Performance Evaluation of Boosted 2-stream TCRNet

Shah Hassan, Md Jibanul Haque Jiban, and Abhijit Mahalanobis

University of Central Florida, Orlando FL 32816, USA
shahhassan@knights.ucf.edu, jibanul@knights.ucf.edu, amahalan@crcv.ucf.edu

Abstract. Target detection in infra-red imagery is a particularly challenging problem due to the presence of terrain clutter. The TCRNet-2 CNN architecture was introduced to combat this issue, and has been shown to perform better than conventional networks such as faster RCNN and YOLOv3 . In this paper, we evaluate the performance of the boosted 2-stream TCRNet in detail (including robustness to range variations, performance under day and night conditions) and compare it with that of YOLOv5. A MWIR data set released by DSIAC is used for training and testing the network. We also propose the MWIR Target classifier that recognizes the 10 classes in the NVESD Dataset and achieves an accuracy of 65.72% which is state-of-the-art to date.

Keywords: TCRNet, Infrared Images, NVESD Dataset, Target Detection, Target Classification.

1 Introduction

Humans have the ability to detect, locate and classify objects in standard environments rather easily. The process is fast and accurate and it allows us to do all sorts of day to day tasks from interacting with our surroundings to any other complex task. Although current detection algorithms have shown great results in many vision related tasks, there is still a long way to go in many other tasks. Detecting targets at distant ranges in a highly cluttered environment in infra-red images is one of such tasks. It is even sometimes very difficult for humans to detect targets in such challenging environments. Researchers have been working on solving this task for quite some time [1, 2]; target detection of Infrared Imagery at low false alarm rates remains a challenging problem.

The TCRNet was introduced to specifically address the problem of finding targets in background clutter. A new loss function, referred to as the *target to clutter ratio* (TCR) was defined as the ratio between the output energies produced by the network in response to target and clutter. The network also employs analytically derived filters in its first layer that optimally represent target and clutter. Using these fixed filters in the first layers imposes strong priors on the rest of the network, forcing the convolution kernels to be learned such that the TCR metric is optimized. This paper builds on the previous work by rigorously analyzing the performance of the TCRNet using DSIAC MWIR data set to evaluate its performance at different times of day, its ability to detect targets at different ranges, and to compare its performance with the state-of-the-art YOLOv5 object detection network.

2 Two-Stream TCRNet: A Review

Figure 1 exhibits the architecture of two-stream TCRNet [4]. Two Stream TCRNet (TCRNet-2) is similar to the original version of TCRNet [3]; however as shown in Figure 1, TCRNet-2 has two separate channels to process the target and clutter information, which ensure the maximum discrepancy between the two sub-spaces, whereas the original TCRNet had only one channel. The number of filters (70 for the target stream and 30 for the clutter stream) are determined based on the dominant eigenvalues found using TCR metric. Two more convolution layers with fifty 3×3 filters are added to each stream and then the output of these two streams is combined. One last convolution layer is added to get the final combined output. Batch Normalization [5] and ReLU [6] are used in all layers. The local maxima in the output activation map are used to determine the target locations in the images.

