

DETECTION AND TRACKING OF SEARCH AND RESCUE PERSONNEL UNDER HINDERED LIGHT CONDITIONS USING HYPERSPECTRAL IMAGING

*Lennert Antson*¹, *Arthur Vandenhoeke*², *Michal Shimoni*³, *Charles Hamesse*^{1,2}, *Hiep Luong*¹

¹ IMEC-IPI-URC research group, TELIN, Ghent University, Ghent, Belgium

² Royal Military Academy (RMA), Brussels, Belgium

³ Kuva Space, Espoo, Finland

ABSTRACT

In the past two decades, advances in cost-effective fabrication techniques, miniature systems, intrinsic detectors and computing technology transformed hyperspectral imaging (HSI) from a bench-top scientific curiosity to a discipline with operational and fielded systems. Despite these advancements, HSI still shows limitations in multiple degraded visual environment (DVE) conditions including rain, haze and smoke. These conditions adversely affect the performance of the sensors by reducing their range of effectiveness in terms of detection, identification, and recognition. Additionally, benchmark detectors and classification techniques are mainly conducted on a pixel-wise basis which has proven to be a solid and effective technique, but often isn't optimized for real-time detection of people or objects. This paper proposes a two-stage neural-net-based detector for real-time detection and tracking of firefighters subjected to DVE using hyperspectral near-infrared (NIR) images. The proposed method first exploits the spatial features through object detection and then analyzes the extracted regions using a spectral-spatial pixel-wise algorithm to validate the presence of firefighters. The robustness of the proposed architecture is successfully demonstrated using various realistic scenarios.

Index Terms— Hyperspectral, detection, tracking, deep learning, Nanodet-Plus, HybridSN, degraded visual environments, near-infrared, real-time

1. INTRODUCTION

Search and Rescue (SAR) operations on land after natural or anthropogenic disasters are frequently affected by smoke, rain, and at sea, they are occluded by mist and haze. To reduce mortality after a disaster, it is crucial to develop methods that overcome Degraded Visual Environment (DVE) conditions and to guarantee rapid detection of survivors and rescue personnel such as firefighters. Hyperspectral imaging (HSI) has proven extremely useful due to its ability to uniquely classify material based on its chemical composition by exploiting the measured spectral signature. Despite the technological maturity HSI has reached in the last decades, it is not applied

in SAR operations due to its limitation in DVE. In this paper, we focus on the real-time detection and tracking of firefighters subjected to DVE, by applying deep learning methods on HSI captured in the near-infrared (NIR) range.

Initially, target detection algorithms such as the Spectral Angle Mapper (SAM), Orthogonal Subspace Projection (OSP), Matched Filter (MF) and Adaptive Cosine Estimator (ACE) have been introduced to classify pixels based on a reference spectrum of the target. Unfortunately, these detection algorithms cannot adapt well to the variation of the spectral signature of the target or its background. Later, machine learning techniques such as the Gaussian Maximum Likelihood Classifier (GMLC) and Support Vector Machine (SVM) have been developed, resulting in increased classification accuracy [1], [2]. However, they also could not overcome light changes caused by DVE or noise induced by the hyperspectral sensor. In the last couple of years, neural networks, and more specifically, convolutional neural networks (CNNs) have been employed in spectral based target detection showing significant improvement in performance. Hu et al. [3] introduced a simple but effective 1D-CNN convolving over the spectral dimension, to outperform SVMs. Slavkovikj et al. [4] implemented a 2D-CNN extracting both spectral and spatial features from the 3x3 neighborhood window of the pixel. More recently, Roy et al. [5] proposed HybridSN, a spectral-spatial 3D-CNN followed by a spatial 2D-CNN, resulting in state-of-the-art performance on various datasets. However, since each pixel in the HSI is detected separately, these pixel-wise detection algorithms are extremely computationally intensive. HybridSN, for example, requires 4.8 seconds (on a GTX 1060 GPU) to classify an Indian Pines (IP) image with a spatial dimension of 145x145 pixels and input patches of shape 25x25x30, making real-time detection challenging.

