10,13] The majority of work considering temporal connectivity in object detection and classification concerns itself primarily with tracking. The tracking problem has a number of complexity constraints such as object occlusion, changing perspective and multiple object association.[21,28] For the task of the simple reporting of target detection at a given position, within the context of the sensor network envisaged, this is somewhat overly complex.[8,14]

In this work we look to investigate the results that can be achieved with a somewhat simpler temporal filtering approach and the resulting impact this has on the required detection approach in use for the task. Firstly we assume, that for a given sensor imaging a true target, the number of samples on target (i.e. images containing the target) will be numerous and temporally consistent (i.e clustered) as the target transits the scene. Subsequently, for a sensitive detection approach we further assume true positives are detected over numerous samples as a target transits the sensor field-of-view. By contrast we also assume, that for such a noisy detector, false positives may be frequent but that the temporal distribution of such false positives will be random. In this way we treat the set of classified image samples akin to a temporally distributed ensemble classifier. The assumption is that although noisy, our automated detection approach is a weak classification approach at best - i.e. at least better than random for true positive target detection. Using this as a base classifier for target detection, we introduce the use of temporally filtered reporting for final confirmation of target presence at the sensor location.

Over a given temporal window, detection of a given target must be statistically better than random for a detection report to be generated by the temporal target detection filter. If the sample rate of the sensor and throughput is high in relation to target motion through the field of view (e.g. 25fps video) then we can validly set the temporal window to a value above the background noise of spurious false positive detections. This can be further tuned to increase the robustness/confidence in reported detections by increasing the size of the temporal window. Furthermore, if we consider multiple modalities of sensor, each with a different likelihood characteristic for either true/false automated detections, this concept can be extended to consider contributing detections from different sensors or even different detection algorithms applied to the same sensor. This approach enables a reduction in both false positive reporting and multi-reporting of target detection events from a given sensor deployment within the network. As a result, the main contribution of this paper is to purport the use of temporal sampling strategies as a conduit to the use of weak or less than optimal target detection approaches within a deployed sensor context.

We illustrate this concept using results obtained from a feature-driven approach for vehicle and people detection in visible (EO) and thermal (IR) imagery from an unattended dual-sensor pod. We use a automated target detection approach based on a codebook of visual features, prone to false positive reporting when considered purely on a per-sample basis, classifying targets extracted from the scene using an adaptive background model. This is illustrated in Figure 1 where we see the detection both people and vehicles within the combined

multi-modal imagery and in addition summary detection reports based on a temporal filter applied over multiple image samples (Figure 1, lower).

2. APPROACH

Our approach is illustrated against the backdrop of a classical two stage automated visual surveillance approach of first detecting initial candidate regions within the scene (Section 2.1), thus facilitating efficient feature extraction over isolated scene regions, to which an identified target type is assigned via secondary object classification (Section 2.2). Whist traditionally such approaches can suffer from higher than acceptable false positive detection rates, when considered for use within a deployed long-term sensing context, we further detail our use of temporal sampling strategies to overcome these issues (Section 2.3).

2.1 Candidate Region Detection

In order to facilitate overall real-time performance, initial candidate region detection identifies isolated regions of interest within the scene. This allows subsequent feature extraction and classification to be performed over isolated region(s) enabling real-time processing. Additionally, this facilitates efficient object localization within the scene. By leveraging the stationary position of our multi-modal sensor pod, this is achieved using a Mixture of Gaussian (MoG) based adaptive background model.[16,29]

Following this approach each image pixel is modeled as a set of Gaussian distributions, commonly termed as a Gaussian mixture model, that capture both noise related and periodic (i.e. vibration, movement) changes in pixel intensity at each and every location within the image over time. This background model is adaptively updated with each frame received and each pixel is probabilistically evaluated as being either part of the scene foreground or background following this methodology.[29] Background modeling facilitates the automated identification of foreground regions within the incoming video imagery such that a) new objects can be isolated from the scene background for efficient feature generation and classification and b) objects that enter and become static within the scene (e.g. parked car) are adaptively learnt as part of the background model.

