A Deep Learning Framework for Detection of Targets in Thermal Images to Improve Firefighting[⋆,⋆⋆]

MANISH BHATTARAI[a,b,*,1], MANEL MARTÍNEZ-RAMÓN[a]

[a]Department of Electrical and Computer Engineering, The University of New Mexico, New Mexico 87106, USA
[b]Los Alamos National Laboratory,New Mexico,USA

ARTICLE INFO

Keywords:
Deep Convolutional Neural Networks
Infrared Images
Firefighting Environment
Firefighters
Situational Awareness.

ABSTRACT

Intelligent detection and processing capabilities can be instrumental to improving the safety, efficiency, and successful completion of rescue missions conducted by firefighters in emergency first response settings. The objective of this research is to create an automated system that is capable of real-time, intelligent object detection and recognition and facilitates the improved situational awareness of firefighters during an emergency response. We have explored state of the art machine/deep learning techniques to achieve this objective. The goal for this work is to enhance the situational awareness of firefighters by effectively exploiting the information gathered from infrared cameras carried by firefighters. To accomplish this, we use a trained deep Convolutional Neural Network (CNN) system to classify and identify objects of interest from thermal imagery in real time. In the midst of those critical circumstances created by structure fire, this system is able to accurately inform the decision making process of firefighters with real-time up-to-date scene information by extracting, processing, and analyzing crucial information. With the new information produced by the framework, firefighters are able to make more informed inferences about the circumstances for their safe navigation through such hazardous and potentially catastrophic environments.

1. Introduction

The application of CNN technology abounds in the Surveillance and Defense fields [4, 8, 16, 26] but very little research is documented in applying these principles to overcoming the navigational challenges faced by firefighters in live fire events. In fact, current firefighting modalities do not involve any automated detection mechanism and the target is identified solely by the firefighter. Detection processes can be adversely affected by environmental factors inherent in active fire scenes. High temperatures, near zero visibility caused by debris, smoke and lack of lighting, and a continuously changing environment can combine to disorient and further inhibit decision making processes, affecting even experienced firefighters. Under such hazardous conditions, lives can be lost due to rescue operation decisions based on incomplete or inaccurate understanding of the most current environmental conditions within the structure. Federal Emergency Management Agency studies [1] show a majority of firefighter mortalities reported, resulted from inefficient decision making protocol. Heightened anxiety levels leading to misinterpretation of the scene, as well as lack of a complete understanding of the environment are cited as factors. We propose an Artificial Neural Network-based system capable of autonomously identifying objects and humans in the scene of the event in real time to improve on-the-ground knowledge that dictates decision making protocol. The artificial intelligence (AI) based results can be used to assist in reducing these mortality statistics by minimizing anxiety induced errors. The AI-based system is also capable of accurately differentiating between human postures. This posture detection can assist firefighters in prioritization of rescue of identified victims through estimation of their health condition based on their posture.

In this paper, we demonstrate a CNN-based autonomous system capable of generating information that can improve situational awareness for firefighters regarding the environment into which they are deployed. The information is generated by classifying fire, objects of interest like doors, windows, and people and other thermal conditions using infrared video that is actively recorded by firefighters on scene. The CNN system can detect and classify desired targets and relay the information back to firefighters, thereby providing crucial information necessary to informing important planning decisions. This enables the firefighters and their commanders access to data collected and processed through an unbiased lens that is both comprehensive and reliable. The improved knowledge of local events and changes across the scene allows leadership to completely assess the local critical conditions and make appropriate decisions based on real time conditions. The improvement in situational awareness provided by the reliable stream of information deduced by the CNN could also assist disoriented firefighters to choose a safer path in a fire environment by autonomously identifying and alerting firefighters to the presence of objects of interest such as doors, windows, human targets, excessive smoke, etc that they may have overlooked in their confusion.

The conventional firefighting system uses different sensors such as temperature, UV, and fire detectors to determine the presence of fire, smoke, and other hazards [12, 17].

⋆This work has been supported by NSF S&CC EAGER grant 1637092.
*Corresponding author
ceodspspectrum@unm.edu (M. BHATTARAI); ceodspspectrum@lanl.gov (M. BHATTARAI)
ORCID(s): 0000-0002-1421-3643 (M. BHATTARAI)
[1] https://www.usfa.fema.gov/downloads/pdf/publications/ff_fat17.pdf