In this paper we propose a two-stage detector (Figure 1), capable of detecting and tracking firefighters subjected to DVE in real-time hyperspectral NIR images. By using object detection in the first stage, regions of interest (bounding boxes) are extracted. The second stage of the model validates these extracted regions by applying pixel-wise detection. The

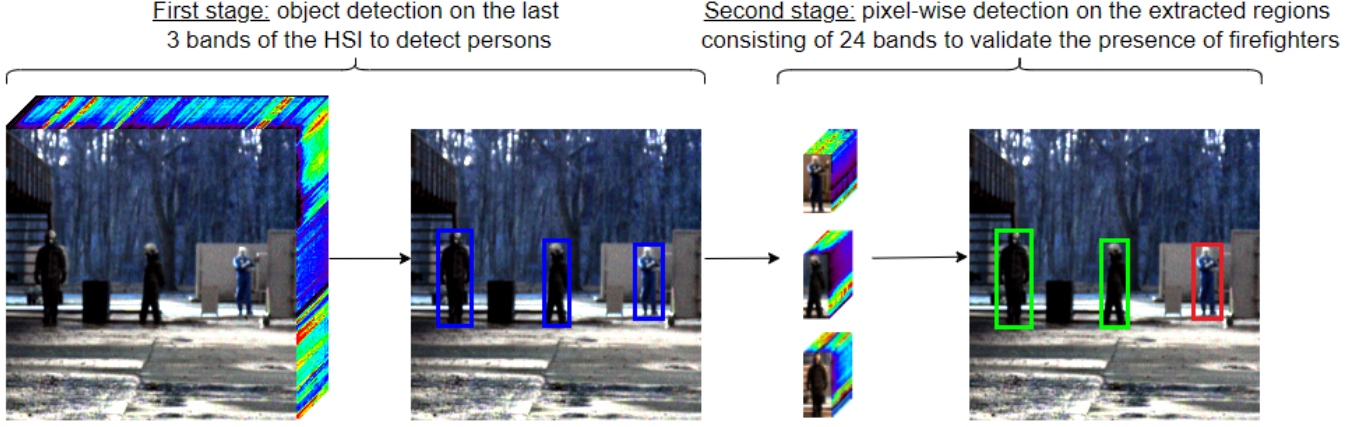


Fig. 1: Overview of the proposed two-stage detector. In the second image, proposed bounding boxes are shown in blue. In the last image, valid and rejected bounding boxes are shown in green and red, respectively.

pixel-wise detector is based on an existing state-of-the-art architecture but uses a smaller window size allowing faster detection. Depending on the fill fraction of detected pixels in the classification map of this region, the bounding box is either kept or discarded.

0 and 1 using the tanh-function instead of the traditional z-score. A benefit of the tanh-normalization method is that it reduces the impact of pixels with high reflectance values that cannot be considered as outliers.

2. STUDY AREA AND PRE-PROCESSING

In a measurement campaign conducted in February 2021, Belgian Royal Military Academy (RMA) collaborated with IMEC to research hyperspectral capabilities in real-time applications under DVE. A hyperspectral snapshot imager, that operates in the NIR range (667 - 941nm), scanned the scene with 24 spectral bands with an average of 25 hyperspectral images per second. Multiple realistic scenarios were captured for various DVE conditions (dark smoke, white smoke, haze and heavy rain) in both static and dynamic settings, forming a total of eight different scenarios. In the static scenarios, firefighters remained stationary, while in the dynamic scenarios, they moved from left to right and back again. The four investigated DVE conditions are presented in Figure 2.

The calibration and pre-processing of the HSI consists of multiple steps [6]. First, a dark reference image is subtracted from the scene image to eliminate the impact of dark noise induced by the sensor. To correct the slight spatial variations across the sensor caused by the CMOS imager and lens vignetting, a calibration measurement is performed using a white reflectance tile. This allows to compute a non-uniformity cube to correct the HSI. Next, a spectral correction method is performed to reduce second-order harmonics and distortions in the spectral responses. These steps are followed by an angularity correction to correct the effects of spectral shifting.

