Noname manuscript No.
(will be inserted by the editor)

A Novel Infrared Video Surveillance System Using Deep Learning Based Techniques

Huaizhong Zhang · Chunbo Luo · Qi Wang · Matthew Kitchin · Andrew Parmley · Jesus Monge-Alvarez · Pablo Casaseca-de-la-Higuera

Received: date / Accepted: date

Abstract This paper presents a new, practical infrared video based surveillance system, consisting of a resolution-enhanced, automatic target detection/recognition (ATD/R) system that is widely applicable in civilian and military applications. To deal with the issue of small numbers of pixel on target in the developed ATD/R system, as are encountered in long range imagery, a super-resolution method is employed to increase target signature resolution and optimise the baseline quality of inputs for object recognition. To tackle the challenge of detecting extremely low-resolution targets, we train a sophisticated and powerful convolutional neural network (CNN) based faster-RCNN using long wave infrared imagery datasets that were prepared and marked in-house. The system was tested under different weather conditions, using two datasets featuring target types comprising pedestrians and 6 different types of ground vehicles. The developed ATD/R system can detect extremely low-resolution targets with superior performance by effectively addressing the low small number of pixels on target, encountered in long range applications. A comparison with traditional methods confirms this superiority both qualitatively and quantitatively.

Keywords Video surveillance · CNN · object detection · ATD/R · Super-resolution

Huaizhong Zhang, Pablo Casaseca-de-la-Higuera, Qi Wang, Jesus Monge-Alvarez
University of the West of Scotland, Paisley PA1 2BE, United Kingdom

Huaizhong Zhang
Edge Hill University, L39 4QP, United Kingdom
E-mail: zhangh@edgehill.ac.uk

Chunbo Luo
University of Exeter, Exeter EX4 4QJ, United Kingdom

Matthew Kitchin, Andrew Parmley
Thales UK, Glasgow G51 4BZ, United Kingdom

Pablo Casaseca-de-la-Higuera
Laboratorio de Procesado de Imagen, Universidad de Valladolid, 47011, Valladolid, Spain

1 Introduction

Infrared thermography (IRT, or thermal video) has been widely used in civilian and military applications such as surveillance, night vision and tracking, weather forecasting, firefighting, facility inspections, etc. for collecting high quality image data that is beyond the human visual perception range. The exceptional capacity of IRT comes from its capability to detect and record radiation in the long-wavelength infrared range of the electromagnetic spectrum [1]. In comparison to conventional night-vision techniques, local illumination or other disturbing factors such as fog or smoke do not become an essential obstacle. Recent advances in IRT cameras have significantly improved the resolution and bit-depth of thermal images, which had previously often been considered inferior to visual images, thereby making IRT images suitable and widely used in scenarios containing high value targets, including remote surveillance applications where distant vehicles, pedestrians or buildings are monitored. For this reason, automatic detection and recognition of these targets has raised increasing interest in both academia and industry [2].

Despite the advances in acquisition technology, long range object detection and recognition in IRT images collected under real-world settings is still a challenging research topic. Such images are usually acquired at a very long distance, leading to extremely low numbers of pixels on target. A further challenge resides in the nature of IRT imaging: if the temperature of the object of interest is similar to the background, the contrast will be low. These adverse effects emerge as significant obstacles that degrade the performance of automatic object detection/recognition (ATD/R) in IRT images and hinder the application in practice. Fig. 1 shows two real-world image examples where the targets (a people carrier ("Bus") in (a) and an estate car ("Skoda") in (b)) bear low resolution and poor contrast, which can lead to high probability of false alarms in our developed ATD/R system (See Table 3 in Section 4). The Bus target in Fig. 1(a) has very low resolution (14×8 pixels), which is barely visible from a distance away. The Skoda target in Fig. 1(b) is almost blended with the background.

A number of techniques have been developed to detect and recognize objects in video surveillance. These include region-based segmentation, background subtraction, temporal differencing, active contour models, and generalized Hough transforms [3]. Due to video sequences in surveillance being obtained through static cameras and fixed background, background subtraction [4] is a commonly used detection approach in this scenario, where a background is modelled and then moving objects in a scene can be identified by comparing key frames with the background. Furthermore, moving object recognition methods rely on low-level feature based methods (Viola-Jones, Histogram of oriented gradients –HoG, Speeded Up Robust Features –SURF, Scale Invariant Feature Transform –SIFT) [5, 6] or texture descriptors (Discrete Wavelet Transform –DWT, Legendre moments and Haralick features) [7] to recognize the objects by classifying them into a certain semantic class or category. Support Vector Machines (SVM) have shown great potential for this classification task, while other methods use ensemble classifiers (e.g. AdaBoost in Viola-Jones) [8].

The schemes above may be able to detect objects to some extent in IRT applications. However, when target signatures are small, the capability of existing methods is uncertain due to the difficulties mentioned previously. Fig. 2 illustrates the DoG (Difference of Gaussian) pyramids of two images with different vehicle

types (Bus and a Transit van ("Van")). These two images are so similar at various scales that the traditional DoG based methods (like SIFT) are prone to making wrong judgements, where Bus and Van are shown in Fig. 2(a) and (b) respectively. This issue was also identified by looking at performance on the canonical visual recognition task, PASCAL VOC object detection [9]. In particular, a pipeline that HoG [6] based feature descriptor combines with a learned SVM classifier is a common way to perform object recognition in the community. This HoG descriptor extracted from an input image can be encoded into a fixed length representation that can be classified by a linear SVM [6, 10] or an additive kernel SVM [11]. We have employed this method as a benchmark for performance comparison (we will refer to it as HoG-SVM). We will show with our experiments that HoG-SVM is ineffective in IRT imaging and the performance is seriously affected by low resolution and poor contrast since these local descriptors obtained from hand-crafted features exhibit limited semantic messages of objects [11, 12]. In particular HoG are greatly dependent on image gradient, i.e. edge information or image contrast. To illustrate this fact, we present in Fig. 3 some issues appearing when applying HoG-SVM to our collected data, where the top 10 detections are kept (some detections may overlap so less than 10 boxes are observable in the figure). We performed two trial acquisitions (hereinafter T_1 and T_2, see Section 2) in different weather conditions, which led to different contrast datasets. For T_1 data, the vehicle with the salient features (see Fig. 3(a1)) can be detected and recognised by HoG-SVM. However, Fig. 3(b1) and Fig. 3(c1) show that HoG-SVM fails to detect the vehicle when applied to cases with weak features and poor contrast. In particular, it is highly ineffective for T_2 data (Fig. 3(c1)) due to the low contrast between vehicle and background. In Section 4, the quantitative comparison demonstrates numerically that our solution scheme is superior to this traditional object recognition method.

