TWO-STREAM BOOSTED TCRNET FOR RANGE-TOLERANT INFRA-RED TARGET DETECTION

Md Jibanul Haque Jiban, Shah Hassan, Abhijit Mahalanobis

Department of Computer Science, University of Central Florida, Orlando, FL 32816

ABSTRACT

The detection of vehicular targets in infra-red imagery is a challenging task, both due to the relatively few pixels on target and the false alarms produced by the surrounding terrain clutter. It has been previously shown [1] that a relatively simple network (known as TCRNet) can outperform conventional deep CNNs for such applications by maximizing a *target to clutter ratio* (TCR) metric. In this paper, we introduce a new form of the network (referred to as TCRNet-2) that further improves the performance by first processing target and clutter information in two parallel channels and then combining them to optimize the TCR metric. We also show that the overall performance can be considerably improved by boosting the performance of a primary TCRNet-2 detector, with a secondary network that enhances discrimination between targets and clutter in the false alarm space of the primary network. We analyze the performance of the proposed networks using a publicly available data set of infra-red images of targets in natural terrain. It is shown that the TCRNet-2 and its boosted version yield considerably better performance than the original TCRNet over a wide range of distances, in both day and night conditions.

Index Terms— TCRNet, Infrared, Target Detection, MWIR, Surveillance

1. INTRODUCTION

The automatic detection of targets in infrared (IR) imagery is an essential capability for long-range surveillance and reconnaissance applications. However, the natural clutter background makes it very difficult to reliably detect targets with a low false-alarm rate [2] [3]. Although the state of the art object detectors (such as YOLOv3 [4] and Faster-RCNN [5]) work very well for many computer vision applications, they are unable to find targets at distant ranges in infrared images [1]. This is mainly due to the poorer resolution and weaker contrast of targets, and the effects of background clutter. The original TCRNet [1] directly addressed this issue by maximizing the TCR metric (defined as the ratio of the energies in the network's response to the target and to clutter) and thus mitigated false alarms produced by clutter while increasing the ability to detect the targets. Notably, the filters in the first layer of the TCRNet were also analytically derived to maximize the same TCR cost function. In turn, this reduced the number of learnable parameters and enabled the rest of the network to learn using relatively few training images.

McIntosh et al. [1] showed that for infrared object detection in clutter, the original TCRNet outperformed Yolo-v3 and Faster-RCNN in terms of probability of detection, with substantial margins of 30% and 38.8%, respectively. Although the TCRNet performed better than other SOTA object detection networks for this application [1], our goal is to further improve the *receiver operating characteristic* (ROC) curve by increasing the probability of detection and reducing false alarms. This is motivated by the fact that target detection for security and surveillance applications requires a high probability of detection with extremely low false alarm rates. The architecture of the TCRNet-2 is described in detail in Section 2. The idea of boosting performance of the TCRNet-2 is described in Section 2.2. In this scheme, a primary TCRNet-2 is used to nominate the potential regions of interest (ROI). The second network is specifically trained to discriminate between the targets and clutter false alarms produced by the first network. The final detection confidence is the sum of the scores produced by both networks at the exact same ROI. In Section 3 we present the results obtained using the same publicly available data set and the testing and training protocols described in [1], followed by a summary and concluding remarks in Section 4.

2. TWO-STREAM TCRNET

The architecture of the TCRNet-2 is shown in Figure 1. Like its predecessor, the two-stream TCRNet architecture also uses fixed *pre-determined* filters in the first layer of the network which are analytically derived to optimize the TCR Metric described in Section 2.1. In a nutshell, these filters are eigen-vectors that discriminate between target and clutter information. Holding the first layer frozen, the rest of the network is trained using gradient descent to also maximize the same TCR metric. However, unlike the original version, the TCRNet-2 processes the target and clutter information in separate channels to further increase the discrimination