This concept is illustrated in Figure 2 where we see isolated scene regions relating corresponding to the vehicle and pedestrains entering the scene (Figure 2 A/B upper) in both the visible spectrum image (Figure 2 A/B left) and thermal image (Figure 2 A/B right). These scene regions (Figure 2 A/B upper) represent pixel values that do not fit the current adaptive background model and are thus identified as foreground pixels within the scene. This set of foreground pixels (Figure 2 A/B upper) is post-processed using morphological dilation and connected components analysis[25] facilitating the rejection of small noisy candidate regions prior to classification. It is notable that some small noise regions remain and overall noise remains in foreground object boundaries (Figure 2 A/B upper).

In Figure 2 A/B (lower) we see a visualization of the current background model in each instance for the optical and thermal images based on a weighted average of the current Gaussian mixture model at each scene pixel. Figure 2B also shows how the representation of the stationary vehicle that was present in the scene for some time (Figure 2A) is incorporated into the adaptive background model and subsequently removed as it departs the scene following the continuous updates to the adaptive background model. Traces of the vehicles prior position in the scene background are clearly visible in Figure 2B (lower) relating to its previous position in Figure 2A. Minor traces of the vehicles previous stationary position (maintained shortly after entering the scene) are also visible in Figure 2A (lower) but have subsequently been removed by updating in Figure 2B which is taken from later in the same sequence (once the vehicle has departed).

Overall, this background modeling approach facilitates the efficient identification of candidate regions for further feature extraction and classification (Section 2.2). As a by-product it readily facilitates the reporting of new, arriving and transiting/moving scene objects as scene events without continual re-reporting of stationary scene objects that were not originally present within the scene. The use of such adaptive background modeling techniques is commonplace in the automated visual surveillance and tracking literature.[16]

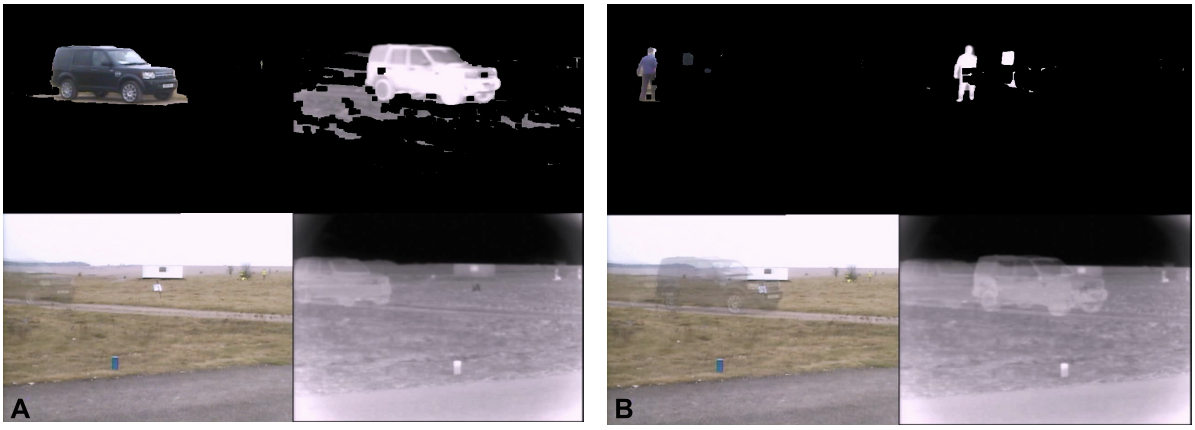

Figure 2. Candidate regions (upper) detected in both visible and thermal imagery against current MoG background model (lower)

2.2 Target Classification

Target classification follows a machine learning driven approach of off-line classifier training and on-line target classification using the trained classifier. Specifically we follow the bag of visual words methodology.[11,22,24] Following this approach abstract multi-dimensional visual features are extracted over the set of training examples. In general a wide range of features are available and widely utilised for this task[19,26] with variation in computational performance, complexity and invariance properties. Here we utilise the Speeded Up Robust Features (SURF) approach[2] that are both viable for real-time performance on full motion video and known to be suitable for cross-modal use as our base features.[3]

Following the bag of visual words (or codebook) methodology, we perform multi-dimensional feature clustering over all of the example training imagery (for all object classes) to produce a set of general feature clusters that characterise the overall feature space. This provides a fixed dimension set of cluster references for all target types and sub-types that we are to classify. Commonly this set of feature clusters is refereed to as a codebook or vocabulary as it is subsequently used to encode the features detected on specific object instances (positive and negative) as fixed length vectors for input to both the off-line training and on-line classification phase of later machine learning classification. A given object instance is encoded as a fixed length vector based on the membership of the features detected within the object to a given feature cluster based on nearest neighbour (hard) cluster assignment. Essentially the original variable number of SURF features detected over each training image or candidate region is encoded as a fixed length histogram representing the membership of these features to each of these clusters. This fixed length distribution of features is then used to differentiate between positive and negative instances of a given class based on a trained classifier.

