Real-time classification of vehicles by type within infra-red imagery

Mikolaj E. Kundegorski, Samet Akçay, Grégoire Payen de La Garanderie, Toby P. Breckon
School of Engineering and Computing Sciences, Durham University, UK
E-mail: toby.breckon@durham.ac.uk

ABSTRACT

Real-time classification of vehicles into sub-category types poses a significant challenge within infra-red imagery due to the high levels of intra-class variation in thermal vehicle signatures caused by aspects of design, current operating duration and ambient thermal conditions. Despite these challenges, infra-red sensing offers significant generalized target object detection advantages in terms of all-weather operation and invariance to visual camouflage techniques. This work investigates the accuracy of a number of real-time object classification approaches for this task within the wider context of an existing initial object detection and tracking framework. Specifically we evaluate the use of traditional feature-driven bag of visual words and histogram of oriented gradient classification approaches against modern convolutional neural network architectures. Furthermore, we use classical photogrammetry, within the context of current target detection and classification techniques, as a means of approximating 3D target position within the scene based on this vehicle type classification. Based on photogrammetric estimation of target position, we then illustrate the use of regular Kalman filter based tracking operating on actual 3D vehicle trajectories. Results are presented using a conventional thermal-band infra-red (IR) sensor arrangement where targets are tracked over a range of evaluation scenarios.

Keywords: vehicle sub-category classification, thermal target tracking, bag of visual words, histogram of oriented gradient, convolutional neural network, sensor networks, passive target positioning, vehicle localization

1. INTRODUCTION

We address the problem of the real-time classification of vehicles into sub-category types within infra-red imagery. Due to the high levels of intra-class variation in thermal vehicle signatures caused by aspects of design, current operating duration and ambient thermal conditions this poses a significant challenge. However, aspects of all-weather operation, invariance to visual camouflage techniques and accepted suitability for the the analogous task of pedestrian detection make sensing within thermal-band infra-red (IR) imagery very attractive.

Within the context of automated visual surveillance from infra-red imagery, our prior work on pedestrians [1, 2] demonstrated that reasonable performance can practically be achieved through the combined use of infra-red imagery (thermal-band, spectral range: 8-12μm) and the application of real-time photogrammetry. A key advantage of such thermal-band infra-red (IR) imagery for pedestrian localization is robust detection of human shape signatures within the scene [3–5]. As such, the principles of photogrammetry can be used to recover 3D pedestrian position within the scene based on a known camera projection model and an assumption that variance in human height is in fact quite small (statistically supported by [6, 7]). In [1] we experimentally investigated the accuracy of classical photogrammetry, within the context of current target detection and classification techniques [3–5], as a means of recovering the true 3D position of pedestrian targets within the scene. A real-time approach for the detection, classification and localization of pedestrian targets via thermal-band (infra-red) sensing was presented with supporting statistical evidence underpinning the key photogrammetric assumptions. Subsequent work in [2] explicitly addressed the remaining issue of correcting for pedestrian posture variation within this localization context. By contrast, here we present an approach for the automatic classification of vehicles by sub-type, such that a similar photogrammetric localization and tracking strategy can be employed. Identifying vehicle sub-type is a key governing factor in determining the suitable height assumption for use in such photogrammetric localization (supported by prior work of [8]) in addition to providing a higher granularity of target reporting within a deployed multi-sensor network.

Optics and Photonics for Counterterrorism, Crime Fighting, and Defence XII, edited by
D. Burgess, G. Owen, H. Bouma, F. Carlysle-Davies, R. J. Stokes, Y. Yitzhaky, Proc. of SPIE
Vol. 9995, 99950T · © 2016 SPIE · CCC code: 0277-786X/16/$18 · doi: 10.1117/12.2241106

Overall, despite extensive work in ground-based sensor networks [9–12], the use of photogrammetry within this context has received only limited attention [1, 13, 14]. The visible-band work of [14] uses a similar approach within a Bayesian 3D tracking framework but does not explicitly address issues of accuracy or its use within a detection filtering framework such as [1].

Prior work on vehicle type classification is dominated by work in visible-band imagery [15] where colour and texture features most often provide the primary conduit to classification by vehicle type [16, 17] and often make/model [18]. Recent work uses a range of feature driven classification approaches [16, 17] and the topic is well established within the domain of urban traffic surveillance [15]. Within this context, the consistency of vehicle appearance (shape outline, colour, texture) albeit under varying illumination conditions is a contrasting challenge to the thermal variations within our task. However, in many contexts it is desirable to perform both pedestrian detection, to which thermal-band sensing is highly suited, and vehicle detection/classification from the same deployed all-weather sensor, operating passively with strong invariance to visual camouflage, within a wider automated surveillance sensor network [19].

Prior work explicitly dealing with thermal-band (IR) imagery within an automated surveillance context is presently largely focused upon pedestrian detection [3, 5, 20–22] and tracking [23, 24]. The work presented in this paper is a direct extension of [1, 2] that demonstrated photogrammetric pedestrian localization within thermal-band imagery incorporating a lightweight tracking solution akin to that of [4]. Building directly on this framework presented in [1], here we present a method that additionally facilitates the passive localization of vehicles within thermal-band (IR) imagery based on prior classification vehicle type. Specifically we evaluate a range of feature detector/descriptor combinations with a traditional feature-driven bag of visual words architecture (akin to [1]), the use of histogram of oriented gradient features within a similar classification framework (building on [2]) and finally two modern convolutional neural network (CNN) architectures (AlexNet [25], GoogLeNet [26]).

A number of classification approaches are compared for this challenging subcategory classification task with results presented within a wider context of photogrammetric target localization and tracking as an enabler to spatio-temporal target reporting in an operational context [19, 27].

2. APPROACH

Our approach is illustrated against the backdrop of classical two stage automated visual surveillance [1]. First we detect 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) [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 facilitating localized feature extraction and classification. By leveraging the stationary position of our sensor, this is achieved using a combination of two adaptive background modeling approaches [28, 29] working in parallel to produce a single robust foreground model over varying environmental conditions and notably within varying ambient thermal/infra-red illumination conditions within complex, cluttered environments.

Within the first model, a Mixture of Gaussian (MoG) based adaptive background model, 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 [28, 30]. 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. The second model comprises the use of Bayesian classification in a closed feedback loop with Kalman filtered predictions of foreground component position [29]. Within this model, each pixel is similarly probabilistically classified as either foreground or background but this is further reinforced via Kalman predictions for the positions of foreground objects (i.e. connected component foreground regions [31]) present in the previous time-step. This object-aware model significantly aids in the recovery of fast moving foreground objects under varying illumination conditions such as the thermal gradients inherent within infra-red imagery. Overall this