The authors gratefully acknowledge the support of Leonardo DRS for the work

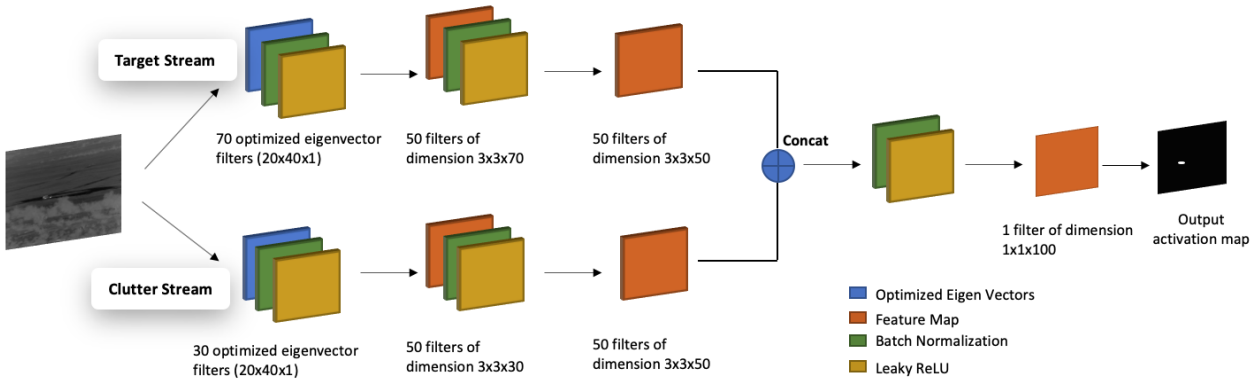


Fig. 1: Architecture of 2-stream TCR Network. Filters of the first layers of both target and clutter streams are analytically derived while the other convolutional layers are iteratively learned to minimize the TCR cost function.

between the two sub-spaces. As shown in Figure 1, the *target stream* has seventy 20x40 dimensional filters in the first layer while the *clutter stream* has thirty filters of the same size in its first layer. The choice of the number of target and clutter filters is based on the distribution of nonzero eigen-values [1]. Within each stream, the output of the first layer is processed by two more convolution layers with fifty 3×3 filters. The output of these two streams is then concatenated, and a final convolution layer is applied to produce the combined output. Batch Normalization [6] and ReLU [7] are applied in all layers. Targets are detected by finding the local maxima in the final output activation map which is produced by the network.

2.1. TCR Metric

We now briefly review the formulation of the TCR metric and how it is used not only as a cost function for training the rest of the network, but also for analytically deriving the filters in the first layer. Let us assume that we have N labeled samples for the target and clutter classes which produce the final outputs of the network denoted by $\{x_1, x_2 ... x_N\}$ and $\{y_1, y_2 ... y_N\}$, respectively. Our objective is to maximize the energy in the output when targets are present and minimize the same in response to clutter; where energy means sum of the square of the pixel values. This is accomplished by minimizing the $J^{'}_{TCR} = \frac{\frac{1}{N}\sum y_i^T y_i}{\sqrt[N]{\prod x_i^T x_i}}$, which is the ratio of the arithmetic mean of the energy of the clutter outputs to the geometric mean of the energy of the target outputs. Minimizing this ratio will make the numerator of $J^{'}_{TCR}$ small, which in turn ensures that all the terms in the summation $\frac{1}{N}\sum y_i^T y_i$ are small. Similarly the denominator of $J^{'}_{TCR}$ must be large to minimize the ratio, which implies that $\sqrt[N]{\prod x_i^T x_i}$ is large, which in turn ensures that all terms in the product are large. It can be shown that the derivative of the log of this function with respect to each class is

$$\nabla_{y_i} \log(J^{'}_{TCR}) = \frac{2y_i}{\sum y_i^T y_i}, \nabla_{x_i} \log(J^{'}_{TCR}) = -\frac{1}{N}\frac{2x_i}{x_i^T x_i} \quad (1)$$

Therefore, as the training images are presented to the network during the learning process, the gradient supplied to the back-propagation algorithm is either $\nabla_{y_i} \log(J^{'}_{TCR})$ for clutter images, or $\nabla_{x_i} \log(J^{'}_{TCR})$ for target images. It should be noted that for one training image considered at a time, the gradient expressions for the two classes reduce to $\nabla_{y_i} \log(J^{'}_{TCR}) = \frac{2y_i}{y_i^T y_i}$ and $\nabla_{x_i} \log(J^{'}_{TCR}) = -\frac{2x_i}{x_i^T x_i}$ which are simply the energy normalized outputs produced by the training images of the respective classes.