Here we perform clustering using the common-place k-means clustering algorithm in 128-dimensional space (i.e. SURF feature descriptor length of 128^2) into 1000 clusters. Support Vector Machine (SVM)[5] and Decision Forest (DF)[4] classifiers are used in this work for the final classification phase. Separate classifiers are trained for vehicle and people detection in each of the imaging modalities (visible and thermal) with additional classifiers for the detection of related sub-categories (people: {military, civilian} and vehicles: {4x4, saloon}) acting on the visible imagery alone. Typically each classifier is trained over a set of images with 100-200 positive examples and several thousand representative negative examples.

2.3 Temporally Filtered Target Reporting

The contribution made here centres around the use of temporal filtering strategies in the onward reporting of automated target detections made in continuous video feeds using the approach outlined in Section 2.2. As we will aim to illustrate, this essentially this acts as an enabler for the use of weaker core classifiers in the detection task. A weak classifier is generally accepted as one where the output that is poorly correlated against the ground truth for a given classification task - in this case target detection from a codebook encoded distribution of features

(Section 2.2). By generality a *weaker* classifier is one that is perhaps less than optimal for a given class when used in a specific environment or circumstance.

A weak or relatively weaker classifier can be constructed in a number of ways. For example either by restricting the complexity of the classification approach (e.g. dimension, number of entities, parameter search space), the amount of training performed or the amount (and variety) of training data used. Essentially a weak classifier is constructed by somehow disobeying one of the established paradigms of good experimental practice in machine learning.[20] Despite the obvious disadvantages of a weak classifier, studies considering ensemble classifiers[4,6] have found the combined use of sets of multiple weak classifiers out perform approaches accepted to operate based on a globally optimal decision boundary (e.g. Support Vector Machine (SVM)). Current state of the art ensemble approaches[6,9] combine a set of such weak classifiers, often trained over varying random subsets of the training examples and/or utilizing a random subset of the available features, using either a bagging (equally weighted) or boosted (performance weighted) voting scheme to arrive a final classification from the set.

Here we extend this concept to the temporal domain. In our target detection task we recognise the high temporal frequency of our sensor (i.e. 25fps video) relative to the transit speed of potential targets through the sensor field-of-view (e.g. Figure 1). As a result, numerous "on-target" samples will be received for classification for a given true target and these will be temporally clustered (over a short time frame relative to the overall surveillance sensing task). For a target in motion, and similarly for a stationary target, the combination of changing scale, perspective onto target and noise will affect the feature distribution extracted from any candidate region in the image (extracted as per Section 2.1) that corresponds to a given target. This variation in features will in the large part be systematic due to changes in scale and perspective and in some part random due to noise. The overall structural variance due to systematic change and the statistical variance due to noise will follow a similar distribution to the encountered in the training data used to train the classifier for a given target type. By contrast, spurious candidate regions isolated in the scene due to noise in the adaptive background model (Section 2.1) should not share these characteristics and should thus be determinable as negative instances of a given target by the classifier. This is essentially the underlying theory of feature based object classification.

However, we must recognise that the perfect classification case of all such spurious candidate regions being correctly rejected as an instance of a given class is unlikely given the current state of the art in object classification.[7,10,12,22,24,27] As a result what we have can be deemed as a weak classifier by some measure of this concept. The contribution here is to then treat this classifier as a weak classifier over time. We accept false positive detections of a given class may occur and they may occur frequently. Similarly false negative (i.e. detection failure of a given class) may also occur frequently but given the number of samples on target their effect is less pronounced. As long as the classifier is truly weak in that its output is correlated, albeit poorly against true target detection, we can make use of it by a performing bagging or boosting approach over a given time (sampling) window. False positives or false negatives will be random and not temporally clustered over any significant sampling period. By contrast, true positive (i.e. correct) detections will be temporally clustered. This key observation allows us to proceed with a less than optimal (weaker) classifier approach and rely on the strength of an ensemble approach to facilitate the desired reduction in false positive reporting.