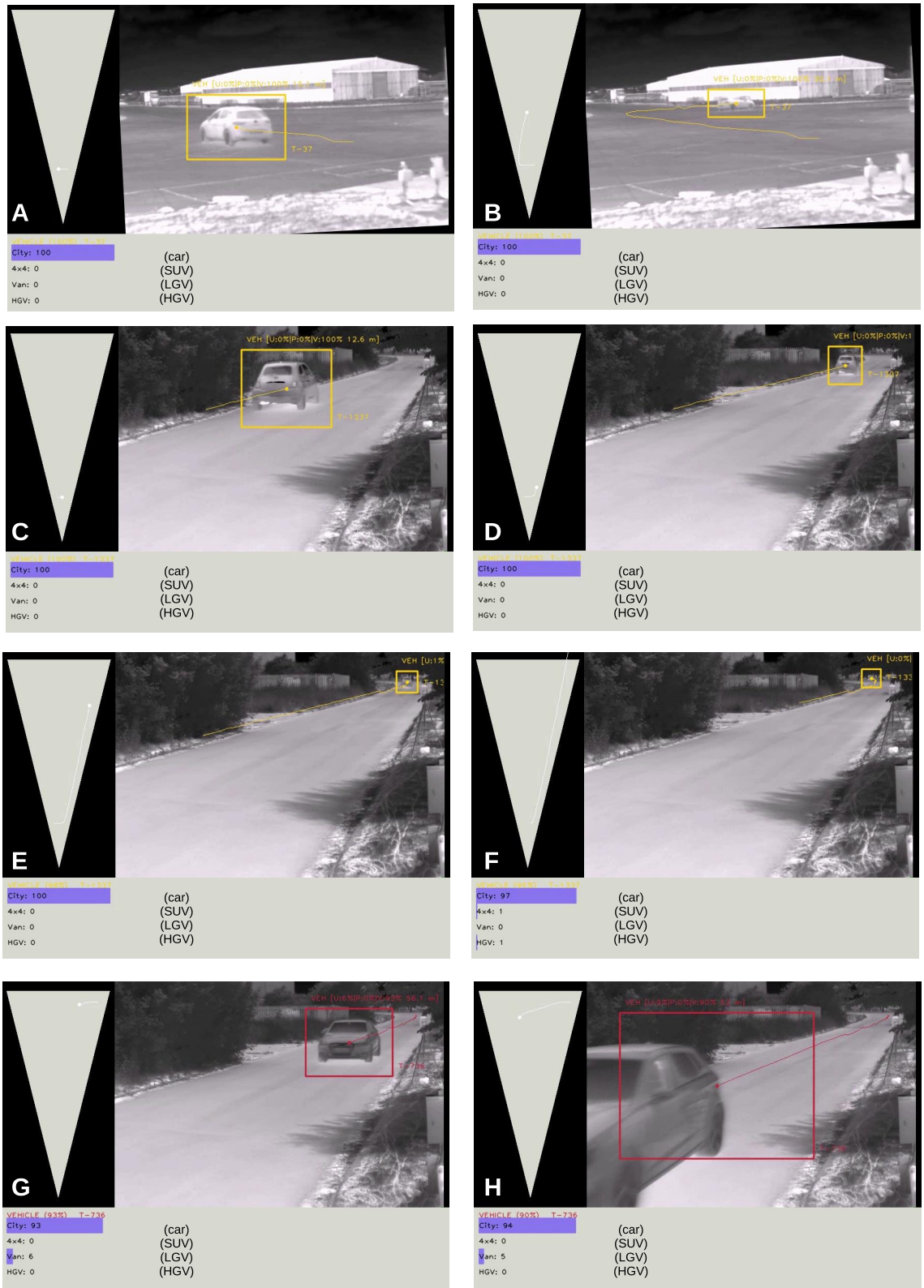

Figure 1. Examples of real-time vehicle detection, type classification and tracking in infra-red imagery with associated geo-referenced 3D track (based on using HOG features with SVM classification).

combined approach provides a slowly-adapting background model in the traditional sense [28], that can be robust to rapid illumination gradients, whilst similarly providing foreground consistency to fast moving scene objects [29]. The binary output of each foreground, based on a probabilistic classification threshold, is combined conjunctively to provide robust detection of both static and active scene objects. For illustrative examples and further discussion the reader is directed to [1].

2.2 Vehicle Classification

We evaluate several variations for the initial vehicle target classification (i.e. vehicle vs. non-vehicle) and subsequent type classification as one of $type = \{car, SUV, LGV, HGV\}$ such that *car* covers city/small/family type saloon cars, Sports Utility Vehicle (SUV) covers conventional (4×4) capable (including pickups) and similarly styled vehicles, Light Goods Vehicle (LGV) covers vans (including small wheelbase trucks) and Heavy Goods Vehicle (HGV) covers articulated trucks (lorries). Based on our detected candidate region (from Section 2.1), we specifically evaluate the use of sparse feature point descriptors using a bag of visual words encoding (Section 2.2.1), the use of dense Histogram of Oriented Gradient (HOG) features (Section 2.2.2) and an end-to-end deep Convolutional Neural Network (CNN) based on the use of transfer learning [25, 32].

2.2.1 Bag of Visual Words

Following a bag of visual words (or code-book) methodology [33], which has been empirically shown to be suited towards thermal infra-red imagery [1, 3, 22, 34], we evaluate a number of feature point detection and descriptor approaches as multi-dimensional features. Image representations based on local feature descriptors are widely applied in image classification and object recognition frameworks due to their robustness to partial occlusion and variations in object layout and viewpoint. Distinctive features of objects are detected at interest point locations which generally correspond to local maxima of a saliency measure calculated at each location in an image. The intensity patterns around these interest points are encoded using a descriptor vector. The most widely followed work in the area of local feature extraction has been Lowe's method of the Scale Invariant Feature Transform (SIFT) [35] which introduced a feature descriptor that is invariant to translation, scale and rotation and robust to image noise. Bay *et al.*'s later work [36] proposed the Speeded Up Robust Features (SURF) algorithm for feature detection and description that is loosely based on SIFT. The computational cost associated with SIFT is dramatically reduced without significant deterioration in performance (as used in prior work on infra-red pedestrian detection [1–3, 22, 34]).

More recently, research in this area led to industrious efforts to optimize sparse feature stability against computational performance leading to a range of local feature and detector variants. A standalone feature detector FAST (Features from Accelerated Segment Test) [37] provides significant number of candidate points for extraction while maintaining low computational cost. The detector-extractor frameworks BRIEF (Binary Robust Independent Elementary Features) [38], and BRISK (Binary Robust Invariant Scalable Key-points) [39] offer integer-space representations, avoiding the floating point operation of earlier SURF/SIFT variants, for faster extraction and subsequent computation on embedded platforms. ORB [40] (Oriented FAST and Rotated BRIEF) extends such methods to address issues of rotation invariance. A recent pairing of floating-point and integer space feature frameworks KAZE [41] and AKAZE [42] aim to improve uniqueness and robustness of features by describing them based on a non-linear model of an image. More recently FREAK (Fast Retina Key-point) [43], following from the earlier DAISY [44], represent feature extractors specifically inspired by retinal sampling in the human visual system originally designed for multiple image matching (i.e. image registration, stereo matching and alike).

Following the bag of visual words methodology, we perform feature extraction and clustering over all of the example training imagery (for all object classes) to produce a set of general feature descriptor clusters that characterise the overall feature space. Commonly this set of feature clusters is referred to as a code-book or vocabulary as it is subsequently used to encode the features detected on specific object instances (vehicle or non-vehicle) as fixed length vectors for input to both the initial off-line classifier training and on-line classification phase of such machine learning driven classification approaches. Here we perform clustering using the common-place k-means clustering algorithm in N-dimensional space (e.g. SURF feature descriptor length, $N = 128$ [36]) into k_v clusters. A given object instance is encoded as a fixed length vector based on the membership

Figure 2. Training data examples - vehicle types {*car*, *SUV*, *LGV*, *HGV*}

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 features detected over each training image or candidate region is encoded as a histogram, of fixed length k_v representing the membership of these features to each of these clusters. This fixed length distribution of features forms a feature vector that is then used to differentiate between labeled instances of a given class based on a trained classifier. Specifically we evaluate a range of such feature point detection and descriptor approaches of varying complexity for this task (namely: FREAK [43], DAISY [44], BRISK [39], ORB [40], KAZE [41], AKAZE [42]) against the mainstay of prior work in the field (i.e. SIFT [35] / SURF [36]) with the default parameter settings from the original works. From this bag of visual words feature encoding of feature descriptors, we have an overall feature vector, $\overrightarrow{v}_{BoVW}$ of dimension k_v (the number of visual code words used in our earlier bag of visual words vocabulary) which forms the input to our classification approach (Section 2.2.3).

2.2.2 Histogram of Oriented Gradient

The Histogram of Oriented Gradient (HOG) feature descriptor [45] is based on histograms of oriented gradient responses in a local region around a given pixel of interest. A rectangular block, pixel dimension $b \times b$, is first divided into $n \times n$ (sub-)cells and for each cell a histogram of gradient orientation is computed (quantised into H histogram bins for each cell, weighted by gradient magnitude). The histograms for all cells are then concatenated and normalised to represent the HOG descriptor as a feature vector, $\overrightarrow{v}_{HOG}$, for a given block (i.e. associated pixel location). For image gradient computation centred gradient filters $[-1, 0, 1]$ and $[-1, 0, 1]^T$ are used as per [45]. To construct our HOG descriptor, the localized vehicle region (from Section 2.1) is first zero-padded to form a square image region and subsequently re-sampled to a uniform 128×64 pixel image size, $(h \times w)$. We then compute the global HOG descriptor of this localized region using a block stride, $s = 8$ ($H = 9$, $n = 2$, $b = 16$ from [45]), to form a 3780 dimensional vector, $\overrightarrow{v}_{HOG}$, (i.e. $H \times n^2 \times (\frac{h}{s} - 1) \times (\frac{w}{s} - 1)$) as an input to subsequent classification (Section 2.2.3).