First Layer Filters: The filters for the first layer of the network (denoted by q_i) also maximize a variant of the same TCR metric expressed as the Raleigh quotient

$$J_{TCR} = \frac{\prod_{i=1}^{M} q_i^T R_1 q_i}{\sum_{i=1}^{M} q_i^T R_2 q_i} \quad (2)$$

Here, $R_1 = E\{x_i x_i^T\}$ and $R_2 = E\{y_i y_i^T\}$ are the correlation matrices of the target (x_i) and clutter (y_i) training vectors, respectively, and q_i are eigen vectors that satisfy $R_2^{-1} R_1 q_i = \gamma_i q_i$. It turns out that the dominant eigen-vectors corresponding to the largest values γ_i best represent targets, while those corresponding to the smallest eigen-values principally represent clutter. As described before, these filters are preloaded into the first layer of the network and then remain fixed while the rest of the network is also trained to optimize the TCR cost function. Details of the derivation of Eq (1) and (2) can be found in [1].

2.2. Boosted TCRNet-2 networks:

The concept of boosting the output of the TCRNet-2 is shown in Figure 2 where the first network is used as a primary detector for nominating the regions of interest (ROI) that may contain potential targets. The primary detector is trained using

all target images, as well as a large set of randomly selected clutter images. The second network is trained on the same target images, but its clutter training set comprises only of *false positives* produced by the primary network. These were obtained by applying the primary network to a separate set of images of natural terrain that did not contain any targets. During operation, the second TCRNet-2 focuses only on the ROIs produced by the primary network. The final detection score is the sum of the confidence values produced by the two networks. The proposed boosting strategy is expected to

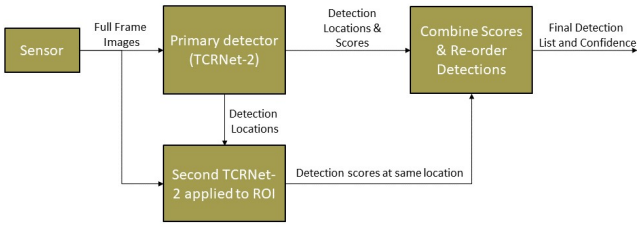


Fig. 2: Boosted Target Detection using two TCRNet-2s

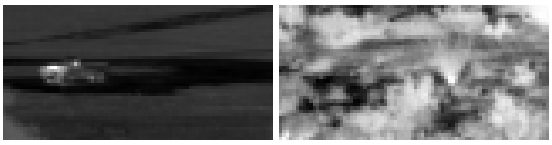
improve the ROC curve because while both networks should respond to targets in a similar way, the second network will produce a weaker response to the false positives produced by the primary detector. A simple justification for the underlying principle is as follows. Let the scores produced by the primary and boosting networks in response to targets be denoted by t_1 and t_2, respectively. The expectation of the *target* energy is $E[t_1^2] = E[t_2^2] = \sigma_t^2$. Since both of the networks are trained on the same target chips, we can assume the target scores are highly correlated so that $E[t_1 t_2] \cong \sigma^2$. Similarly, let energy of the output produced by the two networks in response to clutter be denoted by $E[c_1^2] = E[c_2^2] = \sigma_c^2$. Furthermore, since the two networks are trained on different and unrelated sets of clutter, we can assume that $E[c_1 c_2] = 0$. The output *target to clutter ratio* (TCR) for the primary network is $TCR = \sqrt{\frac{\sigma_t^2}{\sigma_c^2}}$. For the boosted TCRNets, the output TCR is given by

$$TCR_{boosted} = \frac{E[(t_1+t_2)^2]}{E[(c_1+c_2)^2]} = \frac{E[t_1^2]+E[t_2^2]+2E[t_1 t_2]}{E[c_1^2]+E[c_2^2]+2E[c_1 c_2]} = \sqrt{\frac{4\sigma_t^2}{2\sigma_c^2}} = \sqrt{2}TCR \tag{3}$$

Therefore, under the assumption that the response of the two networks to clutter is uncorrelated, the proposed method should improve the output TCR by a factor of $\sqrt{2}$.