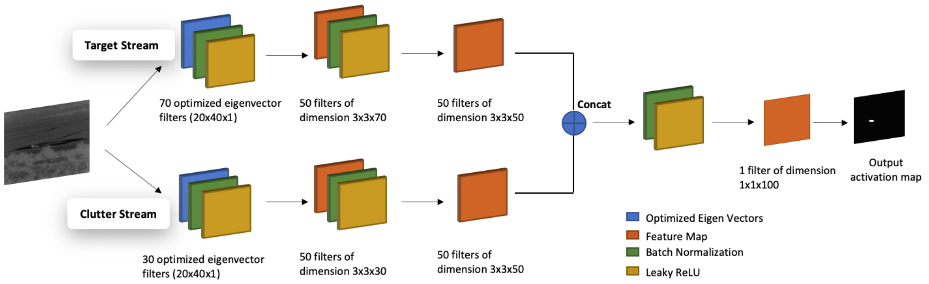
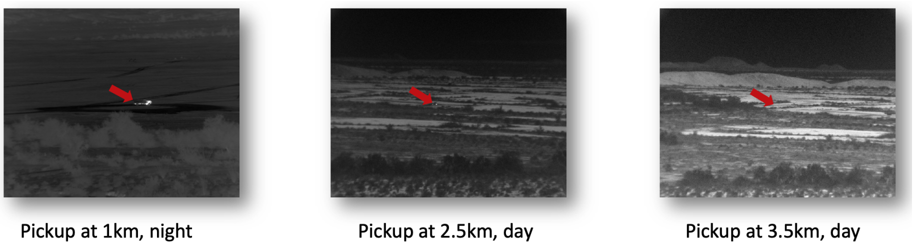

Target Stream
70 optimized eigenvector filters (20x40x1)
50 filters of dimension 3x3x70
50 filters of dimension 3x3x50
Concat
1 filter of dimension 1x1x100
Output activation map
Clutter Stream
30 optimized eigenvector filters (20x40x1)
50 filters of dimension 3x3x30
50 filters of dimension 3x3x50
Optimized Eigen Vectors
Feature Map
Batch Normalization
Leaky ReLU

Fig. 1: The Architecture of TCRNet-2 Network [4].

Boosted TCRNet-2 Networks: A Booster TCRNet model can improve the overall detection rate and reduce false alarm rate. In this process, the primary detector nominates the regions of importance (ROI) that may contain potential targets and the second network focuses only on the ROIs produced by the primary detector. The final detection score is found by adding the scores produced by the primary and secondary networks. Both the primary and secondary network have the same architecture however, the clutter training data is different. The primary detector is trained on target chips that are extracted from full-frame images using ground truth information and clutter chips that are randomly extracted from the same set of full-frame images. The second network uses the same target training chips, but the clutter chip set comprises only of *false* *positives* produced by the primary network. In essence, the clutter chip set for the second network is produced by applying the primary network to the images at ranges 4000m, 4500m, and 5000m images and extracting the false positive regions.


Pickup at 1km, night

Pickup at 2.5km, day

Pickup at 3.5km, day

Fig. 2: Example of full-frame images in dataset.

3 Experiments and Data Set

TCRNet-2 is trained and tested on Automated Target Recognition (ATR) database provided by DSIAC [7]. This dataset contains both visible and mid-wave infrared (MWIR) imagery of people and ten different vehicular (both civilian and military) targets. However, only MWIR imagery of vehicular targets is used to train and test the TCRNet-2 model. The ATR data were collected during both daytime and nighttime from a distance between 1km to 5km with 0.5km increment. The images are 640x514 in size. The dataset contains ground truth information of both target location and target class with much other useful information. This data can be downloaded from [7] which comes with a user reference guide that contains more details about the dataset.

3.1 Result: TCRNet-2 in different time Scenarios

The key performance metrics are the percentage of correct detections (P_d) and the number of false detections per square degree (FAR). Figure 3 shows the performance of TCRNet-2 in day, night and Day & Night scenarios. It is clear that the nighttime performance is better w.r.t both detection probability and false alarm rate. Specifically, at FAR of 2.0 TCRNet-2 is able to achieve P_d of 0.84 for Day time images and 1.00 for nighttime images. Similarly, at a FAR of 1.0, P_d is 0.8 for the day time images, and 1.0 for the night time images. It can be clearly seen from the night-time images that TCRNet-2 is able to quickly achieve the maximum P_d very quickly right after the FAS is 0.2. The reason why TCRNet-2 performance is better in the Night time scenario is that the night time images have a lot less challenging clutter than that of the day time images.