2.2.3 Feature Classification

Support Vector Machine (SVM) [46] classifiers are trained using each of these feature vector representations (Sections 2.2.1 and 2.2.2) over a corpus of exemplar imagery (see training examples in Figure 2). We use 7122 images for initial vehicle target classification (i.e. vehicle vs. non-vehicle) and 3410 images for subsequent type classification as one of *type* = {*car*, *SUV*, *LGV*, *HGV*} using a randomized 66% to 33% training set to validation set split for training and validation. The distribution of vehicle types within the data set used is {*car* = 2158, *SUV* = 315, *LGV* = 431, *HGV* = 506} representing a basic *a priori* likelihood of occurrence. SVM are trained using Radial Basis Function (RBF) kernel $\{SVM_{RBF}\}$ with a grid search over kernel parameter, $\gamma = 2^x : x \in \{-15, 3\}$, and model fitting cost, $c = 2^x : x \in \{5, 15\}$, using $k-$fold cross validation ($k = 5$). The results for the best performing parameter set are reported for each feature configuration in Section 3. Two stages of classification are performed on each feature vector:- primary classification as $\{vehicle, non - vehicle\}$, with non-vehicle encompassing both pedestrians (as per [1, 2]) and other scene objects, then secondary vehicle type classification as one of *type* = {*car*, *SUV*, *LGV*, *HGV*}. Examples of the training images used for this task are shown in Figure 2.

2.2.4 Convolutional Neural Network

Motivated by the work of [25] and current trends in convolutional neural networks (CNN), we evaluate a full CNN pipeline for this task. Unlike traditional feature-driven approaches (Section 2.2.1 & 2.2.2) that rely on a secondary stage of generic classification (Section 2.2.3) (so called *"shallow architectures"*), we employ a CNN approach for the entire feature extraction, representation and classification process (denoted as *"deep architectures"*). More specifically, with the use of a transfer learning approach [47], we optimize the CNN structures designed by Krizhevsky *et al.* [25] and Szegedy *et al.* [26] by fine-tuning its convolutional and fully-connected layers for the full end-to-end feature extraction to classification pipeline within this problem domain.

Unlike the traditional neural networks with conventionally one or two hidden layers, modern CNN can include many more hidden layers [26, 48, 49] comprising varying characteristics: convolutional layers (feature extraction), fully connected layer (intermediate representation), pooling layer (dimensionality reduction) and non linear operators (sigmoid, hyperbolic functions and rectified linear units). This complex of parametrization, and hence representational capacity, make CNN susceptible to over-fitting in the traditional sense. To overcome this issue, a number of techniques are employed to ensure generality of the learned parameterization of the target problem. Within the network, convolutional layers are usually interleaved by pooling layers which down-sample the current representation (image) and hence reduces the number of parameters in-addition to improving overall computational efficiency. Furthermore the use of drop out, whereby hidden neurons are randomly removed during the training process, and shared weights are used to avoid over-fitting such that performance dependence on individual network elements is reduced in favor of collective error reduction. In addition, with the use of the generalized technique called transfer learning, initial CNN parameterization (training) towards a generalized object classification task can then be further optimized (fine tuned) towards a domain specific classification task.

Presently, such CNN are designed manually with the resulting parametrization of the networks performing training using a stochastic gradient descent approach with varying parameters such as batch size, weight decay, momentum and learning rate over a huge data set (typically 10^6 in size). Current state of the art CNN models as such designed by Krizhevsky *et. al.* [25], Zeiler and Fergus [50], Szegedy *et. al.* [26], Simonyan and Zisserman [49] are trained on a huge data-set such as ImageNet [51] which contains approximately a million of data samples and 1000 distinct class labels. However, the limited applicability of such training and parameter optimization techniques to problems where such large data sets are not available gives rise to the concept of transfer learning [52, 53]. The work of [54] illustrated that that each hidden layer in a CNN has distinct feature representation related characteristics among of which the lower layers provide general features extraction capabilities (akin to Gabor filters and alike), whilst higher layers carry information that is increasingly more specific to the original classification task. This finding facilitates the verbatim re-use of the generalized feature extraction and representation of the lower layers in a CNN, whilst higher layers are fine tuned towards secondary problem domains with related characteristics to the original. Using this paradigm, we can leverage the *a priori* CNN parametrization of an existing fully trained network, on a generic 1000+ object class problem (from [55]), as a starting point for optimization towards to the specific problem domain of limited vehicle type classification. Instead of designing a new CNN with random parameter initialization we instead adopt a pre-trained CNN and fine tune its parameterization towards our specific classification domain. Specifically, we make use of the CNN configuration designed by Krizhevsky *et al.* [25], having 5 convolutional layers, 3 fully-connected layer with ∼60 million parameters, ∼650,000 neurons, and trained over the ImageNet data set on an image classification problem in the ILSVRC-2012 competition (denoted as AlexNet). We also employ the network structure proposed by Szegedy *et al.* [26], which won the ILSVRC 2014 competition (denoted as GoogLeNet). This second network is designed using many more layers (22) but with 12 times fewer network parameters compared to AlexNet to reduce the computational complexity of training a wide and deep network, while achieving promising performance results. Their approach is to first convolve each input by 1×1, 3×3, 5×5 filters in parallel (named as the inception module) to perform dimensionality reduction before being fed into subsequent more computationally expensive convolutional layers. From this point we then perform the fine-tuning (transfer learning) approach to both networks to train over the infra-red vehicle type data set (as detailed in Section 2.2.3) using backpropagation via stochastic gradient descent [25].

2.3 Photogrammetric Position Estimation

Firstly, we present a brief recap of our baseline localization approach as presented in [1] and subsequently show how this can be extended to address type variation within detected vehicle targets.

Based on automated detection (Section 2.2), target position is initially known within *"sensor space"* (i.e. pixel position within the image). Consequently, target position is estimated based on the principles of photogrammetry together with knowledge of the perspective transform under which targets are imaged and an assumption on the physical (real-world) dimension of a target in one plane [1]. All targets are imaged under a standard perspective projection [31] as follows:

$$x = f\frac{X}{Z},\ y = f\frac{Y}{Z} \tag{1}$$

where real-world object position, (X, Y, Z), in 3D scene co-ordinate space is imaged at image pixel position, (x, y), in pixel co-ordinate space for a given camera focal length, f. We assume both positions are the centroid of the object with (x, y) being the centre of the bounding box, of the image sub-region, for a target (object) detected in the scene (Section 2.1, e.g. Figure 1).

With knowledge of the camera focal length, f, the original object (target) position, (X, Y, Z), can be recovered based on (assumed) knowledge of either object width, $\triangle X$, or object height, $\triangle Y$ (i.e. the difference in minimum and maximum positions in each of these dimensions for the object). From the bounds of the detected targets (Section 2.2) we can readily recover the corresponding object width, $\triangle x$, and object height, $\triangle y$, in the image. Based on this knowledge, re-arranging and substituting into Eqn. 1 we can recover the depth (distance to target, Z) of the object position as follows:

$$Z = f'\frac{\triangle Y}{\triangle y} \tag{2}$$

Knowing Z via Eqn. 2, we can now substitute back into Eqn. 1 and with knowledge of the object centroid in the image, (x, y), we can recover both X and Y resulting in full recovery of real-world target position, (X, Y, Z), relative to the camera. In Eqn. 2, f' represents focal length, f, translated from standard units, mm, to focal length measured in pixels:-

$$f' = \frac{width_{image} \cdot f}{width_{sensor}} \tag{3}$$

where $width_{image}$ represents the width of the image (pixels), $width_{sensor}$ represents the camera digital (CCD) sensor width (mm).

Crucially, if we now assume a fixed width, $\triangle X$, or height, $\triangle Y$, for our object we can recover complete 3D scene position relative to the camera. For vehicle targets we can assume an average height for a given vehicle type determined from earlier vehicle type classification (as projected vehicle height,$\triangle y$, does not varying with viewing angle of the vehicle in the plane). Despite commonly held beliefs, empirical study has shown height variation within a given type classification of vehicle to be minimal [8]. In this study we use $\triangle Y = \{height_{car}, height_{SUV}, height_{LGV}, height_{HGV}\}$ for $\{height_{car} = 1.5m,\ height_{SUV} = 1.8m,\ height_{LGV} = 2.1m,\ height_{HGV} = 2.9m\}$ based on statistical evaluation of a moderate pool of vehicles. Following in a similar vein to the argument presented in [1] with regard to human height for pedestrians, this translates into a Z position error, attributable to vehicle height variation within a given type class, that is within GPS error tolerances ($\pm 5m$, [56]) for at least ranges up to 60m from the sensor.