Deep learning technique in the form of convolutional neural networks (CNNs) [13], have recently achieved significant progress for object recognition by extracting informative non-linear features with hierarchical, multi-stage processes. The idea of CNNs was originally proposed by LeCun [14] in 1989 but its influence in object recognition was limited due to the popularity of SVM until the $2010s$. In their celebrated paper [13], Krizhevsky et al substantially improved the previous CNN model with several revolutionary inputs, e.g., "ReLU": $max(x, 0)$, "dropout": regularization, and a fast GPU implementation etc. Since then, with the development of CNN architectures [15, 16], CNNs show unrivalled success in object detection that is credited to some new localizing formulations proposed instead of the obsolete sliding-window detectors, such as selective search [11] and region proposal networks [17]. To date, a prevalent CNN framework [17] for object recognition is as follows: first, convolutional layers are employed to acquire region based features for detecting objects of interest; then a region-wise multi-layer perceptron (MLP) classifier is followed to do the classification for recognizing objects. Thus, a complete CNN based ATD/R system can be developed for object detection and recognition in video surveillance. Moreover, CNNs can additionally be used to perform end-to-end mapping between low and high-resolution images [18] in order to achieve a super-resolution version of the input image. This image enhancement process can help to overcome the low pixel-on-target count challenge in long-range IRT acquisitions.

This paper presents an infrared video based surveillance system consisting of a resolution-enhanced ATD/R system that can be widely used in various civilian and military applications. The system has been tested under different seasonal conditions using two datasets featuring pedestrians and 6 different types of vehicle targets. The developed ATD/R system can effectively cope with the small pixel-on-target issue and recognize extremely small-resolution targets with superior performance. A comparison with traditional methods (HoG-SVM) confirms this superiority both qualitatively and quantitatively. Preliminary results on the application of super-resolution have been presented in [19]. This present paper substantially extends the previous study in both the system development and additional experiments with new data. The main contribution in this paper is the threefold as follows.

1. To deal with the small pixel-on-target issue in the IRT imagery, we propose to use a CNN-based super-resolution method [18] for increasing signature resolution and optimising the baseline quality of inputs for object recognition.
2. To fight off the challenges like small signatures and low contrast in the IRT based surveillance, a CNN-based ATD/R system is developed consisting of a novel Region Proposal Network (Faster-RCNN [17]) which detects objects by extracting convolutional feature maps.
3. To evaluate our developed ATD/R system, we have successfully acquired the IRT surveillance data in different weather conditions for 6 types of vehicles and pedestrian. The data is released for the public use.

The remainder of the paper is as follows: Section 2 describes the data preparation and background. Section 3 introduces the framework of our ATD/R system and relevant methodologies involved. Experimental results and system evaluation are presented in Section 4. Finally, Section 5 closes the paper with the main conclusions extracted from this research.

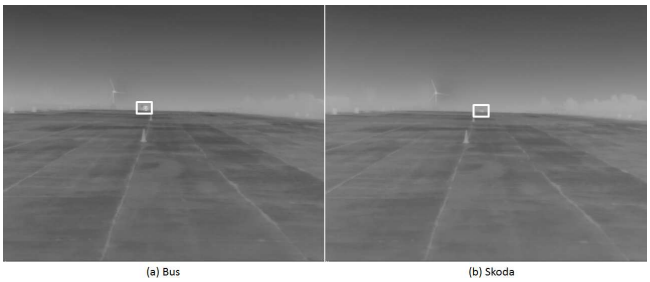

Fig. 1 Two IRT image examples captured at the distance of 600m (camera wide angle of view). The vehicle objects are highlighted in the bounding box.

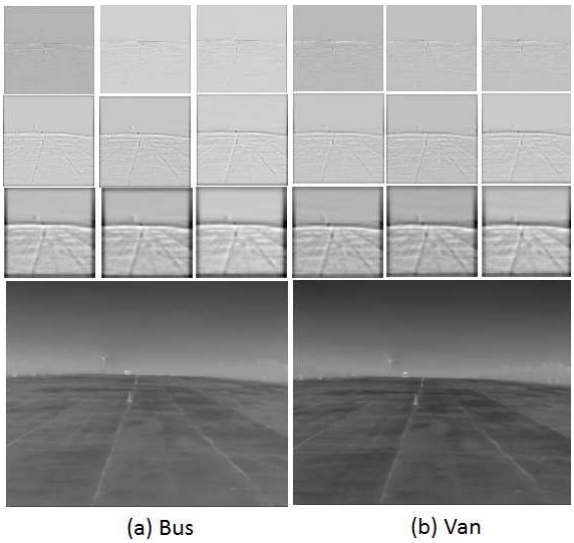

Fig. 2 DoG pyramids for two IRT images with different objects.

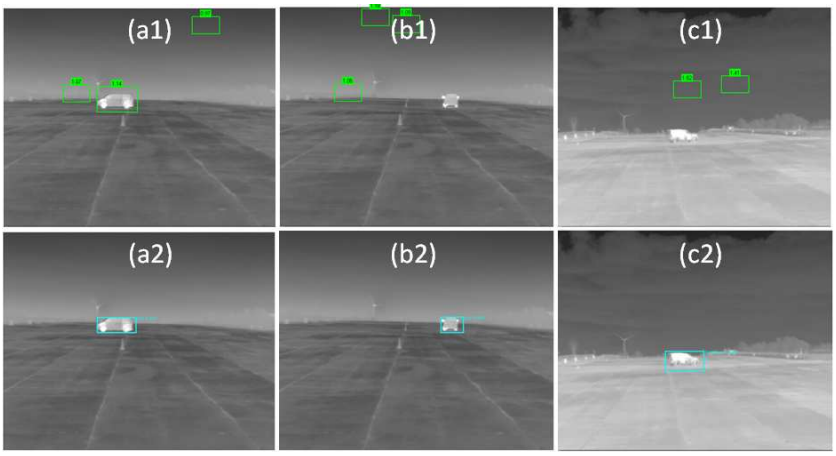

Fig. 3 Three resulting examples using Hog-SVM (a1,b1,c1) and our method (a2,b2,c2).

2 Data acquisition and pre-processing

2.1 Data acquisition and preparation

We extracted raw images from surveillance video clips acquired using Catherine MP LWIR camera (Thales UK Ltd) [20], which is a specialised thermal camera using micro-scanning technology to combine the fields of resolution (640×512). We have carried out data collection on two separate occasions, termed Trial 1 (T_1) and Trial 2 (T_2), with 6 types of vehicles involved in total as well as pedestrian targets. T_1 data was acquired in winter time, where there was significant thermal contrast between background and targets. Three types of vehicles, Bus, Skoda, and Van,

were employed as targets. The targets were acquired using the camera wide field of view at $100m, 200m, 300m, 400m, 500m$ and $600m$, and six groups of video clips were collected, each consisting of approximately $8,000$ IRT images. Fig. 1 shows two example images from a video clip collected at $600m$ range with a people carrier (left) and an estate car (right) in the respective scenes. As for T_2 scenes, they were acquired in spring time with lower contrast between the background scenery and the targets, with a distant cold blue sky expanding the image dynamic range, so the contrast levels and absolute target signatures are lower compared with T_1 scenes. The vehicles involved are Landrover, Saloon car ("Saloon"), Pickup truck ("Truck"). The datasets were acquired with the same camera setup as in T_1. Fig. 4 presents several images in T_1 and T_2. The Bus training and test datasets can be downloaded with the link, http://www.lpi.tel.uva.es/AALARTDATA. The whole data will be released soon.