For a classifier detecting the presence of target class c based on a given feature distribution x, represented by a binary classification function $k_c(x) \in \{0, 1\}$, this results in a classification result integrated and normalized over a temporal window of w at time t as follows:

$$A_{k_c} = \frac{\sum_{t-w}^{t} k_c(x_t)}{w} \tag{1}$$

This results in a real-valued integrated classification in the range $\{0 \rightarrow 1\}$ which if treated akin to a probability can be considered to give a likelihood of detection. Applying a threshold to this parameter, τ_k, facilitates the translation of this to a detection report, $vehicle = \{yes|no\}$, for onward transmission within the sensor network. In essence, as Eqn. 1 gives equal weighting to each time step within the temporal window, this can be directly translated to state - *if target is detected greater than τ_k% of the time within w sequential image samples we then report it as a confirmed target detection.* In a practical sense, assuming a weak classifier that is significantly better than random, this can be readily implemented as - if a given target is detected greater than 50% of the time (i.e. better than random) over w sequential images we then report it as a detected target.

Target Type	Sensor Modality	%True Positive (TP)	%False Positive (FP)	Precision	Recall	Accuracy
people	visible	96.3%	8.1%	0.95	0.96	0.94
vehicle	visible	98.5%	6.6%	0.94	0.98	0.96
people	thermal	93.0%	13.2%	0.92	0.93	0.91
vehicle	thermal	100%	0.75%	0.93	1.00	0.99

Table 1. Validation results for SVM RBF classifiers on test imagery samples

Our formulation (Eqn. 1) can be further expanded as follows to consider a non-binary classification function, $f_c(x)$, as follows:

$$A_{f_c} = \frac{\sum_{t-w}^{t} f_c(x_t)}{\beta w} \tag{2}$$

where β is the maximal value of $f_c(x)$ and thus A_{f_c} can be treated analogous to A_{k_c}. In practice our non-binary classification function may itself return a probability directly from an underlying Bayesian classifier or perhaps the number of positive votes within a decision forest classifier (where $\beta = \#trees\, in\, forest$)[6] or the distance of the feature distribution instance from the decision boundary in an SVM (where $\beta = max(f_c(i))\, \forall\, i,\, i \in \{training\, examples\}$).[5]

Here, where multiple sensors are deployed in a given target detection task we can extend both approaches to accumulate contributions from each sensor at each time step and normalize appropriately for the number of sensors contributing to the over classification (e.g. two sensors for our dual sensor experimental setup, see Figure 1). In theory the performance of the classifier function need only be greater than 50% in terms of true positive detection to give an aggregated performance that is better substantially better than random but in practice we use a classification approach with significantly higher performance (~93-98% true positive, 6-13% false positive on the test examples used for training validation - see Table 1).

3. EVALUATION

Evaluation was carried out using Support Vector Machine (SVM) classifiers trained using a Radial Basis Function (RBF) kernel and cross-validation driven grid-based parameter search.[5] Separate binary classifiers were trained for each target type in each sensing modality (4 classifiers in total - one per modality per class type). These were trained using training image sets of approximately 100-200 positive example images and several thousand negative image examples. Validation testing was purposefully performed over separate test image sets of a similar size with the results, on per test image sample basis, reported in Table 1.

From Table 1 we can readily see this classification approach performs extremely well over the limited test set examples with high true positive detection and low false positive detections in almost all cases. This is similarly reflected in the associated precision, recall and accuracy statistics calculated for these approaches over this test set. However, these results are on a per image (i.e. per-sample) basis over test sets that, although typical for object detection/classification tasks,[10] are relatively small compared to the number of images frames received from a 25fps video feed over a long duration surveillance task. A reported 6-8% false positive rate could directly translate as 6-8 false positive target detection reports 100 video frame images - i.e. 6-8 false positive detection reports every 4 seconds for 25fps video. Under these conditions the classifier could be considered as less than optimal (or weak) under our earlier definition despite the statistical results shown in Table 1.

These individual SVM classifiers are used to form the input to temporally filtered target reporting. Classifier results are combined on a per target type basis, from both sensors, into a single accumulated target classification following the approach outlined in Section 2.3. The reporting threshold was set as $\tau_k = 0.5$ following a simple better than random bagging paradigm with the time window w varied as set out in the results of Table 2. Separate window periods, $w_{vehicles}$and w_{people} are outlined for each of the target types under consideration, recognizing *a priori* differing likely transition speeds through the sensor field of view.