Due to the skewed and non-Gaussian pixel distribution of the NIR images, all HSI in this study are normalized between



Fig. 2: False RGB color NIR images of shape 208x208x24 for each type of simulated DVE.

3. METHODOLOGY

3.1. Object detection

In the first stage of our proposed model, an object detection algorithm is trained. Initially, the YOLOv3-Tiny [7] architecture was implemented and tested. It is a lightweight object detection model able of making predictions at two different scales. Since the model makes use of anchor boxes in the training stage, these were calculated on the training set using K-means clustering. However, since in the dataset all firefighters and persons have approximately the same shape, there was little to no variation in the calculated anchor boxes. Training with these newly calculated anchor boxes did not produce satisfactory results, and neither did training with the default anchor boxes. Therefore, we decided to adopt an anchorless object detection approach such as NanoDet-Plus [8]. This one-stage anchorless object detection model uses ShuffleNet V2 [9] as the backbone, introducing depth-wise convolutions and channel shuffle operations with the direct purpose of improving detection speed. The outputs of the backbone are connected to a Feature Pyramid Network (FPN) called Ghost-PAN, which uses Ghost modules [10] to enhance multi-layer feature fusion. Finally, detections are done using four NanoDet-Plus 'heads' allowing to make predictions at multiple scales. Due to the optimizations such as depth-wise convolutions, and the small size of the model (4,163,661 parameters), it is capable of making predictions at high speed.

The object detector in the first stage, focuses on the spatial features in the images and predicts regions where a possible firefighter is located. The last 3 bands of the NIR images are selected as input to the NanoDet-Plus model. As illustrated for heavy white smoke in Figure 3, the spectral wavelengths of these bands show better penetration through the DVE. Annotations were made for all types of persons and not only people wearing firefighter uniforms.

3.2. Pixel-wise detection

In the first stage, persons were detected using bounding boxes. In the second stage, these bounding boxes are used to extract HSI or regions consisting of the original 24 spectral bands. Next, pixel-wise detection is performed on the extracted regions to verify the presence of a firefighter.

By applying pixel-wise detection only to the extracted regions instead of the whole image, the computational load is significantly reduced. This model is trained on the spectra of the firefighter's uniform. Due to the complex mixture of different types of fabric in the firefighter's uniform and thus different spectral responses, the model has to rely on both the spatial and spectral features. Especially when the firefighter is subjected to DVE, it is hard to rely only on spectral features to classify a pixel. Inspiration for the pixel-wise detection algorithm is taken from the HybridSN model. Originally, HybridSN uses input samples of shape $25 \times 25 \times K$ with

K representing the number of bands. This fairly large spatial window size of 25×25 resulted in excellent results but is relatively slow due to the large number of parameters in the model. For this reason, multiple spatial window sizes were tested and compared. To make a final decision about the validity of the extracted regions by the object detection model, the fill fraction of firefighter-classified pixels in the classification map of the extracted region is measured. If the fraction is larger than a certain threshold, the bounding box is kept, otherwise it is discarded.