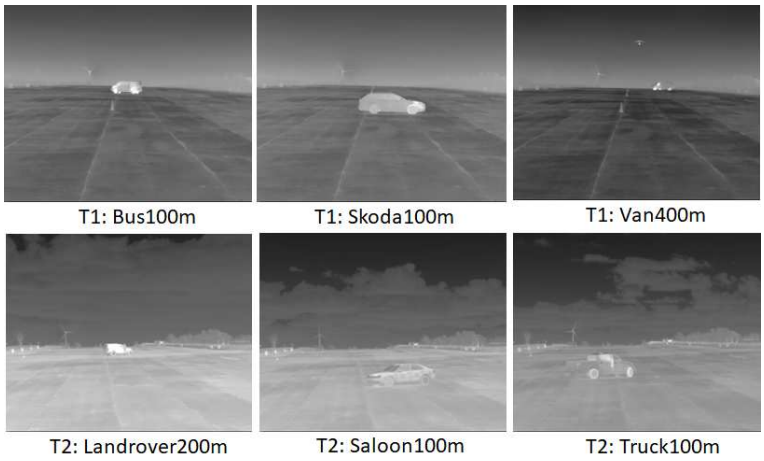

Fig. 4 Example images for T_1 and T_2.

2.2 Image preprocessing and camera bias correction

The video clips were initially acquired as "*.vstream". Raw frames were subsequently extracted and converted to the png format to feed into our ATD/R system. IRT images have a small amount of random noise uncorrelated from pixel to pixel and small-scale non-uniformities from pixel to adjacent pixel. The effect of these perturbations was negligible in the subsequent process. To decrease variability during model training and CNN calculation, the input images were average-subtracted before entering the pipeline.

3 ATD/R structure and methodology

Our ATD/R system consists of two main stages as shown in Fig. 5. At the first stage as shown in the blue block in Fig. 5, a contemporary CNN-based super-

resolution method (SRCNN) [18] is applied to improve the signature of small-number-of-pixels-on-target objects in the original IRT images. The CNN weights were trained using the raw IRT data randomly selected from the dataset. At the second stage, a state-of-the-art CNN model, faster RCNN [17], is applied to perform object detection and recognition as shown in the green block in Fig. 5. The architectures of these stages are unified into the overall CNN framework and the implementation is based on the Caffe [21] development environment.

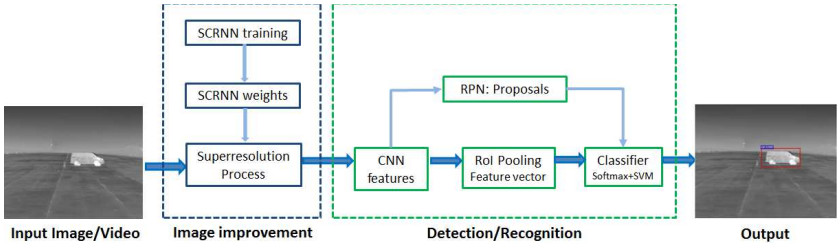

Fig. 5 Block diagram of our developed ATD/R system.

3.1 Convolutional Neural Networks

Findings on the mechanisms in the visual cortex of the brain have successfully driven CNN design, in order to address pattern based problems. Similar to a traditional neural network architecture, a CNN is made up of layers that aim to obtain a set of locally connected neurons between two layers by learning data-specific kernels. Three main types of layers are employed to build a CNN model: convolutional layer, pooling layer and fully-connected layer. A typical CNN for object recognition is illustrated in Fig. 6. The input is an image and output is a single vector of class scores. The role of each type of layer can be described as follows:

(i) The convolutional layer will generate a volume of feature maps by computing a dot product between the weights used and the region connected to the input volume.
(ii) The pooling layer will downsample the feature map along the spatial dimensions.
(iii) The fully-connected layer will create the class scores according to the given categories.

In addition, in order to make the CNN model robust, a ReLu layer is used to apply an elementwise activation function. A dropout strategy is used to perform the action of randomly ignoring neurons for preventing inter-dependencies between neurons.

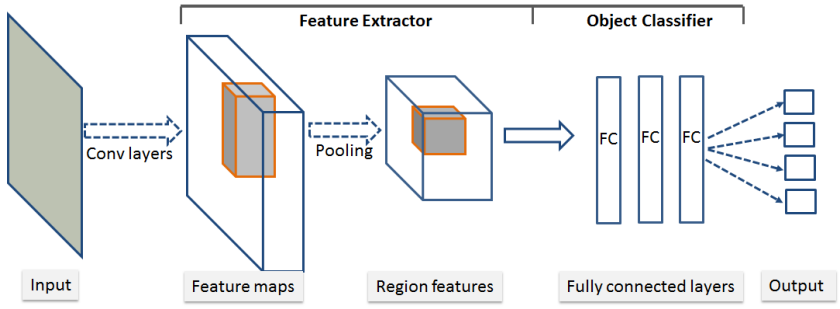

Fig. 6 A typical CNN architecture.

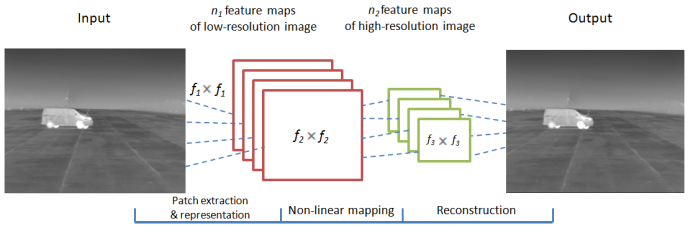

Fig. 7 Block diagram of the super-resolution process using SRCNN. Here f1, f2, f3 are the digital matrices. "×" denotes the convolutional operation.

3.2 Superresolution process using CNNs

3.2.1 Introduction to the SRCNN method

Fig. 7 shows the block diagram of processing our data using SRCNN [18]. The overall idea of super-resolution is that a low-resolution image I is upscaled to a new image Y using bicubic interpolation and then a mapping function F is employed to recover the high-resolution image X from Y. To obtain F, a popular strategy is the following: first, generate patches from Y and represent them by a set of pre-trained bases, and thus obtain the feature maps of low-resolution images; second, a non-linear mapping is applied to the feature maps so that the representation of a high-resolution patch is generated; finally, the predicated high-resolution patches are averaged to produce the final full image. In SRCNN, these traditional operations are implemented by creating a three-layer CNN. The mapping F is conceptually obtained by a CNN framework, which consists of the following three operations:

(Operation 1) Patch extraction and representation

This is the implementation of the first layer in Fig. 7. It can be described as an operation F_1:

$$F_1(Y) = max(0, W_1 * Y + B_1) \tag{1}$$

where W_1 and B_1 are the filters and biases, respectively. W_1 applies n_1 convolutions on the input image, where the kernel size is $c \times f_1 \times f_1$, with c the image channel. The output includes n_1 feature maps. B_1 is an n_1-D vector associated with the filters.