2.4 3D Tracking

Unlike conventional tracking approaches that track 2D position, (x, y), within the image itself [57], our photogrammetric recovery of target position within the scene, (X, Y, Z) (Section 2.3) facilitates 3D tracking within scene space. This can be accomplished as tracking *"within the plane"* based on horizontal target position within the scene, X, and distance to target, Z, or full 3D scene space tracking including target elevation (vertical position), Y.

		SVM_{RBF}, $k_v = 500$					SVM_{RBF}, $k_v = 1000$					SVM_{RBF}, $k_v = 1500$					SVM_{RBF}, $k_v = 2000$				
Detector	Descriptor	TP	FP	P	A	F	TP	FP	P	A	F	TP	FP	P	A	F	TP	FP	P	A	F
SURF [36]	SURF [36]	81.8	18.0	0.81	0.82	0.81	83.8	19.0	0.80	0.82	0.82	85.3	23.7	0.77	0.81	0.81	82.0	15.3	0.83	0.83	0.83
SIFT [35]	SIFT [35]	86.6	17.5	0.82	0.85	0.84	86.3	19.5	0.80	0.83	0.83	88.7	18.9	0.81	0.85	0.85	91.2	18.4	0.82	0.86	0.86
ORB [40]	ORB [40]	66.2	19.6	0.76	0.74	0.71	69.8	18.9	0.77	0.76	0.73	68.1	20.1	0.76	0.74	0.72	69.3	20.4	0.76	0.75	0.72
KAZE [41]	KAZE [41]	87.3	12.4	0.87	0.88	0.87	85.4	15.8	0.83	0.85	0.84	87.0	12.6	0.86	0.87	0.87	89.2	13.6	0.86	0.88	0.88
FAST [37]	SURF [36]	89.4	13.8	0.86	0.88	0.88	89.6	15.6	0.84	0.87	0.87	89.7	14.5	0.85	0.88	0.87	89.1	15.7	0.84	0.87	0.87
FAST [37]	SIFT [35]	**95.8**	**7.5**	**0.92**	**0.94**	**0.94**	**96.6**	**8.9**	**0.91**	**0.94**	**0.94**	**95.6**	**8.5**	**0.91**	**0.94**	**0.93**	**95.9**	**8.8**	**0.91**	**0.94**	**0.93**
FAST [37]	ORB [40]	87.4	16.9	0.83	0.85	0.85	87.9	16.5	0.83	0.86	0.85	86.5	17.4	0.82	0.84	0.84	88.0	14.0	0.85	0.87	0.87
FAST [37]	FREAK [43]	65.7	20.8	0.75	0.73	0.70	65.9	24.0	0.72	0.71	0.69	64.0	24.1	0.71	0.70	0.67	62.2	19.6	0.75	0.72	0.68
FAST [37]	DAISY [44]	91.2	15.2	0.85	0.88	0.88	94.2	12.4	0.88	0.91	0.91	94.2	11.2	0.89	0.91	0.91	93.1	9.9	0.90	0.92	0.91
FAST [37]	BRISK [39]	84.9	12.4	0.86	0.86	0.86	85.5	10.9	0.88	0.87	0.87	83.7	11.8	0.87	0.86	0.85	84.5	12.0	0.87	0.86	0.86
BRISK [39]	BRISK [39]	78.7	18.2	0.80	0.80	0.79	75.8	17.5	0.80	0.79	0.78	77.5	18.2	0.80	0.80	0.79	79.3	15.9	0.82	0.82	0.81
AKAZE [42]	AKAZE [42]	44.5	10.6	0.80	0.68	0.57	47.0	9.9	0.81	0.69	0.60	45.3	7.8	0.84	0.70	0.59	48.5	9.7	0.82	0.70	0.61
		SVM_{RBF}																			
HOG [45]		**97.9**	**2.2**	**0.98**	**0.98**	**0.98**															
		AlexNet [25]					GoogLeNet [26]														
CNN		**99.9**	1.3	0.99	0.99	0.99	99.7	**1.0**	**0.99**	**0.99**	**0.99**										

Table 1. Results of feature and classification variants for primary vehicle classification.

For each candidate region identified as a new foreground object (Section 2.1), we initially created a new 2D track-let based on localized frame to frame connectivity derived from sparse optic flow [58, 59]. If one of the frame samples for this object is subsequently classified as vehicle (via the approach outlined in Section 2.2), this target transitions from a 2D tracked instance within image space to a 3D tracked vehicle within scene space. The tracked position, based on photogrammetric position recovery (Section 2.3) can then be propagated, over earlier instances of the same object similarly transitioning the motion history of this instance from 2D image position to 3D scene position. If an identified foreground object is not classified as being a vehicle its tracking remains within 2D image space until either its spatio-temporal filtered classification returns a vehicle classification (as per [1, 3] or it leaves the scene. Tracking within 3D scene space is performed using Kalman filter based tracking [60] on either a state vector comprising position and velocity *"within the plane"*, $\vec{s} = (X, Z, vX, vZ)^T$, or within $\mathbb{R}^3$ scene space, $\vec{s} = (X, Y, Z, vX, vY, vZ)^T$. Scene and measurement noise within the Kalman formulation are estimated empirically.

3. EVALUATION

Our results are presented using both quantitative measures of classification accuracy (Table 1 and 2) and qualitative assessment classification performance over a range of exemplar scenarios (Figures 1, 4). All evaluation imagery is captured using an un-cooled infra-red camera (*Thermoteknix Miricle 307k*, spectral range: 8-12μm) with statistical performance measured using validation test set of 2351 vehicle/non-vehicle images and 1126 vehicle sub-type images drawn from the same variation and environmental conditions as used for training (random 33% validation, as detailed in Section 2.2.3). Evaluation was performed around a variety of urban/industrial (cluttered) and suburban environments as part of work carried out in [19]. Within the feature detector, descriptor and classification variants outlined, we consider the comparison of True Positives Rate (TP), False Positives Rate (FP) (as percentages) together with the Precision (P), accuracy (A) and F-score (F) (harmonic mean of precision and true positive rate) for primary vehicle target classification (Table 1) and mean average precision (mAP) (mean of precision across all possible class labels) in addition to both mean accuracy (A) and F-score (F) for secondary vehicle type classification ($type = \{car, SUV, LGV, HGV\}$, Table 2).

From Table 1 we can see that CNN offer the best performance for primary vehicle classification (GoogLeNet, F-score of 0.993 and FP of only 1.0% followed closely by AlexNet with slightly higher FP, 1.3%). Traditional HOG with SVM classification also gives very strong results (F-Score of 0.98, FP of 2.2%) with the best bag of visual words approach (FAST feature detection with (slow) SIFT feature descriptor) coming in 4% lower across

		SVM_{RBF}, $k_v = 500$			SVM_{RBF}, $k_v = 1000$			SVM_{RBF}, $k_v = 1500$			SVM_{RBF}, $k_v = 2000$		
Detector	Descriptor	mAP	A	F	mAP	A	F	mAP	A	F	mAP	A	F
SURF [36]	SURF [36]	0.67	0.74	0.54	0.66	0.75	0.55	0.68	0.75	0.57	0.64	0.74	0.54
SIFT [35]	SIFT [35]	0.59	0.71	0.50	0.63	0.73	0.53	0.63	0.72	0.53	0.67	0.74	0.56
ORB [40]	ORB [40]	0.36	0.40	0.34	0.38	0.46	0.37	0.40	0.44	0.37	0.36	0.43	0.35
KAZE [41]	KAZE [41]	0.72	0.78	0.63	0.71	0.79	0.65	0.72	0.78	0.65	0.64	0.75	0.61
FAST [37]	SURF [36]	0.52	0.69	0.49	0.53	0.69	0.52	0.50	0.68	0.49	0.53	0.70	0.50
FAST [37]	SIFT [35]	**0.78**	**0.84**	**0.74**	**0.83**	**0.86**	**0.75**	**0.80**	**0.85**	**0.75**	**0.84**	**0.87**	**0.78**
FAST [37]	ORB [40]	0.43	0.50	0.42	0.41	0.51	0.41	0.41	0.49	0.41	0.39	0.43	0.37
FAST [37]	FREAK [43]	0.33	0.32	0.28	0.31	0.32	0.26	0.33	0.32	0.28	0.33	0.33	0.28
FAST [37]	DAISY [44]	0.61	0.71	0.49	0.71	0.74	0.54	0.66	0.76	0.61	0.66	0.74	0.54
FAST [37]	BRISK [39]	0.38	0.46	0.38	0.39	0.47	0.38	0.38	0.44	0.37	0.39	0.45	0.38
BRISK [39]	BRISK [39]	0.34	0.38	0.31	0.35	0.39	0.33	0.38	0.43	0.36	0.37	0.40	0.34
AKAZE [42]	AKAZE [42]	0.35	0.31	0.27	0.36	0.34	0.30	0.40	0.33	0.32	0.41	0.38	0.35
		SVM_{RBF}											
HOG [45]		**0.94**	**0.93**	**0.88**									
		AlexNet [25]			GoogLeNet [26]								
CNN		0.85	0.89	0.83	**0.94**	**0.95**	**0.92**						