From the results shown in Table 2 we can observe the resulting number of generated target detection reports against the ground truth number of target events present within three surveillance sequences (A,B,C). These

	Ground Truth		Vehicles($w_{vehicles}$)				People (w_{people})		
Sequence	Vehicle	People	1	5	10	50	1	25	50
A (Fig. 3)	9	6	87	22	5	4	39	5	4
B (Fig. 4)	0	8	71	18	0	0	163	12	12
C (Fig. 5)	12	0	72	13	6	1	150	7	0

Table 2. Temporal filtering results for target detection reporting

surveillance sequences are 1-2 hours in duration over the environment shown in Figures 1, 3 - 5. The number of ground truth targets is defined on a per event basis - i.e. each occurrence of a target of the specified type entering the sensor field of view and either leaving or becoming stationary for a significant period. A stationary object leaving the scene is counted as a separate target event, requiring detection, for the purposes of ground truth accounting. This methodology for the treatment of stationary objects follows inherently from the MoG background modeling approach used to generate candidate region in the first stage of target detection (see Section 2.1).

For comparison, temporal filtering of target detection reporting is shown under a range of window parameter settings including for $w = 1$ (Table 2) which essentially gives us the total number of reported detection without any temporal filtering being performed (see Eqn. 1). As can be see from the results (Table 2) the number of reports without any temporal filtering is extremely high in comparison to the number of ground truth target events occurring within the scene. This high rate of target detection reports is directly attributable to the re-reporting of targets as multiple occurrences due to the lack of any temporal cohesion and the false positive rate incurred from direct application of the SVM classifiers on a per-sample (i.e. per video image frame) basis (see Table 1). A general trend can be observed that the number of reported targets is significantly reduced as the size of the temporal window increases and more closely resembles the level of ground truth target events (Table 1). The number of false positive reports is reduced as is the number multiple many-to-one reports. Although the number of both remains high relative to the ground truth, it has been significantly reduced from a much higher level of reporting over this elongated period. The exact setting required appears to vary per target and is dependent on the detection sensitivity required.

We further illustrate the use of this technique in Figures 3 - 5 which show still image examples taken from the test sequences A-C respectively. In all of these figures each candidate region identified from the earlier MoG background modeling (Section 2.1) is shown with a bounding red box. Within each of these highlighted regions, green text in the top right corner identifies the classification of this region within this image sample using a set of abbreviated encodings as *{V- = vehicle, P- = people, U-=unknown}* where *U-* is a region identified as neither vehicle or person. These per-image region classifications feed into the temporal reporting structure with the current temporally filtered target detection reports displayed in red/green text below each optical/thermal image pair. It is the change in state of these target detection reports which is collated in Table 2. Within these examples the detected vehicle and people targets are also sub-classified as $\{4x4|regular\}$ or $\{military|civilian\}$using separate classification approach operating on the pre-classified candidate regions of the visible band video imagery (indicated as subtype classifications in Figures 3 - 5). This subtype technique, making use of decision forest classification[6] over the same feature distribution, will be fully reported in future work.

In Figure 3 we see the successful detection of a vehicle partially entering the scene (top) including the rejection of several other candidate regions (top right, thermal image, foreground), the subsequent reporting of the same vehicle as it moves again to the centre of the field of view and the additional identification of people within the scene (including in front of the stationary vehicle). Notably, the stationary vehicle is added to the background model after a period of time (Section 2.1) which allows detection of additional foreground objects (people, Figure 3 lower) within this region. Figure 4 shows the successful identification of people in the environment and here we also see the fused contribution of both sensing modalities to the over temporal target detection reporting. A consistent state of target presence reporting is maintained despite a break in the per-sample detection in the thermal imagery (Figure 4, middle). Figure 5 similarly shows vehicle detection for a vehicle transiting and re-transiting the environment generating a separate, isolated report for each transit. In addition to these figures, illustrative video examples are provided as supplementary material from:
`http://www.cranfield.ac.uk/~toby.breckon/demos/multimodaldetection/`