3. RESULTS

The experiments reported in this paper are conducted using a public domain dataset released by the US Army Night Vision and Electronic Sensors Directorate (NVESD) [8]. The data contains mid-wave infrared (MWIR) imagery of ten different vehicular targets collected during the day and night, and at different ranges between 1000m and 5000m in increments of 500m. Following the same protocol in [1], the images at ranges of 1000m, 1500m, and 2000m were used to train the networks, while tests were conducted using the images at ranges of 2500m, 3000m, and 3500m. This was done to ensure that both the target and background in the testing and training sets are sufficiently different. All training images were resized using range information such that the apparent distance to the target was 2500m. Thereafter, training chips of size 40 $\times$ 80 were extracted for targets and clutter, examples of which are shown in Figure 3. While the target training chips were extracted using ground truth information, the clutter chips for training the primary TCRNet-2 were selected randomly from the full-frame images. For boosting, however, the clutter chips were centered at the location of the false alarms produced by the primary network on the images at the farthest ranges of 4000m, 4500m, and 5000m. This ensured that the clutters used for training the two networks are different and that the boosting network learns the false alarm space of the primary network. In total, there were 10800 training image chips each for target and random clutter for training the primary network, 6725 detected clutter chips for training the boosting network, and 9720 full-frame images for testing the networks. A batch size of 100, epoch size of 15, Adam optimizer, 0.0001 learning rate, and 0.002 weight decay are used



(a) Target chip. (b) Clutter chip.

Fig. 3: Example of a typical target and clutter training chips.

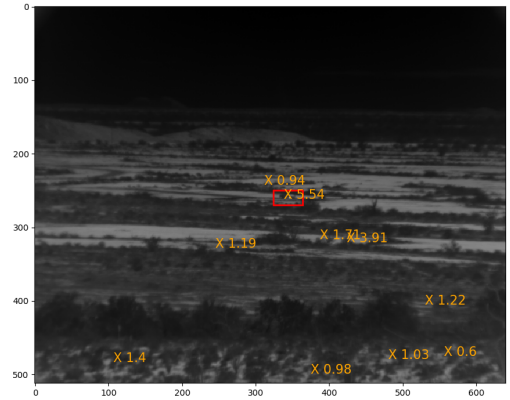
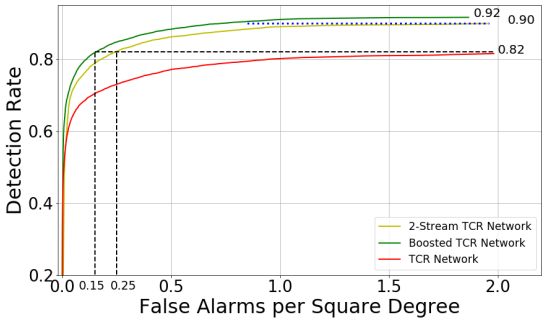
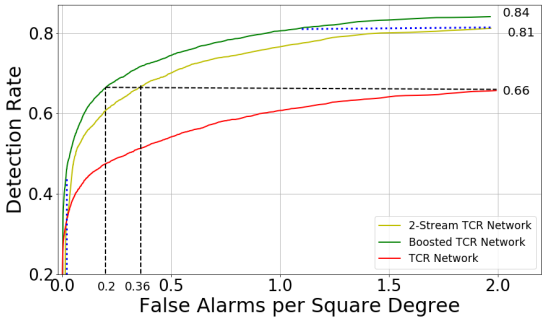


Fig. 4: A typical full-frame image showing the output of the TCRNet-2. The $\times$ marks are the detections and their confidence scores, while the red boundary is the ground truth.



(a) Both day and night images



(b) Only day images

Fig. 5: ROC Curves comparing orginal TCRNet, TCRNet-2, and its boosted version

as hyperparameters in both primary and boosting networks.