(Operation 2) Non-linear mapping

The second layer in Fig. 7 is applied to implement the following operation:

$$F_2(Y) = max(0, W_2 * F_1(Y) + B_2) \tag{2}$$

where W_2 is a matrix of $n_1 \times 1 \times 1 \times n_2$ dimensions and B_2 is an n_2-D vector. Each of the outputs is an n_2-D vector that conceptually represents a high-resolution patch.

(Operation 3) Reconstruction

This convolutional layer in Fig. 7 produces the final high-resolution image by applying the following operation:

$$F(Y) = W_3 * F_2(Y) + B_3 \tag{3}$$

where W_3 is a matrix of $n_2 \times f_3 \times f_3 \times c$ dimensions, and B_3 is a c-D vector.

3.2.2 Training the model weights with the acquired IRT images

The model weights, W_1, W_2, W_3, in Eqs. (1-3), are calculated by applying the standard stochastic gradient descent (SGD) algorithm. This is a back-propagation CNN process for the 3-layer CNN. The training set of 100 IRT images is randomly selected from our created IRT database. The following steps are performed to obtain the model weights.

(i) The ground truth images are prepared as 32×32-pixel sub-images randomly cropped from the training set.
(ii) Low resolution images are pre-processed using Bicubic interpolation.
(iii) The initial filter weights of each layer are generated by drawing randomly from a Gaussian distribution with zero mean and standard deviation 0.001. The learning rates are 0.0001 for the first two layers and 0.00001 for the last layer.

3.2.3 Applying the obtained model weights in SRCNN

In order to adapt the model to fit our IRT data and enhance the images more effectively, we integrate the trained model weights into the SRCNN model to replace the original default weights. Thus, the collected IRT images can be improved properly according to the acquisition environment and modality properties in practice.

3.3 Object detection and recognition using Faster-RCNN

Faster-RCNN [17], a most recently developed object detection system, aims to integrate traditional region proposals and an object detector/classifier into one CNN. It is composed of two main components. The first one is region proposal networks (RPN), a fully convolutional network, that produces region proposals where objects therein are likely. The second one is Fast RCNN [22], which uses the proposed regions to do classification and make a final decision on the existence of those objects. Our ATD/R system employs Faster-RCNN to carry out object detection and recognition as shown in the green block of Fig. 5. The following subsections introduce how Faster-RCNN works.

3.3.1 RPN for generating region proposals

This region proposal network is constructed as a fully convolutional network (FCN) [23] that produces region bounds and objectiveness scores simultaneously at each location. As shown in Fig. 8, the RPN architecture is actually composed of an $n \times n$ convolutional layer ($L_1, n = 3$ used) and two sibling 1×1 convolutional layers for box regression (reg) and box classification (cls) respectively.

(L1 layer): an $n \times n$ spatial window of the feature map of the last shared convolutional layer is input into this layer for generating region proposals. At each sliding position, $k(= 9)$ region proposals (the green part in Fig. 8) are created by using 3 scales and 3 aspect ratios. These k proposals are mapped to a feature vector that is fed into the reg layer and cls layer.

(reg layer): this layer performs the regression process for the input feature vector in terms of each sliding position. The outputs are the coordinates of k region proposals.

(cls layer): this layer estimates the object probability for each region proposal. The outputs are the scores of each proposal.

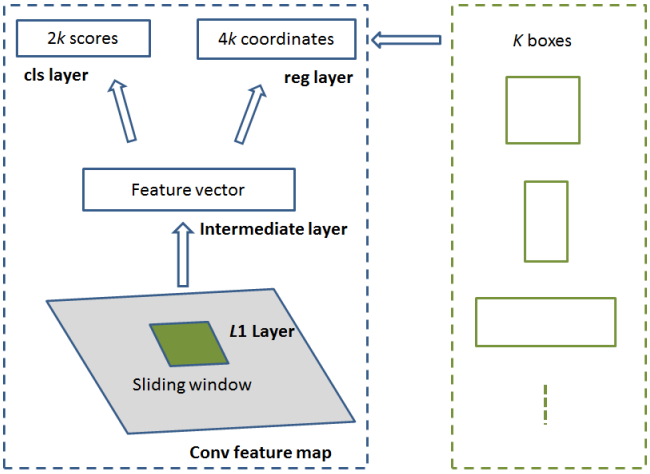

Fig. 8 Region Proposal Network (RPN)

3.3.2 Fast RCNN object detection network

Fast RCNN is an improved version of RCNN [24] for accelerating the detection process. It uses bounding box proposal methods [11] to create bounding boxes. Then, Region of Interest (RoI) pooling is applied to generate a feature vector for each bounding box. Afterwards, the feature vector is input to a 2-layer regression network and a classification network for fine-tuning the bounding boxes and obtaining class scores. Finally, non-maximum suppression (NMS) is applied over all boxes to eliminate the redundant bounding boxes.

3.3.3 Model training

For creating region proposals, the RPN is trained end-to-end by back propagation and stochastic gradient descent (SGD). For recognising objects, fast RCNN is adopted and can be trained independently. In Faster RCNN, a unified network is learnt from RPN and fast RCNN by sharing convolutional layers (see the green block in Fig. 5). This is implemented by a pragmatic 4-step training algorithm as follows.

(i) The RPN is trained as above, which is initialised with an ImageNet-pre-trained model, primarily trained on Visible Band imagery.
(ii) A detection network is trained within fast RCNN using the proposals generated in the trained RPN, which is also initialised by the ImageNet-pre-trained model.
(iii) The initialisation of RPN training is through the detection network by fixing the shared convolutional layers and only fine-tuning the layers unique to RPN.
(iv) The shared convolutional layers is kept fixed, the layers unique to fast RCNN are fine-tuned.

Thus, a unified network is formed because the same convolutional layers are shared by both of the RPN and detection networks.

4 Experimental results

The developed ATD/R system is evaluated with the collected T_1 and T_2 datasets. The evaluation results are presented both in qualitative and quantitative aspects. For demonstrating the performance of our ATD/R, we implemented the HoG-SVM method, which adopts a HoG-based sliding-window detector to localise objects in images and an SVM classifier is trained for classification while a non-maximum suppression (NMS) algorithm is used to eliminate redundant detections. As discussed in Section 1, HoG-SVM is incapable of dealing with low contrast imagery like T_2 data. So, our comparison is only presented for 438 T_1 test images. Due to HoG being sensitive to image contrast, and in order to fairly show the performance of HoG-SVM, we define that it has detected and recognised the vehicle when its detection has at least 10% overlap with the ground truth. In addition, the same training set as used in our ATD/R is employed to train the HoG model.

The following two models have been generated from the training datasets:

(i) **Raw model** − the original training dataset is used.
(ii) **Superresolution model** − the superresolution training dataset is used.