Figure 3. Sequence A: people and vehicles sequence

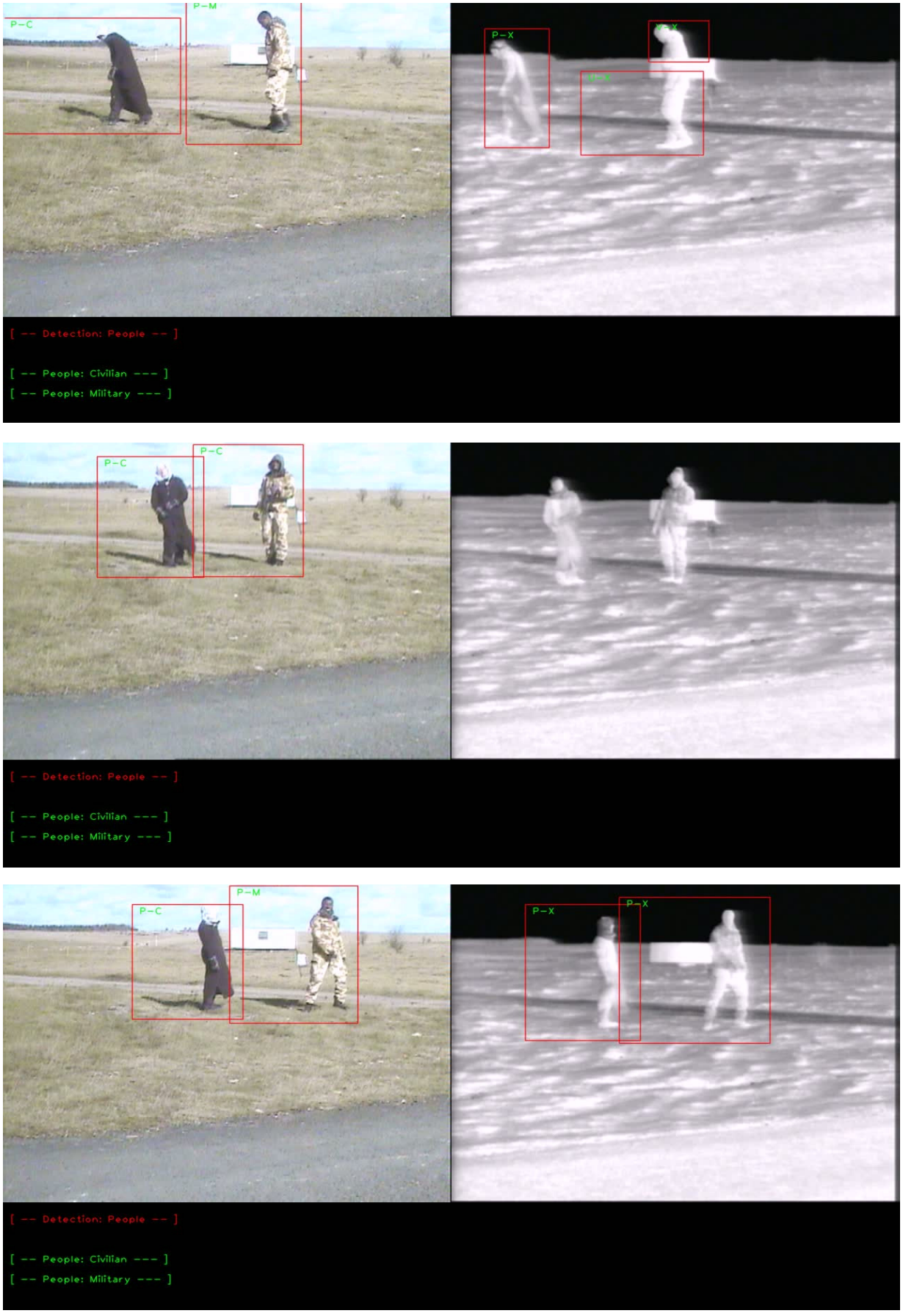

Figure 4. Sequence B: people sequence

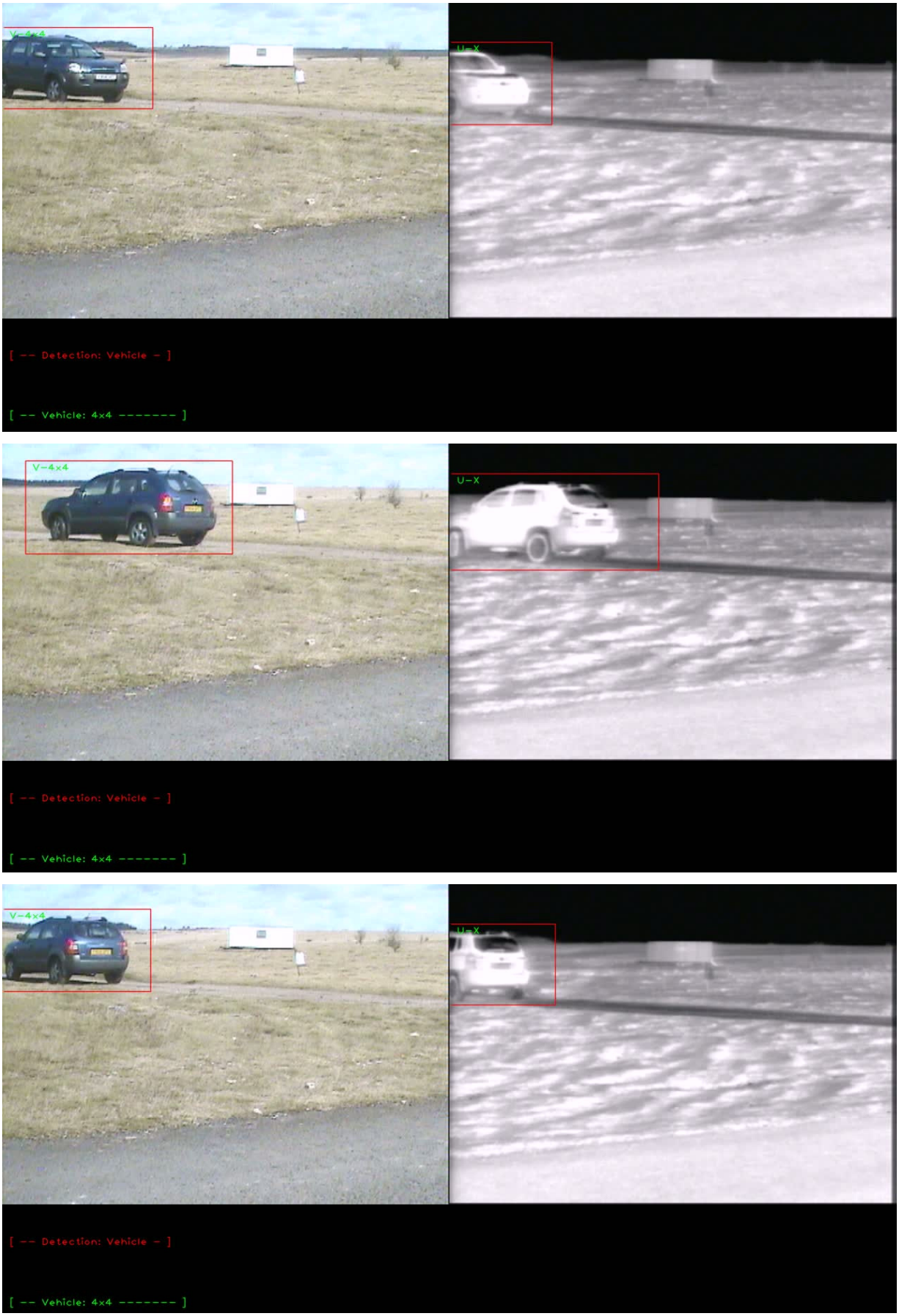

Figure 5. Sequence C: vehicles sequence

4. CONCLUSIONS

Overall we have shown that the impact of a simple temporal filtering approach for target detection reporting within a deployed sensing context is significant. It reduces both target re-reporting and false positive detections whilst additionally facilitating the use of a weaker underlying core classification approach with greater viability in the type of deployed sensing context typical of sensor network applications. Furthermore it provides a framework for the integration of multi-modal target detection information at the post-classification stage in a methodology that is robust to sensor drop-out type failure in any one of these modalities. This is illustrated through the use of a bag of visual words approach for the detection of people and vehicles from a dual visible/thermal video sensor pod deployed as an unattended sensor in a representative environment. The use of temporally filtered reporting is shown to significantly reduce spurious false positive and re-reporting within this context.

Future work will look to investigate the use of recent advances in real-time salient object detection[18] for initial candidate region detection and the relationship between the concepts presented here and the episodic evaluation criteria proposed in[14] for wide-area search and surveillance.

ACKNOWLEDGMENTS

This work was funded by the UK MoD Centre for Defence Enterprise (CDE) (research contract number: 11657) and carried out in collaboration with the Defence Science and Technology Laboratory (DSTL).