Table 2. Results of feature and classification variants for vehicle type classification.

all vocabulary sizes (k_v) explored. The next best bag of visual words approach, FAST feature detection with DAISY feature descriptor, gives a 2.5% lower score despite the density of DAISY features. It can be observed that variation in vocabulary size generally appears to make negligible difference to performance.

From Table 2 we can see that the more difficult task of recognizing vehicle sub-types leads to a greater spread of performance between varying approaches. Again, we see that CNN offer the best performance (GoogLeNet, mAP of 0.94 / accuracy of 0.95) but that traditional HOG with SVM classification (mAP of 0.94 / accuracy of 0.93 / F-score of 0.88) outperforms the CNN AlexNet architecture (mAP of 0.85). However, all three approaches (GoogLeNet, HOG-SVM and AlexNet) significantly outperform the best bag of visual words approach (FAST feature detection with SIFT feature descriptor, mAP of 0.78). Within this bag of visual words approach (and some other) we can see that increasing vocabulary size appears to make notable difference to performance. The normalized inter-class confusion matrices presented in Figure 3 show the greatest cross-label confusion for the {*SUV*, *LGV*} type vehicles against the *car* vehicle type and additionally between the *LGV* and *HGV* vehicle types with the CNN approach (Figure 3 left) notably outperforming the HOG-SVM combination (Figure 3 right) in these cases.

Overall we see the prevalence of dense features (i.e. CNN, HOG) over the traditional bag of visual words approaches for these two classification tasks with the best performing bag of visual words approach also using FAST feature detection which is known to produce a higher density of feature points within the image. Within the two stage automated visual surveillance framework used here (Section 2.1), with features extracted only within the isolated candidate regions of the scene, all are achievable within the bounds of real-time operation [19, 27] (∼10fps+) based on CPU computation for the bag of visual words / HOG techniques and GPU-based computation for CNN based techniques.

This quantitative statistical evaluation (Table 1 and 2) is further supported by the qualitative results presented in Figures 1, 4-6 which illustrate extracts from vehicle type classification using HOG features with SVM classification and subsequent tracking sequences (using only the CPU computation available with the deployed sensor nodes [27]). These images are sequentially sub-sampled from the test scenarios with tracking and spatio-temporal detection performed as outlined in [1]. Within each sub-figure (Figures 1, 4-6 A-H) we present the detected vehicle(s) using a bounding box, associated 2D image projection of the track (A-H insets, right), the planar view of the $\{Y/Z\}$ tracked position relative to the camera (A-H insets, left) and the resulting temporally filtered vehicle type classification distribution (A-H, inset bottom).

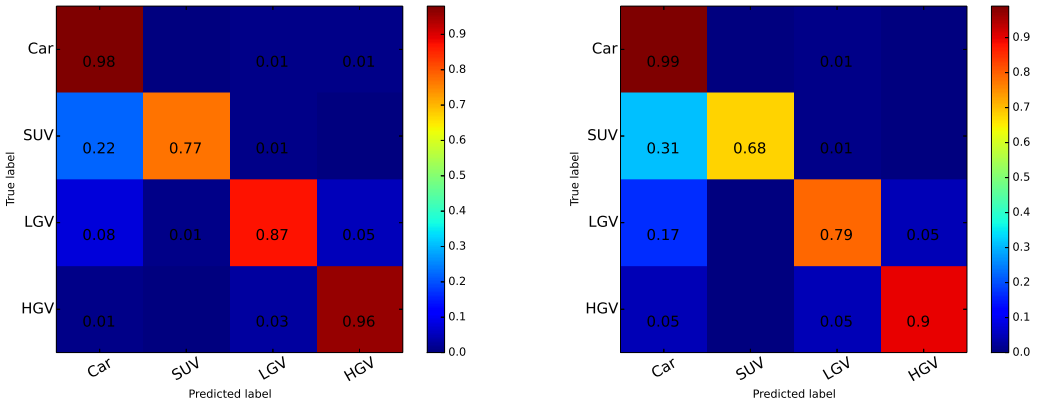

Figure 3. Normalized inter-class confusion matrices for vehicle type classification for both CNN (GoogLeNet, left) and HOG features with SVM classification (right).

From Figures 1 and 4 we can see that the accuracy and continuity of the $\{Y/Z\}$ position localization of the vehicle from standard photogrammetric techniques [1] (shown in A-H left, Figures 1 & 4) is consistent over varying vehicle types. Variation in vehicle viewing angle to the sensor in Figure 1 (e.g. transitions A →B, C →F and G →H) and Figure 4 (e.g. transitions A →D, E →F) show no significant erroneous jumps in the spatial locality of vehicle target when the planar view of the $\{Y/Z\}$ tracked position history is considered. This is further illustrated in Figures 5 and 6 where we see two sequences of consistent HGV type vehicle tracking from differing viewpoints (Figure 5, transitions A →E, F →H) and consistent tracking of a larger HGV over an extended distance including change in viewpoint (Figure 5, transitions A →E, F). As shown in Figure 4 (transition E →F, G) and Figure 6 (G, H) vehicle type miss-classification (confusion) largely occurs between the $\{car, SUV, LGV\}$ vehicle types dependent on viewpoint and distance to target in the scene. Intra-class variation between these classes is clearly visible in the vehicle configurations of Figure 4 (transition E →F, G) and Figure 6 (G, H) where the configuration of the vehicle type is either ambiguous due to viewpoint (Figure 4) or unusual to any such vehicle type (Figure 6).

Overall, our use of a vehicle classification by type is shown to facilitate effective compensation for variations in both vehicle dimension and viewing angle for the purposes of photogrammetric based localization (Figures1, 4-6). Under evaluation conditions GPS accuracy locally was found to be $\pm 5m$, based on a consumer GPS unit [56] and secondary verification of vehicle position from a concurrently deployed active range sensor [61] (as part of [27]) showed the photogrammetric localization recovered to be within this bound in the majority of test cases.

4. CONCLUSIONS

Overall we have shown that the use of Convolutional Neural Networks (CNN) or Histogram of Oriented Gradient (HOG) feature based classification facilitate the most effective determination of vehicle type to enable improved 3D localization and tracking within infra-red imagery based on the principles of photogrammetry. This directly advances the generality of prior work in field for pedestrian localization in the presence of posture variation [1–3] by additionally facilitating vehicle localization from the same infra-red sensing modality within a deployed sensor network [19]. Within the context of passive target localization in infra-red thermal imagery, and the general use of passive sensing for geo-located target tracking in wide-area sensor networks [19], this work similarly extends the argument in favour of passive sensor utilization within the bounds of acceptable accuracy. This is supported by a strong statistical evaluation over a number of variations on current state of the art classification approaches with CNN and HOG features outperforming traditional bag of visual words based approaches for this task. This work further strengthens the application of generalized target tracking within 3D scene-space that facilitates the ready disambiguation of multiple target tracking scenarios using low-complexity approaches with reduced computational overheads [1]. Our approach is demonstrated over multiple scenarios in cluttered environments where a clear capability in vehicle type classification is clearly illustrated as an enabler to the passive localization of vehicles.