For the purposes of running our ATD/R system and HoG-SVM, our chosen PC configuration was an HP Pavilion 550 with Intel $i7-6700$ processor and Nvidia Quadro $K4200$ GPU.

4.1 Training set and test set

The ground truth data are created manually by marking the datasets of various vehicles with 100m, 200m and 400m distances. For obtaining the training and test

Table 1 ROI size statistics (pixels) for person targets in the collected data

Person	Width				Height			
Dist. (m)	Mean	SD	Min	Max	Mean	SD	Min	Max
100	32.82	6.36	11.92	55.00	66.74	15.13	11.24	94.00
200	18.32	2.99	6.18	29.00	35.18	7.38	6.75	47.00
400	12.46	2.60	4.33	41.38	21.35	3.27	4.77	29.00
600	9.08	2.19	4.14	30.18	13.81	2.99	4.14	21.00

Table 2 ROI size statistics (pixels) for vehicle targets in the collected data

Vehicle	Width				Height			
Dist. (m)	Mean	SD	Min	Max	Mean	SD	Min	Max
100	150.75	48.61	37.00	249.00	81.05	14.24	42.00	123.00
200	79.13	16.55	14.00	125.00	39.05	6.41	21.00	64.00
400	42.12	8.79	10.00	66.00	22.99	4.07	12.00	38.00
600	27.29	7.11	6.00	47.00	15.52	3.52	6.00	24.00

datasets, the ground truth data (3780 images in total) are segregated into different subsets with a given class and a camera distance such as Bus200m. Then, for each subset, a random selection with an $80-20$ ratio of training and test data is made. Afterwards, all training and test subsets are aggregated separately for generating the training set (3025 images, T_1 : 1758, T_2 : 1266) and test set (755 images, T_1 : 438, T_2 : 317). Tables 1 and 2 present the relevant statistics on the target pixel values in our collected datasets. In addition, the datasets of 300m, 500m and 600m are reserved for testing and are not involved into the training process. Note that our ADT/R system is trained using the pre-trained VGG16 model to be the initialisation, which was trained on the PASCAL VOC 2007 detection benchmark [9].

4.2 Performance results

We employ receiver operating characteristic (ROC) [25] to analyse the performance of our object detection and recognition system. The ROCs in Fig. 9 illustrate that both the raw and superresolution models of our ATD/R can perform much better than HoG-SVM in T_1 test images. The area under curve (AUC) is 0.70 for HoG-SVM (green) in Fig. 9 whereas the obtained AUCs in Fig. 9 are 0.92, 0.96 for our two models respectively. Fig. 9 shows that True Positive Rate (TPR) approaches 100% when False Positive Rate (FPR) surpasses 15% for the superresolution model (cyan) and 45% for the raw model (blue). However, for HoG-SVM, FPR is at least 95% when TPR reaches 100% as shown in Fig. 9.

Fig. 10 presents specific ROC curves for our blind test data acquired from 300m (upper row) and 500m (lower row) in T_1. Data acquired from these distances were not involved in the training process, so the relevant performance can truly reflect the ability of our ATD/R to recognise objects. Perfect AUC values are obtained for all cases at 300m, meaning that almost all vehicles are correctly recognised with negligible FPR. For the case of Skoda at 500m, we can see our approach also achieves comparable results, even though the training dataset just comes from 100m to 400m, particularly in the case of Van (Fig. 10 (f)). It is worth noting that

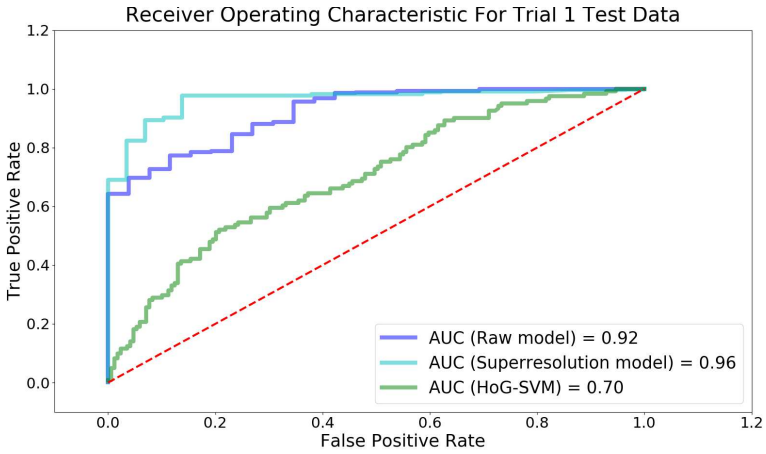

Fig. 9 ROCs for Trial 1 test data generated from the HoG-SVM method and our ATD/R.

the superresolution model improves the performance significantly compared with the raw model in the cases at 500m.

In comparison, we apply our ATD/R to the blind datasets in T_2. Fig. 11 illustrates the obtained ROCs. Due to the weather and seasonal conditions leading to low contrast in the recorded imagery, the overall performance is inferior to that in T_1. Fig. 11 (a,d) show that our method can recognise Landrover well in both 300m and 500m. For the case of Truck, Fig. 11 (b,e) demonstrate that superresolution can help increase the AUC value from 87% to 94% at 300m, 82% to 90% at 500m respectively. It is worth noting that the AUC value obtained in the case of Saloon can keep around 91% for both 300m and 500m.

Regarding the datasets acquired at 600m, the signatures of objects in images are extremely low: the smallest size for vehicles is 8×12 and for pedestrians 8×5. Fig. 12 illustrates that our ATD/R can deal with the difficulties of low resolution and poor contrast well. Particularly, a significantly high performance has been achieved in the cases of Van, Saloon and Landrover ((c),(e),(f)) when applying the superresolution model.

To examine the overall performance of our ATD/R in both T_1 and T_2 as a whole, we apply our two generated models to the entire test set ($1266+317 = 1583$ images). In addition, for a more accurate study of the role of the superresolution in the overall ATD/R chain, we applied superresolution to the test data so that we can see whether the improved test data can help enhance the performance of the final ATD/R. Fig. 13 illustrates that the superresolution model performs better than the raw model, regardless of whether or not the input data is super-resolved. However, the raw model applied to the super-resolved dataset does not provide better performance in comparison with the raw data input.