REFERENCES

1. H K Aghajan and A Cavallaro. *Multi-camera networks: principles and applications*. Academic press, 2009.
2. H Bay, A Ess, T Tuytelaars, and L Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.
3. B Besbes, A Rogozan, and A Bensrhair. Pedestrian recognition based on hierarchical codebook of SURF features in visible and infrared images. In *2010 IEEE Intelligent Vehicles Symposium*, pages 156–161. IEEE, June 2010.
4. L Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001.
5. C J C Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.
6. A Criminisi, J Shotton, and E Konukoglu. Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. Technical report, Tech. Rep. MSR-TR-2011-114, Microsoft Research, Cambridge, 2011.
7. N Dalal and B Triggs. Histograms of Oriented Gradients for Human Detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 886–893. IEEE.
8. S Denman, T Lamb, C Fookes, V Chandran, and S Sridharan. Multi-spectral fusion for surveillance systems. *Computers & Electrical Engineering*, 36(4):643–663, 2010.
9. T G Dietterich, P Domingos, L Getoor, S Muggleton, and P Tadepalli. Structured machine learning: the next ten years. *Machine Learning*, 73(1):3–23, August 2008.
10. M Everingham, L Van Gool, C K I Williams, J Winn, and A Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, September 2009.
11. L Fei-Fei and P Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *Computer Vision and Pattern Recognition*, volume 2, pages 524–531. IEEE, 2005.
12. P F Felz, R B Girshick, D McAllester, and D Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (99):1–1, 2010.
13. A García-Martín, J M Martínez, and J Bescós. A corpus for benchmarking of people detection algorithms. *Pattern Recognition Letters*, 33(2):152–156, January 2012.
14. A Gaszczak, T P Breckon, and J W Han. Real-time people and vehicle detection from UAV imagery. In *Proc. SPIE Conference Intelligent Robots and Computer Vision XXVIII: Algorithms and Techniques*, page Vol. 7878 Number 78780B, January 2011.

15. A Gilbert, J Illingworth, and R Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *12th International Conference on Computer Vision*, pages 925–931. IEEE, September 2009.
16. D Hall, J Nascimento, P Ribeiro, E Andrade, P Moreno, S Pesnel, T List, R Emonet, R B Fisher, J S Victor, and J L Crowley. Comparison of target detection algorithms using adaptive background models. In *Proc. 2nd Joint IEEE Int. W'shop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 113–120, 2005.
17. Z Kalal, K Mikolajczyk, and J Matas. Tracking-Learning-Detection. *IEEE transactions on pattern analysis and machine intelligence*, 34, December 2011.
18. I Katramados and T P Breckon. Real-time Visual Saliency by Division of Gaussians. In *Proc. International Conference on Image Processing*, pages 1741–1744, September 2011.
19. K Mikolajczyk, T Tuytelaars, C Schmid, A Zisserman, J Matas, F Schaffalitzky, T Kadir, and L Van Gool. A Comparison of Affine Region Detectors. *International Journal of Computer Vision*, 65(1):43–72, 2005.
20. T M Mitchell. *Machine Learning*. McGraw-Hill, March 1997.
21. J J Pantrigo, J Hernández, and A Sánchez. Multiple and variable target visual tracking for video-surveillance applications. *Pattern Recognition Letters*, 31(12):1577–1590, September 2010.
22. J Philbin, O Chum, M Isard, J Sivic, and A Zisserman. Object Retrieval with Large Vocabularies and Fast Spatial Matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
23. W Ren, S Singh, M Singh, and Y Zhu. State-of-the-art on spatio-temporal information-based video retrieval. *Pattern Recognition*, 42(2):267–282, February 2009.
24. J Sivic, B C Russell, A A Efros, A Zisserman, and W T Freeman. Discovering Object Categories in Image Collections. In *Proceedings of the International Conference on Computer Vision*, 2005.
25. C J Solomon and T P Breckon. *Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab*. Wiley-Blackwell, 2010.
26. T Tuytelaars and K Mikolajczyk. Local Invariant Feature Detectors: A Survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280, January 2007.
27. P Viola and M J Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, May 2004.
28. A Yilmaz, O Javed, and M Shah. Object tracking - a survey. *ACM Computing Surveys*, 38(4):13–es, December 2006.
29. Z Zivkovic and F van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters*, 27(7):773–780, 2006.