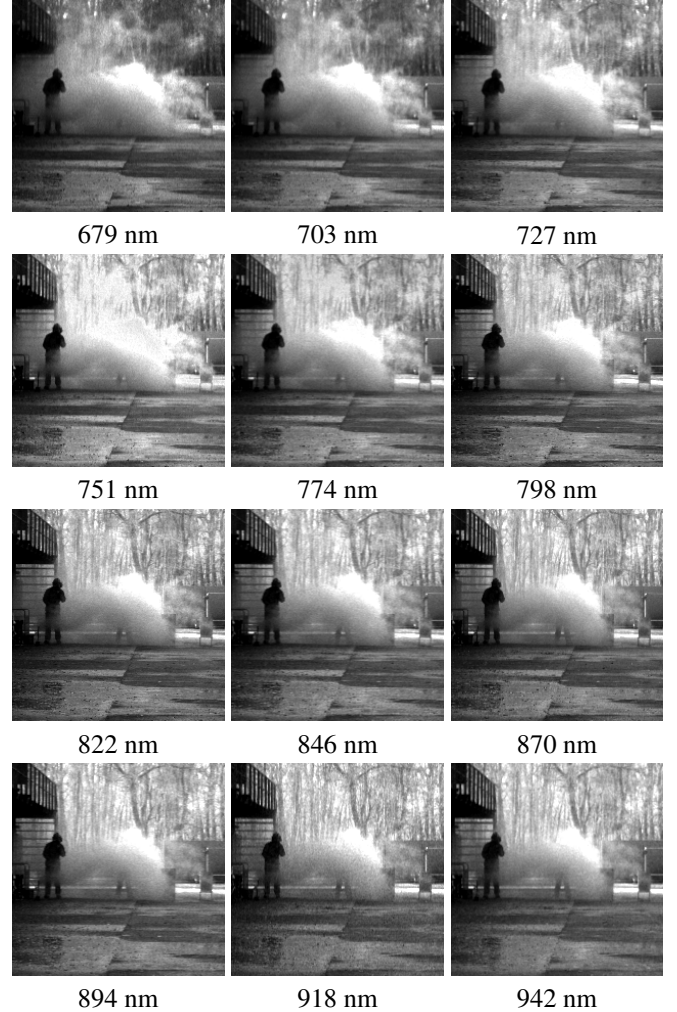


Fig. 3: Penetration through heavy white smoke shown for bands 2, 4, 6, ..., 24. Two firefighters are present in the NIR image, one on the left and one in the center behind the smoke.

4. EXPERIMENTAL RESULTS

4.1. Object detection

The captured NIR images have a spatial dimension of 407×215 but were cropped to 208×208 for training purposes. In total

Table 1: Overview of the object detection results.

Model	AP ₅₀	AP ₇₅	Inference time (ms)
NanoDet-Plus	0.9528	0.8243	13.8
YOLOv3-Tiny	0.8558	0.4997	22.0

1886 images from all eight scenarios including the four types of DVE were used. Annotations (bounding boxes) were made for all persons in the images. The training-test data, contained a proportional amount of images from each type of DVE, and has been respectively split using the following ratio: 75-25%.

The YOLOv3-Tiny model was trained during 200 epochs with various batch sizes (32, 64 and 96) using the Adam optimizer with a learning rate of 0.001. Transfer learning and multiple augmentations were applied such as a horizontal flip, rotation and scaling. Due to the limited variation in the anchor boxes calculated prior to training, augmentation did not improve the results. The loss failed to converge in many cases, resulting in difficulties while training. Next, the NanoDet-Plus model was trained during 300 epochs with a batch size of 96. Multiple augmentations were implemented such as saturation, brightness, contrast, scaling, horizontal flip and translation. This model was trained using the Adam optimizer with a weight decay of 0.05 and a learning rate of 0.001.

The results, measured on the test set, are shown in Table 1. To evaluate the models, the average precision (AP) at 50% and 75% Intersection-over-Union (IoU) is used. Average precision is the area under the precision-recall curve, evaluated at a certain IoU threshold. The inference time is measured by iterating over the test set and is the average time in milliseconds, to predict the bounding boxes in a single image. The results of NanoDet-Plus are superior, achieving an AP@50 of 0.9528, AP@75 of 0.8243, and requiring only 13.8 ms on average to detect persons in an image.

4.2. Pixel-wise detection

Training and test data of the firefighter uniforms were collected for all types of DVE and without the influence of DVE. In total, 647K firefighter uniform samples and an equal number of random background samples were collected from 812 images. Just as for the object detection, a 75-25% train-test split was employed. Models were trained and compared for various window and batch sizes. A relatively large batch size of 1024 reduced the effects of overfitting and allowed more stable training. All models were trained using the Adam optimizer with a learning rate of 0.0001 during 400 epochs. The results are reported in Table 2. To evaluate the models, the accuracy and Cohen’s Kappa score are used. To determine the model’s speed, the average bounding box size was calculated from the annotated bounding boxes. On average, a bounding box produced by NanoDet-Plus consists of 1034 pixels. This means that the measured inference time in Table 2 represents

Table 2: Overview of the pixel-wise detection results for different window sizes for the HybridSN model.