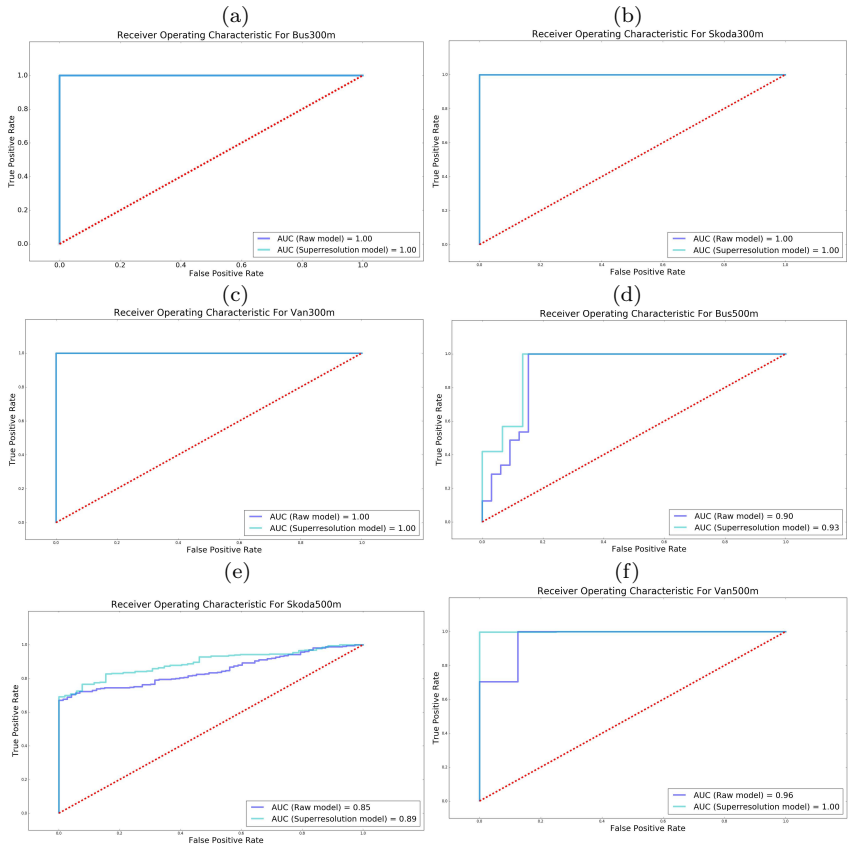

Fig. 10 ROCs for the 300m (a,b,c) and 500m (d,e,f) datasets of Bus, Skoda and Van in Trial 1. Each graph features two ROCs corresponding to raw model and super-resolution model-derived results.

4.3 Difficult cases

This subsection presents qualitative and quantitative results on the performance of our ATD/ATR system in some difficult cases, in order to further illustrate its application.

Table 3 specifically gives the detection confidences of the objects shown in Fig. 1, and provides some insight about the method. Before applying the proposed image enhancement method, the top three detection probabilities are: Bus, 0.311699; Skoda, 0.875782; and Van, 0.0154172. Under these results, the target would be classified as a Skoda since it has the highest detection confidence. However, if we check the ground truth, the target is actually a Bus. This error is clearly corrected by adopting the proposed methodology. The second row in Table 3 shows that, after image enhancement, the detection probability of Bus is the highest among the three. Therefore, the ATD/R system can correctly recognise the target that was wrongly interpreted in the absence of image enhancement. Another point worth noting is the following. In Table 3, for Skoda in Fig. 1(b), the

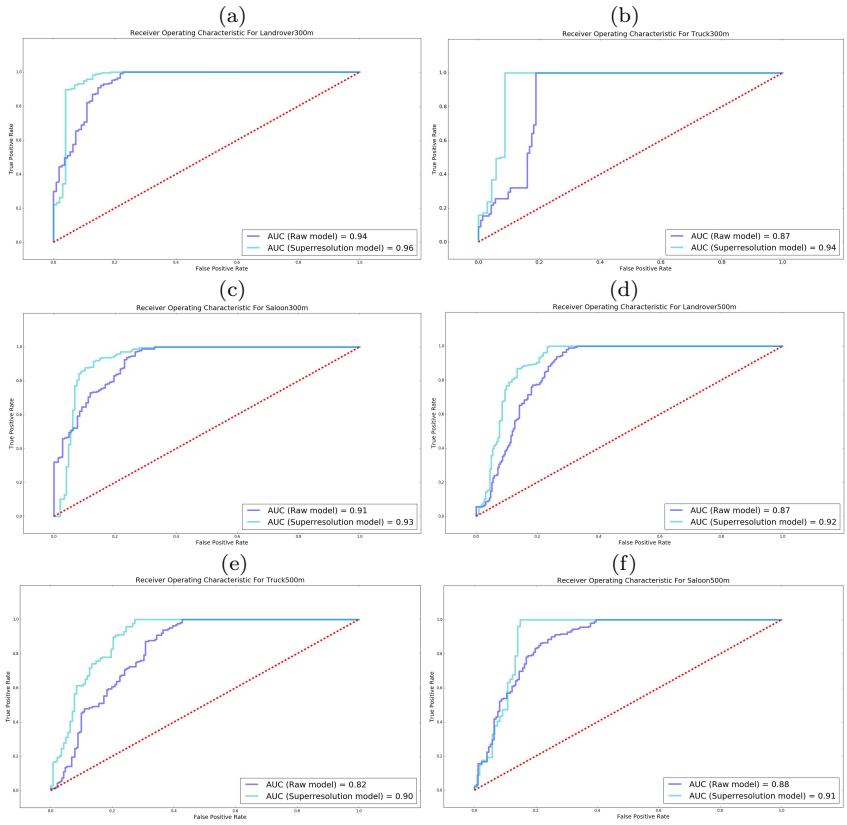

Fig. 11 ROCs for the 300m (a,b,c) and 500m (d,e,f) datasets of Truck, Saloon and Landrover in Trial 2. Each graph features two ROCs corresponding to raw model and super-resolution model-derived results.

Table 3 The recognition results of Fig. 1 using raw training dataset and enhanced training dataset, respectively. The values presented denote the probabilities of each recognised vehicle type. The ground truths of Fig.1(a) and (b) are Bus and Skoda, respectively.

	Bus	Skoda	Van
Using raw training dataset (Fig. 1(a))	0.274491	**0.875782**	0.0154172
Using enhanced training dataset (Fig. 1(a))	**0.311699**	0.275520	0.0202263
Using raw training dataset (Fig. 1(b))	0.000419	**0.205708**	0.0014541
Using enhanced training dataset (Fig. 1(b))	0.001150	**0.48366**	0.0025878