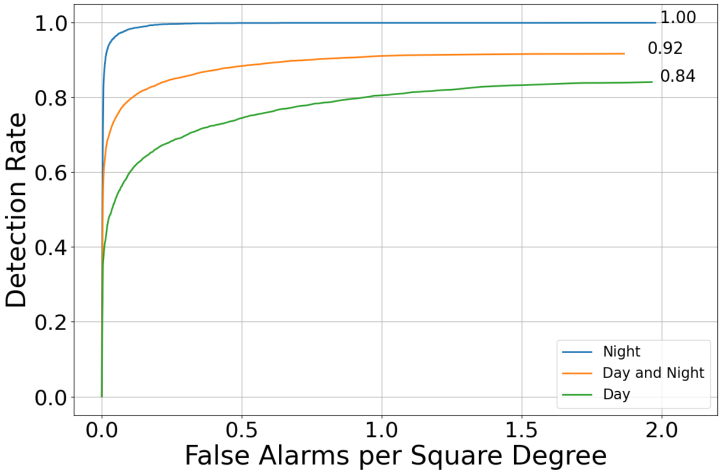


Fig. 3: ROC Curves of Boosted 2-stream TCRNet comparing day, night, & day and night times

3.2 Comparison with YOLOv5:

You Only Look Once (YOLO) [11] is a very popular method that does both detection and classification in one step. The latest version is YOLOv5 [12] that we used to compare with Boosted 2-stream TCRNet for target detection in the cluttered environment.

YOLOv5 is fine tuned on the NVESD dataset using the same training and testing protocols. The model is trained with batch size 8, image size 640x514, 300 epochs, and Adam optimizer with learning rate 0.001. We trained and tested YOLOv5 on Pytorch 1.8.1 using an NVIDIA GeForce RTX 2080 Ti - 11GB GDDR6 GPU.

Fig. 4 shows performances of YOLOv5 compared to boosted TCRNet-2. There are different sizes (s,m,l,x) of YOLOv5 models. We fine tuned both size s and x. Among these two models, size s is found to perform better for this dataset. The maximum Yolov5 P_d is 0.62 whereas at same false alarm rate the P_d for Boosted TCRNet-2 is around 0.78.

3.3 Range Invariance:

Fixed Detection Window Analysis: TCRNet-2 was tested on resized images for 2500m range with a detection window of 20 pixels radius. However, the manual resizing is not a desirable approach. Therefore, we trained a multiple streams TCRNet-2 on different ranges. The first experiment was conducted with three streams; first stream with training chips of range 1000m, second stream with training chips of range 1500m and the third stream for training chips of

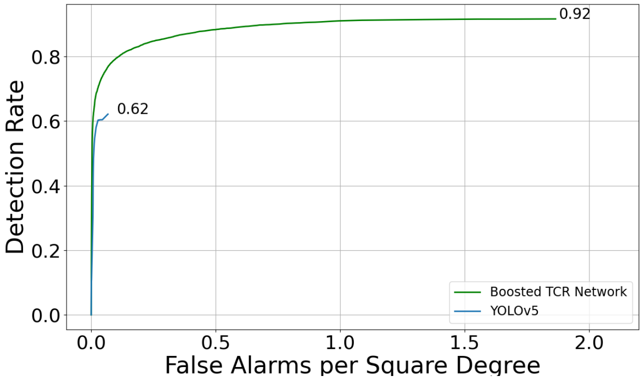


Fig. 4: ROC curves comparing performaces of Boosted TCRNet-2 and YOLOv5 both day and night test images. The boosted TCRNet curve shows significant better performance in terms of detection rate with a margin of 30% over YOLOv5.

range 2000m. The second experiment was conducted on 7 streams; with training chips of ranges 1000m, 1250m, 1500m, 1750m, 2000m, 2250m, 2500m. The third experiment used 3 ranges i.e. 1250m, 1750 and 2250 ranges.