An example of a test image with detections is shown in Figure 4. For evaluation, we initially treat a detection as correct if it is within a nominal distance of 20 pixels from the center of the ground truth box. The key performance metrics are the *probability of detection* (P_d) defined as the percentage of correctly detected targets, and the *false alarm rate* (FAR) defined as the number of false detections per square degree. Figure 5a compares the ROC curves of the original TCRNet, TCRNet-2, and its boosted version using the test images during both day and night. It is clear these outperform the original TCRNet in terms of both P_d and FAR. Specifically, the maximum P_d is 0.90 for the TCRNet-2, while that of the original TCRNet is 0.82 at a FAR of 2.0. Similarly, at a FAR of 1.0, P_d is 0.8 for the original network, and 0.9 for the TCRNet-2, and 0.92 for its boosted version. Both methods also yield lower FAR at any given P_d. Specifically at $P_d = 0.82$, the black dashed lines indicate that the TCRNet-2 and its boosted version achieve at FAR values of 0.25, and 0.15 respectively, which are $8\times$ and $13\times$ lower than FAR = 2.0 for the original TCRNet. The day time performance of the network is shown in Figure 5b when clutter conditions are challenging. At a FAR = 1.0, P_d is 0.61 for the original network, 0.77 for the TCRNet-2, and 0.81 after boosting. Similarly, the TCRNet-2 and its boosted version yield FAR values of 0.36 and 0.2, respectively, at $P_d = 0.66$, which are a factor of $5\times$ and $10\times$ lower compared to FAR = 2.0 for the original

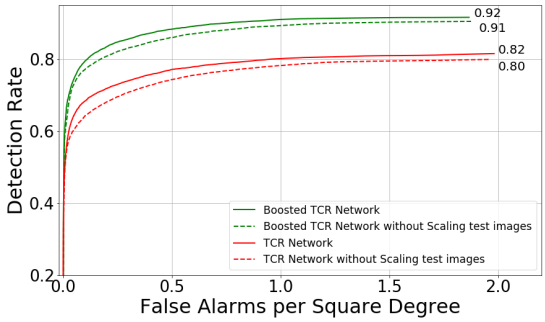


Fig. 6: The ROC curves for the TCRNet and boosted TCRNet-2 do not change substantially when test images are not resized using range information

TCRNet at the same P_d. During the night, all methods perform well due to less intense clutter conditions (i.e. because the terrain is cooler) as shown in [1] that the original TCRNet achieves 90% detection at night with 0.05 FA/sq degree.

Range Tolerance: As stated earlier, the results discussed thus far used range information to resize all test images such that the targets appear to be at a distance of 2500m. This permits a fixed radius of 20 pixels from the ground truth position to be used as a valid detection window in all test scenarios. However, it is desirable to not explicitly resize the test image both for computational reasons as well the potential lack of range information in some tactical scenarios. Instead, the radius of detection can be easily varied depending on the distance at which we wish to detect targets. Very simply, this is specified as $Detection\ Window = 2500/(Range) \times 20$. Although the networks are trained using images scaled to 2500m, the results of the unscaled test images are surprisingly good, shown in Figure 6. Both the TCRNet and the TCRNet-2 are fairly insensitive to the range, and ROC curves obtained using scaled and unscaled test images are comparable.

4. CONCLUSION

We have introduced the TCRNet-2, a new two-stream architecture, for detecting targets in MWIR , surrounded by natural terrain clutter. The TCRNet-2 substantially outperforms the original single-stream version of the network (which was previously shown to outperform the Faster RCNN and YOLOv3 for this application [1]). The overall performance was further improved by using a second network to attenuate the false alarm produced by the primary detection network. We showed that the TCRNet-2 and its boosted version not only achieve higher P_d, but also reduce FAR by factors of $8\times$ and $13\times$ during night and day, and by $5\times$ and $10\times$ specifically during the day when clutter conditions are more challenging. Finally, we showed that the TCRNet-2 is inherently robust to range variations, and the test images do not need to be explicitly scaled. Instead, a variable size detection window ensures that targets at different distances are successfully detected.

5. REFERENCES

[1] B. McIntosh, S. Venkataramanan, and A. Mahalanobis, "Infrared target detection in cluttered environments by maximization of a target to clutter ratio (tcr) metric using a convolutional neural network," *IEEE Transactions on Aerospace and Electronic Systems*, pp. 1–1, 2020.

[2] E. Gundogdu, A. Koç, and A. A. Alatan, "Automatic target recognition and detection in infrared imagery under cluttered background," in *Target and Background Signatures III*. International Society for Optics and Photonics, 2017, vol. 10432, p. 104320J.

[3] J. A. Ratches, "Review of current aided/automatic target acquisition technology for military target acquisition tasks," *Optical Engineering*, vol. 50, no. 7, pp. 072001, 2011.

[4] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[6] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[7] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010, pp. 807–814.

[8] DSIAC, "Atr algorithm development image database," .