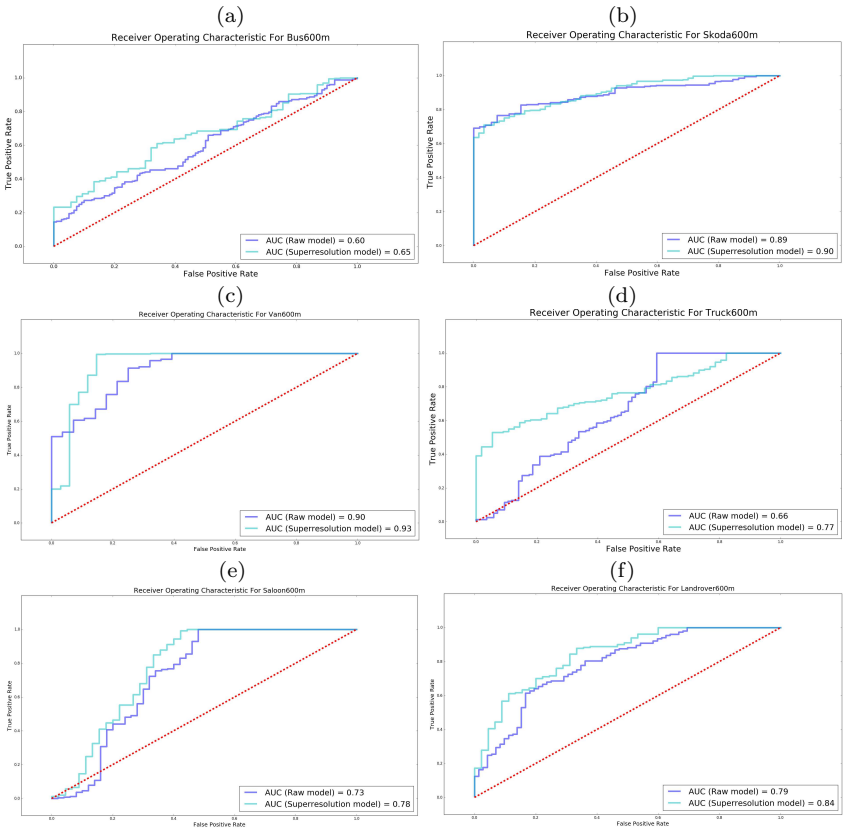

Fig. 12 ROCs for the 600m datasets of Bus, Skoda, Van, Truck, Saloon and Landrover. Each graph features two ROCs corresponding to raw model and super-resolution model-derived results.

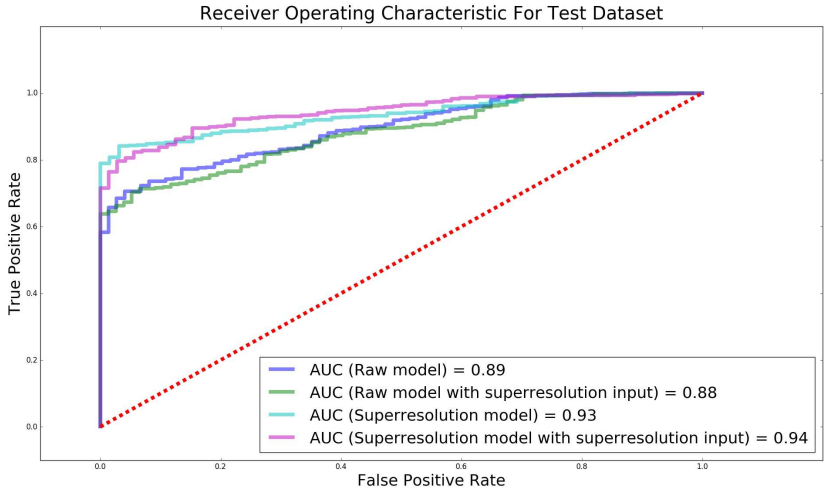

Fig. 13 ROCs for the results using different models in the test data.

system exhibits detection confidence increasing from 0.205708 to 0.483660 after data enhancement.

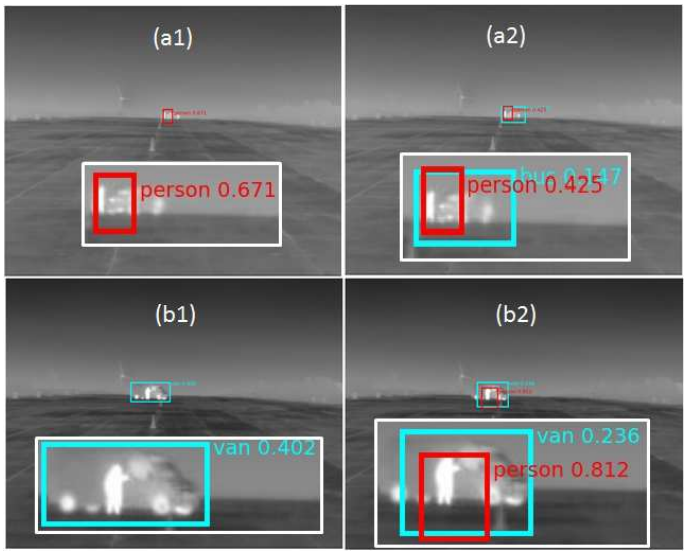

Fig. 14 (a1-a2) Results with the raw-model and the superresolution-model for Bus500m respectively, (b1-b2) Result with the raw-model and the superresolution-model for Van300m respectively.

Fig. 14 shows detection performance when the raw model is used to detect objects for the cases of Bus 500m (a1) and Van300m (b1) respectively. It is worth noting that Bus is not detected in (a1) and the pedestrian is missed in (b1). However, the superresolution model can detect all objects in both of these cases as shown in (a2) and (b2).

Concerning the cases with extremely small signatures, three examples from the 600m distance, Bus, Van with pedestrian, and Saloon, are presented in Fig. 15. The resulting images using the raw model are shown in Figs. 15(a1)(b1)(c1), and indicate that both Bus and Saloon are wrongly detected in (a1)(c1) and the pedestrian is not detected in (b1). Figs. 15(a2)(b2)(c2) show the resulting images where the superresolution model is applied to the same images. We can clearly see that the relevant objects are correctly recognised accordingly.

Fig. 16 illustrates an example of a pick up truck from 600m distance in T_2. The detection confidence using the superresolution model is 0.896 while it is 0.856 with the raw model. This visually demonstrates that the superresolution model can improve the performance in comparison to the raw model.

4.4 Discussion

Due to the discrepancy in image quality between two Trials, Figs. 10 and 11 show that our ATD/R has performed much better in T_1 than in T_2. However, it is

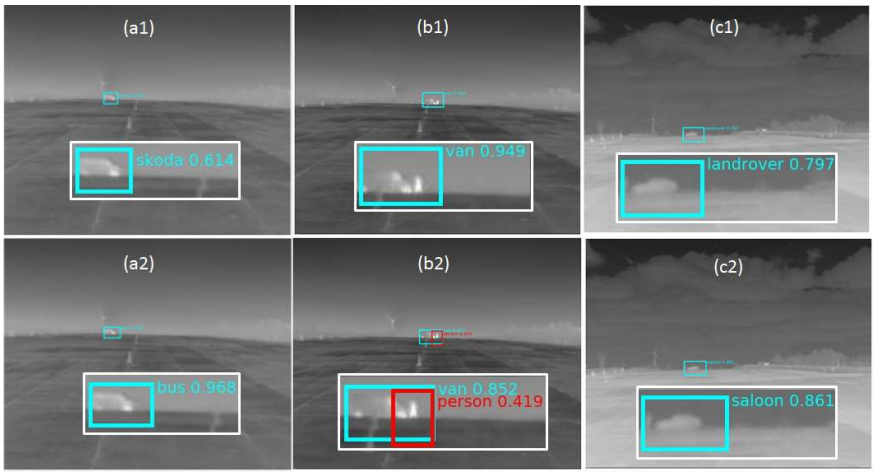

Fig. 15 Examples showing the performance of the superresolution-model. Upper row: the results for Skoda600m, Van600m, Saloon600m with the raw-model. Lower row: the results for the same images with the superresolution-model.