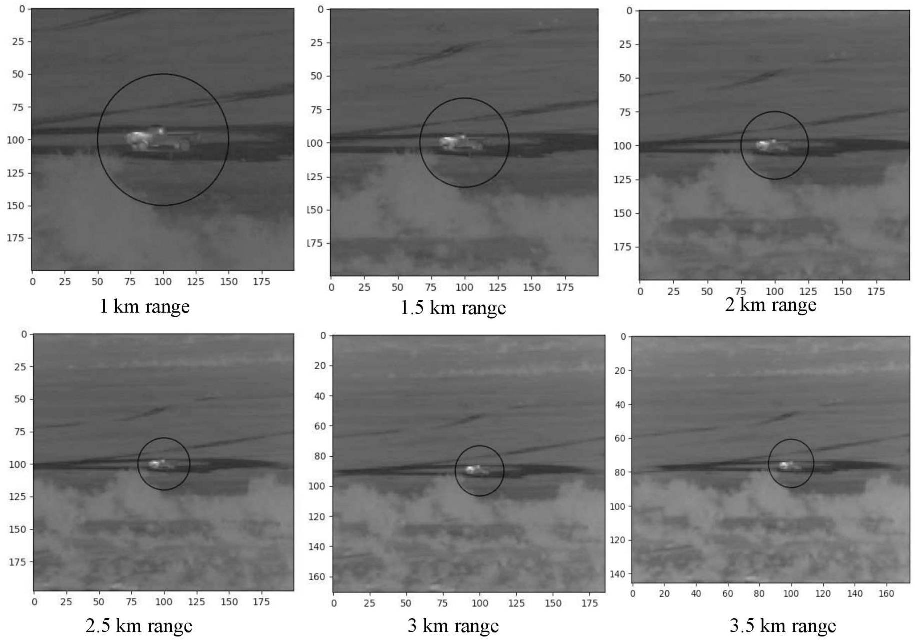
Results for these experiment show that having training images of various ranges can help in achieving almost equal performance on the actual non-resized images. However, defining a detection window for every range is crucial.

Variable Detection Window: As mentioned earlier, the testing images are manually resized for 2500m range. However, the manual resizing of the test images is not desired. Therefore, we present a formula to vary the detection radius with respect to the range given in equation 1. As seen in the Fig. 5, the formula provides a way to estimate the detection window with respect to the size of the target. When the range is 1km the radius is 50; it decreases to 33.3 as the range increases to 1.5km and it keeps on decreasing as the range increases. The radius is 25 for 2000m range, 20 for 2500m range, 16.7 for 3000m range and 14.3 for 3500m range. We found that the result without manually resizing the test data is comparable to the scaled images as shown in Fig. 6.

$$DetectionWindow = 2500/(Range) \times 20 \tag{1}$$

4 MWIR Target Classifier:

In order to classify the kind of vehicle in the test images, we also propose a MWIR Target Classifier. The classifier consists of an input layer, followed by 5 convolution layers, followed by a fully connected layer and a classification layer.



1 km range

1.5 km range

2 km range

2.5 km range

3 km range

3.5 km range

Fig. 5: Variable detection windows.

Fig. 7 depicts the architecture of the classifier. Each convolutional layer has the filter size of 3x3 and padding of 2. 65 filters are used in the first convolutional layer, 32 in the second layer, 64 in the third, and 32 filters are used in the 4th and 5th convolutional layer. 5 convolutional layers are followed by a fully connected layer with ReLu as activation function which is followed by another fully Connected Layer followed by Softmax as the activation function.

We train the MWIR Target Classifier on 32x64 chips with vehicular target at the center. The chips are extracted from images of the ranges 1km, 1.5km and 2km. We test the classifier on Boosted TCRNet-2 detections in order to complete the pipeline. MWIR Target Classifier achieves 65.72% accuracy which is state-of-the-art on NVESD dataset.