Window size	Accuracy	Kappa	Inference time (ms)
9x9	0.8647	0.7292	6.0
11x11	0.8862	0.7722	12.4
13x13	0.9678	0.9356	21.3
15x15	0.9780	0.9561	31.2
15x15 (reduced)	0.9789	0.9579	30.8

the amount of time needed to validate a single bounding box. The measured time is an average taken by iterating through the test set with batch size 1034.

The results demonstrate that smaller window sizes produce lower classification accuracy but faster predictions. The best performance is observed for the model with a spatial window size of 15x15. Since a smaller window size is employed than in the original HybridSN model, experiments were also conducted on smaller versions of HybridSN with fewer parameters. Best results were achieved on the 15x15 window-sized model with a reduction in the number of neurons in the dense or fully connected layers, of which the architecture is shown in Table 3. By halving the number of neurons from 256 to 128 in layer *dense_1*, and from 128 to 64 in layer *dense_2*, the number of parameters are reduced from 1,077,489 to 651,249, for input patches with 24 bands. This resulted in a slightly higher accuracy and Kappa score.

4.3. Two-stage detector: object and pixel-wise detection combined

The complete architecture was tested with the NanoDet-Plus detector in the first stage, and the reduced 15x15 window-sized HybridSN model, in the second stage. A total of 40 randomly selected HSI were used, representing all types of DVE, consisting of 63 firefighters and 17 non-firefighter personnel wearing other types of clothing. An object confidence threshold of 0.5 was employed. To decide the threshold for the fill fraction for the pixel-wise detection, we experimented with different values. A fill fraction of 0.3 led to the best results. Out of the 63 firefighters, 62 were correctly detected. The single false negative was a firefighter at the border of the image only being half-visible. Out of the 17 non-firefighters, one person was mistakenly classified as a firefighter due to the large overlap with another firefighter. The speed of the two-stage detector was measured and runs at 12.4 FPS.

Combining object detection and pixel-wise detection as in the proposed two-stage detector creates a robust architecture. If there are many false negative detections in the first stage of the model, the object confidence score could potentially be thresholded to a lower value. This results in higher recall in return for lower precision. But this is not an issue since the pixel-wise detector in the second stage, can discard the

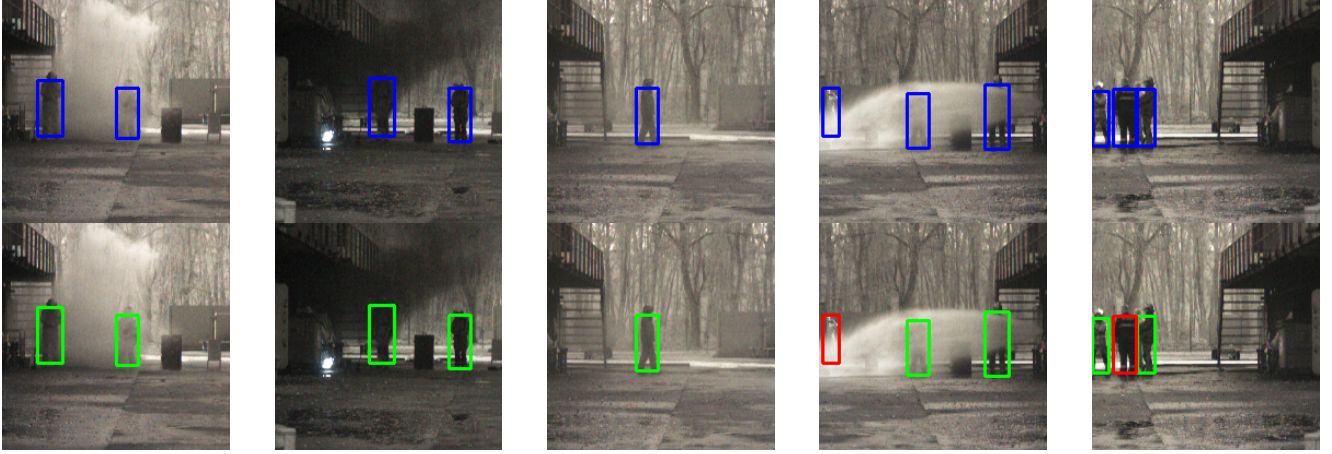


Fig. 4: Results of the two-stage model for various types of DVE. In the top images, proposed bounding boxes are shown in blue. In the bottom images, valid and rejected bounding boxes are shown in green and red, respectively. The two bounding boxes in red are correctly rejected, because in both cases the persons do not wear firefighter uniforms: in the penultimate image the person wears normal civilian clothing and in the last image, the person wears a Belgian defence uniform.

falsely proposed regions. All experiments in this paper were conducted on an RTX 2080 GPU and 32 GB of RAM.