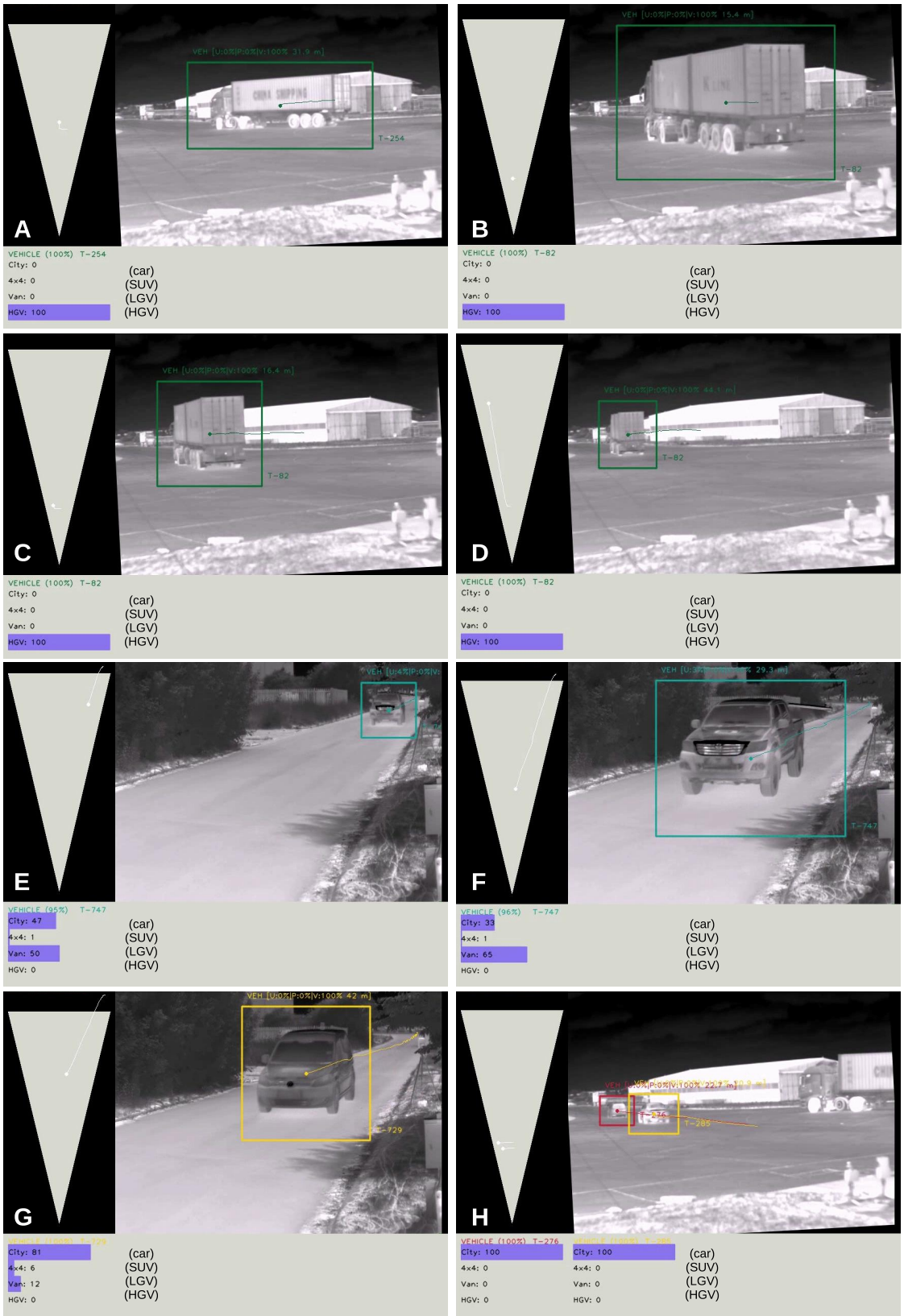

Figure 4. Examples of real-time vehicle detection, type classification and tracking in infra-red imagery with associated geo-referenced 3D track (based on using HOG features with SVM classification).

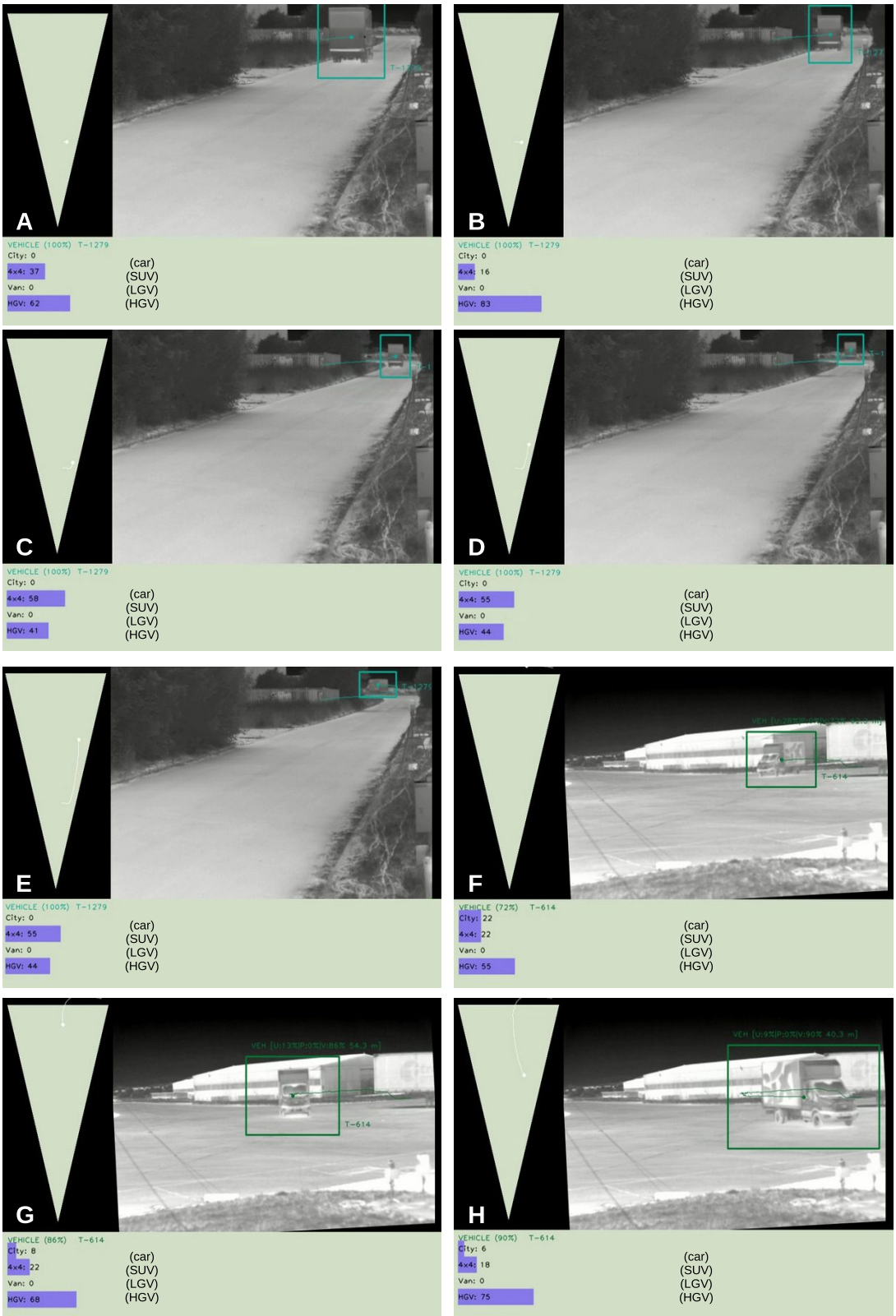

Figure 5. Examples of real-time vehicle detection, type classification and tracking in infra-red imagery with associated geo-referenced 3D track (based on using HOG features with SVM classification).