5 Conclusion

This paper evaluates performance of the boosted two-stream TCRNet for target detection in challenging clutter terrain in MWIR images. First, it significantly outperforms well-known deep learning technique YOLOv5 . We showed that the two-stream TCRNet is robust to range variations by testing it on unscaled test images. We also suggest to use a variable detection window ensuring that the targets at variable ranges are correctly detected. Moreover, we showed that a simple CNN classifier can achieve 65.72% accuracy on the NVESD dataset.

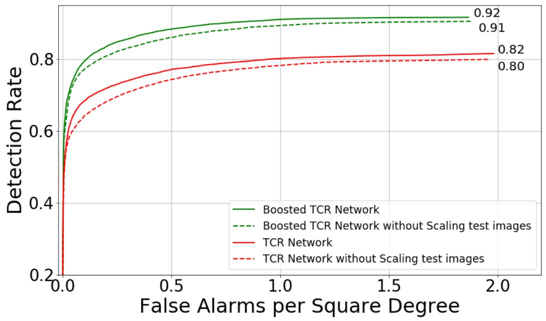


Fig. 6: The ROC curves with variable detection window

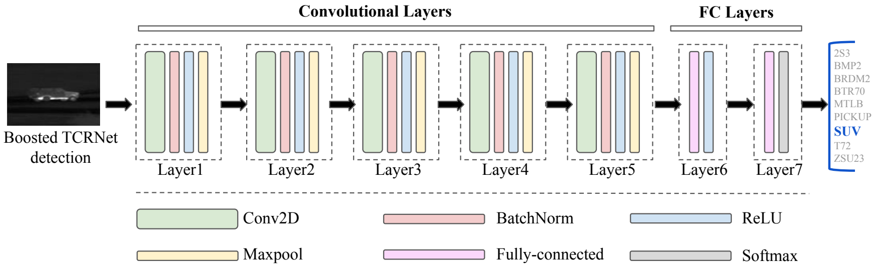


Fig. 7: Architecture of target classifier.

References

1. Ratches, J.A.: Review of current aided/automatic target acquisition technology for military target acquisition tasks. Optical Engineering, 50(7), p.072001 (2011).
2. Gundogdu, E. and Koç, A. and Alatan, A. A.: Automatic target recognition and detection in infrared imagery under cluttered background. In: Target and Background Signatures III (Vol. 10432, p. 104320J). International Society for Optics and Photonics (2017).
3. McIntosh, B., Venkataramanan, S. and Mahalanobis, A.: Infrared target detection in cluttered environments by maximization of a target to clutter ratio (TCR) metric using a convolutional neural network. In: IEEE Transactions on Aerospace and Electronic Systems, 57(1), pp.485-496 (2020).
4. Jiban, M.J.H., Hassan, S. and Mahalanobis, A.: Two-Stream Boosted TCRNet for Range-Tolerant Infra-Red Target Detection. In: IEEE International Conference on Image Processing (ICIP) (pp. 1049-1053). IEEE. (2021).
5. Ioffe, S. and Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning (pp. 448-456). PMLR (2015).
6. Nair, V. and Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML (2010).

7. DSIAC: ATR Algorithm Development Image Database. https://dsiac.org/databases/atr-algorithm-development-image-database/
8. Reynolds, W.R.: Toward quantifying infrared clutter. In Characterization, Propagation, and Simulation of Infrared Scenes (Vol. 1311, pp. 232-240). International Society for Optics and Photonics (1990).
9. Redmon, J. and Farhadi, A.: Yolov3: An incremental improvement. In arXiv preprint arXiv:1804.02767 (2018).
10. He, K., Gkioxari, G., Dollár, P. and Girshick, R.: Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969) (2017).
11. Redmon, J., Divvala, S., Girshick, R. and Farhadi, A.: You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788) (2016).
12. YOLOv5-5.0. Ultralytics. https://github.com/ultralytics/yolov5 (2020). Accessed 26 July, 2021.