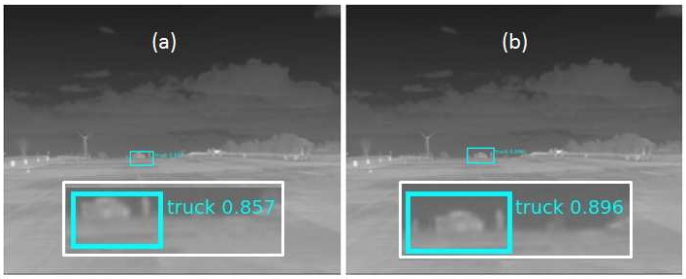

Fig. 16 Examples for Truck600m: (a) Result with the raw-model, (b) Result with the superresolution-model

apparent that Fig. 9 demonstrates that our ATD/R can deal with the issue of low contrast much better than HoG-SVM.

The success of superresolution on improvement of ATD/R in long distant cases is generally reduced, since feature information extracted may be limited. Figs. 12 (a,b,d,e) indicates that the superresolution model performs similarly to the raw model at the 600m range. However, in Figs. 12 (c,f), we can see that the superresolution technique helps improve the performance significantly. A plausible reason is that the vehicle signature in these cases of Van and Landrover has larger dimensions compared with other vehicles.

As shown with the presented ROCs, although the overall detection and recognition performance is improved, the small size of the objects at distant views can still cause some false positives. Due to the limited number of pixels on the small object signature being presented, the enhancement process may generate inaccurate feature information, leading to false positives. For example, a Bus target is initially detected as a Bus correctly with a confidence of 0.729734 in the raw model; however, it is wrongly detected as a Skoda with a confidence of 0.772193 after the enhancement processing. This issue may potentially be improved by refining the

training set in future work so that the ATD/R system can obtain more accurate feature information for recognition. In our system, such false positives introduced by the enhancement process are rare and considerably outweighed by the significant performance gain. In fact, the ROCs show that the true positive ratios have been greatly improved after the enhancement process.

The use of Faster-RCNN is justified in our work due the difficulty of annotating very small signature objects at long distances. The proposed regions in the feature space allow training at shorter distances where the annotations are feasible ($100-400m$) and then deploy the system for small signatures ($600m$) with high accuracy. This can be achieved since the RPN applies a multi-scale method with a sliding window associated to a scale and aspect ratio. In terms of CNN-based object recognition, there are several recently proposed techniques which have shown good performance in both accuracy and speed. Particularly, the single shot mulitbox detector (SSD) [26] and YOLO [27] can be highlighted. YOLOs model training is based on the entire image rather than the region proposal network (RPN) used in Faster-RCNN. So, YOLOs loss function deals equally with all bounding boxes. This leads to YOLO underperforming for small objects if accurate annotations cannot be provided (as in our case at $600m$). As for SSD, it employs multi-scale and data augmentation to enhance the detection accuracy. In addition, SSD uses bank filters to subsample data for faster calculations. However, SSD is insensitive to detect smaller objects since SSD features a class-aware RPN with a lot of bells and whistles. Therefore, Faster-RCNN is better fitted to our IRT images where very small objects are difficult to annotate and detect accurately.

5 Conclusions

This paper presents an ATD/R system that is able to deal with the main difficulties in IRT video surveillance. First, we propose a CNN-based super-resolution method to improve IRT images, especially in long-distance view cases where the small target signature can hinder the detection/recognition process. Then, a state-of-the-art object detection method, faster RCNN, is employed to carry out object detection and recognition. We integrate these two approaches into our system to produce a robust and accurate surveillance system. Evaluation results show that the performance of the developed ATD/R system can efficiently deal with the obstacles in IRT video, and thus validates the surveillance system in practice. The study suggests that further work including developing advanced super-resolution methods, incorporating appropriate denoising techniques and geometrical feature extraction like [28], and integrating the methods to create a fully deployable system can be valuable extensions.

Acknowledgements This work was funded by Thales UK, the Centre of Excellence for Sensor and Imaging System (CENSIS), and the Scottish Funding Council under the project "AALART. Thales-Challenge Low-pixel Automatic Target Detection and Recognition (ATD/ATR)", ref. CAF-0036. Thanks are also given to the Digital Health and Care Institute (DHI, project Smartcough-MacMasters), which partially supported Mr. Monge-Alvarez's contribution, and to the Royal Society of Edinburgh and National Science Foundation of China for the funding associated to the project "Flood Detection and Monitoring using Hyperspectral Remote Sensing from Unmanned Aerial Vehicles", which partially covered Dr. Casaseca-de-la-Higuera's, Dr. Luo's, and Prof. Wang's contribution. Dr. Casaseca-de-la-Higuera would also like to acknowledge the Royal Society of Edinburgh for the funding associated to project "HIVE".

References

1. S. Bagavathiappan, B. Lahiri, T. Saravanan, J. Philip, and T. Jayakumar. Infrared thermography for condition monitoring: A review. *Infrared Physics and Technology*, 60:35–55, 2013.
2. D. Manolakis, D. Marden, and G. Shaw. Hyperspectral image processing for automatic target detection applications. *Lincoln Laboratory Journal*, 14(1), 2003.
3. J. Nascimento and J. Marques. Performance evaluation of object detection algorithms for video surveillance. *IEEE Trans On Multimedia*, 8(4), 2006.
4. C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, 1999.
5. D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004.
6. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.
7. L. Fernandez Robles. Object recognition techniques in real applications. In *PhD Thesis, University of Groningen*, 2016.
8. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.
9. M. Everingham, L. van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 80(2):303–338, 2010.
10. D. McAllester P. Felzenszwalb, R. Girshick and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 32(9):1627–1645, 2010.
11. van de Sande K. Gevers T. Uijlings, J. and A. Smeulders. Selective search for object recognition. *Int. J. Comput. Vis.*, 104(3):154–171, 2013.
12. L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*, 2015.
13. A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Adv. Neural Inf. Process. Syst.*, pages 1106–1114, 2012.
14. Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1(4):541–551, 1989.
15. M. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, 2014.
16. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
17. S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *PAMI*, 2016.
18. C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE T-PAMI*, 38(2):295–307, 2016.
19. H. Zhang, P. Casaseca de-la Higuera, C. Luo, Q. Wang, M. Kitchin, A. Parmley, and J. Monge-Alvarez. Systematic infrared image quality improvement using deep learning based techniques. In *SPIE*, pages 515–522, 2016.
20. R. Craig and J. Parsons. Thermal imaging for current d&s priorities. In *SPIE*, 2012.

21. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *arXiv preprint arXiv:1408.5093*, 2014.
22. R. Girshick. Fast r-cnn. In *ICCV*, 2015.
23. J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.
24. R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *TPAMI*, 38(4):142–158, 2016.
25. T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874, 2006.
26. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.
27. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.
28. H. Xu, J. Yan, N. Persson, W. Lin, and H. Zha. Fractal dimension invariant filtering and its cnn-based implementation. In *CVPR*, 2017.