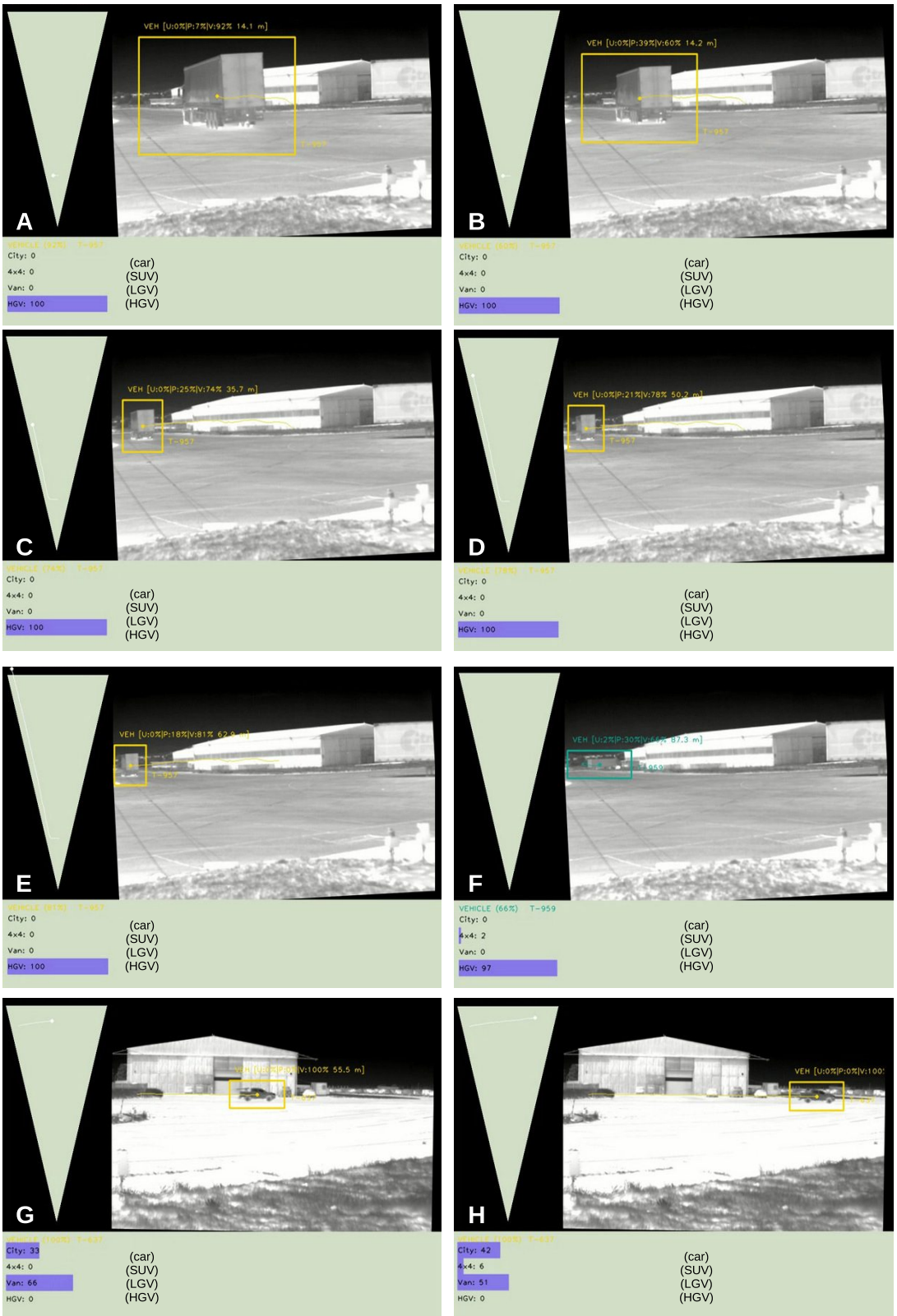

Figure 6. Examples of real-time vehicle detection, type classification and tracking in infra-red imagery with associated geo-referenced 3D track (based on using HOG features with SVM classification).

Future work will look to investigate the extension of this approach to the recovery of vehicle and pedestrian interactions for inform human/vehicle activity classification [4, 59, 62] and also the applicability within the context of mobile platform navigation [63–66], driver assistance systems [67, 68] and for multi-platform, multi-modal wide-area search and surveillance tasks [5, 69, 70].

Acknowledgments: *This work was supported by the Defence Science and Technology Laboratory (UK MOD) and the Innovate UK. Supporting parts of this work, forming background material to the contribution made on the topic of vehicle type classification here, were originally published as part of [1, 2, 5, 32] by the same authors.*

This work forms part of the wider SAPIENT (Sensing for Asset Protection using Integrated Electronic Networked Technology) programme in collaboration with Defence Science and Technology Laboratory (DSTL) [19, 27]*, QinetiQ [71], Cubica Technologies [72], Createc [61] and AptCore [73].*

REFERENCES

1. M. Kundegorski and T. Breckon, "A photogrammetric approach for real-time 3D localization and tracking of pedestrians in monocular infrared imagery," in *Proc. SPIE Optics and Photonics for Counterterrorism, Crime Fighting and Defence*, vol. 9253, pp. 1–16, 2014.
2. M. Kundegorski and T. Breckon, "Posture estimation for improved photogrammetric localization of pedestrians in monocular infrared imagery," in *Proc. SPIE Optics and Photonics for Counterterrorism, Crime Fighting and Defence*, vol. 9652, pp. 1–12, SPIE, September 2015.
3. T. Breckon, J. Han, and J. Richardson, "Consistency in muti-modal automated target detection using temporally filtered reporting," in *Proc. SPIE Electro-Optical Remote Sensing, Photonic Technologies, and Applications VI*, vol. 8542, pp. 23:1–23:12, November 2012.
4. J. Han, A. Gaszczak, R. Maciol, S. Barnes, and T. Breckon, "Human pose classification within the context of near-ir imagery tracking," in *Proc. SPIE Optics and Photonics for Counterterrorism, Crime Fighting and Defence*, vol. 8901, pp. 1–10, SPIE, September 2013.
5. T. Breckon, A. Gaszczak, J. Han, M. Eichner, and S. Barnes, "Multi-modal target detection for autonomous wide area search and surveillance," in *Proc. SPIE Emerging Technologies in Security and Defence: Unmanned Sensor Systems*, vol. 8899, pp. 1–19, SPIE, September 2013.
6. R. Craig, J. Mindell, and V. Hirani, "Health survey for England," *Obesity and Other Risk Factors in Children. The Information Centre*, vol. 2, 2006.
7. P. M. Visscher, "Sizing up human height variation.," *Nature genetics*, vol. 40, pp. 489–90, may 2008.
8. I. Urazghildiiev, R. Ragnarsson, P. Ridderstrom, A. Rydberg, E. Ojefors, K. Wallin, P. Enochsson, M. Ericson, and G. Lofqvist, "Vehicle classification based on the radar measurement of height profiles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, pp. 245–253, June 2007.
9. A. Yilmaz, O. Javed, and M. Shah, "Object tracking - a survey," *ACM Computing Surveys*, vol. 38, pp. 13–es, dec 2006.
10. H. K. Aghajan and A. Cavallaro, *Multi-camera networks: principles and applications*. Academic press, 2009.
11. G. Doretto, T. Sebastian, P. Tu, and J. Rittscher, "Appearance-based person reidentification in camera networks: problem overview and current approaches," *J. of Ambient Intelligence and Humanized Comp.*, vol. 2, no. 2, pp. 127–151, 2011.
12. X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognition Letters*, 2012.
13. S. Iwasawa, K. Ebihara, J. Ohya, and S. Morishima, "Real-time estimation of human body posture from monocular thermal images," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp. 15–20, 1997.
14. E. Brau, J. Guan, K. Simek, L. D. Pero, C. R. Dawson, and K. Barnard, "Bayesian 3D Tracking from Monocular Video," in *Int. Conf. Computer Vision*, pp. 3368–3375, 2013.
15. Z. Chen, T. Ellis, and S. A. Velastin, "Vehicle type categorization: A comparison of classification schemes," in *Proc. Int. Conf. on Intelligent Transportation Systems*, pp. 74–79, Oct 2011.
16. Z. Chen, T. Ellis, and S. A. Velastin, "Vehicle detection, tracking and classification in urban traffic," in *Proc. Int. Conf. on Intelligent Transportation Systems*, pp. 951–956, Sept 2012.
17. B. Zhang, "Reliable classification of vehicle types based on cascade classifier ensembles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, pp. 322–332, March 2013.
18. L. Yang, P. Luo, C. Change Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. Conf. on Computer Vision and Pattern Recognition*, pp. 3973–3981, 2015.
19. P. Thomas, G. Marshall, D. Faulkner, P. Kent, S. Page, S. Islip, J. Oldfield, T. Breckon, M. Kundegorski, D. Clarke, and T. Styles, "Towards sensor modular autonomy for persistent land intelligence surveillance and reconnaissance," in *Proc. SPIE Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent Intelligence Surveillance and Reconnaissance VII*, vol. 9831, pp. 1–18, SPIE, May 2016.