Table 3: Adapted HybridSN with an input shape of 15x15x24 and reduction in the number of neurons in the dense layers.

Layer (type)	Output Shape	#Parameters
input_1 (Input)	(15, 15, 24, 1)	0
conv3d_1 (Conv3D)	(13, 13, 18, 8)	512
conv3d_2 (Conv3D)	(11, 11, 14, 16)	5776
conv3d_3 (Conv3D)	(9, 9, 12, 32)	13856
reshape_1 (Reshape)	(9, 9, 384)	0
conv2d_1 (Conv2D)	(7, 7, 64)	221248
flatten_1 (Flatten)	(3136)	0
dense_1 (Dense)	(128)	401536
dropout_1 (Dropout)	(128)	0
dense_2 (Dense)	(64)	8256
dropout_2 (Dropout)	(64)	0
dense_3 (Dense - Output)	1	65

Total Trainable Parameters: 651,249

5. CONCLUSIONS

In this paper, a two-stage detector is proposed able of detecting and tracking firefighters under multiple DVE conditions in hyperspectral NIR images. Despite the complex nature of the spectra of the firefighter’s uniform in DVE, excellent results were achieved using the proposed model. By combining NanoDet-Plus with a reduced version of HybridSN with a smaller window size of 15x15, we created a robust detec-

tor able of tracking firefighters in realistic DVE scenarios, in real-time (12.4 FPS) and with high accuracy. Out of the 63 firefighters present in 40 randomly selected HSI, 62 firefighters were successfully detected, confirming the working of the proposed two-stage detector.

6. REFERENCES

- [1] Davood Akbari, Mohammad Reza Saradjian, and Mina Moradizadeh, “Building detection using hyperspectral images by support vector machines,” 03 2012.
- [2] Gaurav Hegde, J Mohammed Ahamed, R Hebbar, and Uday Raj, “Urban land cover classification using hyperspectral data,” *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 8, no. 8, pp. 751–754, 2014.
- [3] Wei Hu, Yangyu Huang, Li Wei, Fan Zhang, and Hengchao Li, “Deep convolutional neural networks for hyperspectral image classification,” *Journal of Sensors*, vol. 2015, pp. 1–12, 07 2015.
- [4] Slavkovikj, Viktor and Verstockt, Steven and De Neve, Wesley and Van Hoecke, Sofie and Van de Walle, Rik, “Hyperspectral image classification with convolutional neural networks,” in *PROCEEDINGS OF THE 2015 ACM MULTIMEDIA CONFERENCE*. 2015, pp. 1159–1162, ACM.
- [5] Swalpa Kumar Roy, Gopal Krishna, Shiv Ram Dubey, and Bidyut B. Chaudhuri, “HybridSN: Exploring 3-d-2-d CNN feature hierarchy for hyperspectral image classi-

fication,” *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 2, pp. 277–281, feb 2020.

- [6] Kathleen Vunckx and Wouter Charle, “Accurate video-rate multi-spectral imaging using imec snapshot sensors,” in *2021 11th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2021, pp. 1–7.
- [7] Joseph Redmon and Ali Farhadi, “Yolov3: An incremental improvement,” *arXiv*, 2018.
- [8] RangilYu, “Nanodet-plus: Super fast and high accuracy lightweight anchor-free object detection model,” <https://github.com/RangilYu/nanodet>, 2021.
- [9] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun, “Shufflenet v2: Practical guidelines for efficient cnn architecture design,” 2018.
- [10] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu, “Ghostnet: More features from cheap operations,” 2019.