20. J. W. Davis and V. Sharma, "Robust detection of people in thermal imagery," in *Proc. Int. Conf. Pattern Recognition*, vol. 4, pp. 713–716, 2004.
21. J. W. Davis and V. Sharma, "Background-subtraction in thermal imagery using contour saliency," *Int. Journal of Computer Vision*, vol. 71, no. 2, pp. 161–181, 2007.
22. B. Besbes, A. Rogozan, and A. Bensrhair, "Pedestrian recognition based on hierarchical codebook of SURF features in visible and infrared images," in *Proc. Intelligent Vehicles Symp.*, pp. 156–161, IEEE, jun 2010.
23. M. Yasuno, S. Ryousuke, N. Yasuda, and M. Aoki, "Pedestrian detection and tracking in far infrared images," in *Proc. Int. Conf. Intelligent Transportation Systems*, pp. 182–187, 2005.
24. J. Wang, D. Chen, H. Chen, and J. Yang, "On pedestrian detection and tracking in infrared videos," *Pattern Recognition Letters*, vol. 33, pp. 775–785, apr 2012.
25. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.
26. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014.
27. P. A. Thomas, G. F. Marshall, D. J. Stubbins, and D. A. Faulkner, "Towards an autonomous sensor architecture for persistent area protection," in *Proc. SPIE Optics and Photonics for Counterterrorism, Crime Fighting and Defence*, no. 9995-27, SPIE, September 2016.
28. Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773–780, 2006.
29. A. Godbehere, A. Matsukawa, and K. Goldberg, "Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation," in *American Control Conference*, pp. 4305–4312, IEEE, 2012.
30. 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. Int. W'shop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 113–120, 2005.
31. C. Solomon and T. Breckon, *Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab*. Wiley-Blackwell, 2010.
32. S. Akcay, M. Kundegorski, M. Devereux, and T. Breckon, "Transfer learning using convolutional neural networks for object classification within x-ray baggage security imagery," in *Proc. Int. Conf. on Image Processing*, pp. 1057 –1061, 2016.
33. J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Int. Conf. Computer Visio*, pp. 1470–1477, 2003.
34. P. Ricaurte, C. Chilán, C. A. Aguilera-Carrasco, B. X. Vintimilla, and A. D. Sappa, "Feature point descriptors: infrared and visible spectra.," *Sensors*, vol. 14, pp. 3690–701, jan 2014.
35. D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
36. H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.
37. E. Rosten, R. Porter, and T. Drummond, "Faster and better: a machine learning approach to corner detection.," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 32, no. 1, pp. 105–19, 2010.
38. M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "Brief: Computing a local binary descriptor very fast," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 34, no. 7, pp. 1281–1298, 2012.
39. S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. Int. Conf. Computer Vision*, pp. 2548–2555, 2011.
40. E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," *Proc. Int. Conf. Computer Vision*, pp. 2564–2571, 2011.
41. P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "KAZE features," in *Proc. Euro. Conf. Computer Vision*, pp. 214–227, 2012.
42. P. F. Alcantarilla, J. Nuevo, and A. Bartoli, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," in *Proc. British Machine Vision Conf.*, pp. 13.1–13.11, 2013.
43. A. Alahi, R. Ortiz, and P. Vandergheynst, "Freak: Fast retina keypoint," in *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 510–517, 2012.
44. E. Tola, V. Lepetit, and P. Fua, "DAISY: an efficient dense descriptor applied to wide-baseline stereo.," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 32, no. 5, pp. 815–830, 2010.
45. N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp. 886–893, 2005.
46. A. Ben-Hur and J. Weston, "A user's guide to support vector machines," *Methods in Molecular Biology*, vol. 609, pp. 223–239, 2010.
47. M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Computer Vision and Pattern Recognition*, pp. 1717–1724, IEEE, 2014.

48. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, Nov 1998.
49. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
50. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *CoRR*, vol. abs/1311.2901, 2013.
51. J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," pp. 248–255, Computer Vision and Pattern Recognition, 2009.
52. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition*, pp. 580–587, IEEE, 2014.
53. K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.
54. J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in Neural Information Processing Systems*, pp. 3320–3328, 2014.
55. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014.
56. M. G. Wing, A. Eklund, and L. D. Kellogg, "Consumer-Grade Global Positioning System (GPS) Accuracy and Reliability," *Journal of Forestry*, vol. 103, no. 4, p. 5, 2005.
57. E. Maggio and A. Cavallaro, *Video tracking: theory and practice*. Wiley, 2011.
58. J. Bouguet, "Pyramidal implementation of the Lucas Kanade feature tracker," tech. rep., Intel Corporation, 2001.
59. X. Li and T. Breckon, "Combining motion segmentation and feature based tracking for object classification and anomaly detection," in *Proc. 4th European Conference on Visual Media Production*, pp. I–6, IET, November 2007.
60. E. Maggio and A. Cavallaro, "Accurate appearance-based Bayesian tracking for maneuvering targets," *Computer Vision and Image Understanding*, vol. 113, pp. 544–555, apr 2009.
61. D. J. Clark, S. L. Prickett, A. A. Napier, and M. P. Mellor, "SLATE: scanning laser automatic threat extraction," in *Proc. SPIE Optics and Photonics for Counterterrorism, Crime Fighting and Defence*, no. 9995-28, SPIE, September 2016.
62. W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, pp. 18–32, jan 2014.
63. I. Katramados, S. Crumpler, and T. Breckon, "Real-time traversable surface detection by colour space fusion and temporal analysis," in *Proc. International Conference on Computer Vision Systems*, vol. 5815 of *Lecture Notes in Computer Science*, pp. 265–274, Springer, 2009.
64. M. Breszcz and T. Breckon, "Real-time construction and visualization of drift-free video mosaics from unconstrained camera motion," *IET J. Engineering*, vol. 2015, pp. 1–12, August 2015.
65. T. Kriechbaumer, K. Blackburn, T. Breckon, O. Hamilton, and M. Riva-Casado, "Quantitative evaluation of stereo visual odometry for autonomous vessel localisation in inland waterway sensing applications," *Sensors*, vol. 15, pp. 31869–31887, December 2015.
66. P. Cavestany, A. Rodriguez, H. Martinez-Barbera, and T. Breckon, "Improved 3d sparse maps for high-performance structure from motion with low-cost omnidirectional robots," in *Proc. International Conference on Image Processing*, pp. 4927–4931, IEEE, September 2015.
67. O. Hamilton, T. Breckon, X. Bai, and S. Kamata, "A foreground object based quantitative assessment of dense stereo approaches for use in automotive environments," in *Proc. International Conference on Image Processing*, pp. 418–422, IEEE, September 2013.
68. O. Hamilton and T. Breckon, "Generalized dynamic object removal for dense stereo vision based scene mapping using synthesised optical flow," in *Proc. International Conference on Image Processing*, pp. 3439–3443, IEEE, September 2016.
69. P. Pinggera, T. Breckon, and H. Bischof, "On cross-spectral stereo matching using dense gradient features," in *Proc. British Machine Vision Conference*, pp. 526.1–526.12, September 2012.
70. M. Magnabosco and T. Breckon, "Cross-spectral visual Simultaneous Localization And Mapping (SLAM) with sensor handover," *Robotics and Autonomous Systems*, vol. 63, pp. 195–208, February 2013.
71. P. J. Kent, G. F. Marshall, and D. A. Faulkner, "An autonomous sensor module based on a legacy CCTV camera," in *Proc. SPIE Optics and Photonics for Counterterrorism, Crime Fighting and Defence*, no. 9995-31, SPIE, September 2016.
72. S. F. Page, J. Oldfield, S. Islip, B. Benfold, R. A. Brandon, P. A. Thomas, and D. J. Stubbins, "Threat assessment and sensor management in a modular architecture," in *Proc. SPIE Optics and Photonics for Counterterrorism, Crime Fighting and Defence*, no. 9995-32, SPIE, September 2016.
73. T. Styles, "Radar based autonomous sensor module," in *Proc. SPIE Optics and Photonics for Counterterrorism, Crime Fighting and Defence*, no. 9995-30, SPIE, September 2016.