Their long-established usage is evidenced by the prevalence of such sensors in all buildings, and is a requirement in building codes, to generate timely response for first responders. However, such detectors typically have a long response time in large spaces [27]. Furthermore, they do not provide any spatial information regarding the presence of hazards in the given scenario. More contemporary firefighting modalities for detection of fire, smoke and other targets in a fire environment rely on color [2, 21, 27, 31], motion [3, 19] and texture [20, 31] features of the captured image. These vision-based approaches use histogram thresholding, optical flow- based motion vector computations, and texture analysis. These more modern techniques provide enhanced performance using RGB imagery and are an improvement over the more conventional techniques described above. These algorithms do not perform as robustly on Infrared (IR) or darker imagery and also require longer computational time. IR imagery lacks sufficient complexity needed by such algorithms to perform well. Conversely, RGB images are hard to classify in active fire environments due to heavy smoke and poor lighting. IR image technology fills this gap. Furthermore, the presence of fire and smoke, by its very nature, creates a non-stationary environment and renders most existing stationary vision-based detection systems ineffectual in informing decision-making processes in real time. To address this issue, a robust real-time detection system based on CNN is proposed which is able to detect and localize the target of interest instantaneously. The research presented found in [13], describes the usage of infrared images to extract motion and statistics features in real time using a Bayesian classifier for multi-class identification. Significant research has also been done in human detection in other dark/ low visibility environments utilizing a single visible camera and fusion of the generated RGB image with an IR data set. Single IR camera based detection mechanisms have been presented in [9, 10, 22, 32]. Most of these approaches use HOG-based feature extraction and a classifier using SVM or other ML-based techniques [6, 22]. Paper [9] presents the use of a GMM system in human detection. Template matching techniques and thresholding techniques have also been reported in [10, 11, 22].

Other published works that have influenced our research perform classification tasks on thermal imaging. In [5], a CNN is used in material recognition with non-firefighter grade thermal cameras. A related study is reported in [29], where transfer learning is used to detect objects in a fireground. Nevertheless, the number of data samples used in the Vandecasteele paper is low, so deep learning cannot be applied. Saliency detection and convolutional neural networks are applied to detect wildfires in [36]. In [14] a Bayesian procedure to detect fire and smoke and discriminate them from thermal reflections in infrared is used in [33]. In the recent work [1], authors introduce a methodology based on Random Markov fields to segment fire, smoke and background in a sequence of images.

Further works that use thermal imagery and deep learning include [34], where authors use a CNN to detect known objects in infrared surveillance cameras. In [30], a CNN is applied to the detection of vehicles in thermal imagery. Researchers in [7] introduce the use of CNNs to detect pedestrians in order to apply detection to unmanned aerial vehicles (UAV). Another application in UAV is presented in [23], where authors train a CNN structure to detect objects of interest, such as bodies or body parts and other objects related to victims of avalanches. A similar approach uses CNNs on long wave infrared imagery to detect objects. Work [28] uses deep learning to detect people with a semi-supervised approach that takes advantage of a large quantity of non-labeled images containing humans.

In spite of the large quantity of works related to the processing of infrared images in fireground or related scenarios, to our knowledge, no work has been published that attempts to construct a system that integrates online detection of targets of interest in these scenarios, including humans, objects, poses or the presence of fire, with a large quantity of images recorded in real fire training situations by firefighters. Thus, there is a need for effective automatic target detection generated in real time in a firefighting environment as well as the associated need for a highly accurate classifier. Our research seeks to address these needs. To do so, we have adapted and enhanced an existing state of the art CNN based automatic classifier system to improve its efficacy in identifying and classifying humans and objects of interest in real time in a firefighting environment. Also, we have improved upon the creation process of a data set so that it may be used to effectively train the neural network (NN) system. We have also trained the system to detect objects and humans simultaneously. Prior technique capabilities were limited to one or the other. To further assist rescue operations, we also added a posture detection element in which the CNN further distinguishes whether a person is in a sitting, crawling, prone, or upright position. The intent of this detection series is to allow rough estimates of health condition or panic state of the victim to be approximated and used in the prioritization of rescue operations.

Panic detection work has also utilized machine learning techniques to analyze video footage, typically acquired from stationary surveillance cameras. Two papers provide excellent summaries of the techniques authored to date, focused on crowd analysis and panic detection. [35] covers crowd state detection methods utilizing stationary RGB video surveillance footage capable of detection at micro/macro levels to analyze local/global individual movements in the frame. The analyses usually focus on small subsets of a crowd which are then aggregated together to achieve global attributes. The analyses summarized discussed two possible approaches in their frameworks, physics-based or machine learning-based. Physics-based models use collected motion such as velocity, correlation function, fluid dynamics, energy and entropy, force model and complex systems. Machine learning-based models utilize features extracted through signal processing or computer vision tools to detect crowd state. The panic detection algorithms are applied to three different types of datasets which are generated under three situations: 1) con-

trolled experiment 2) crowd model and simulation 3) crowd video surveillance. [15]covers crowd state detection and compares a stationary crowd (the body movements of people who do not move) to a dynamic crowd (people moving from one place to another). The analysis on dynamic crowds focuses on movement patterns of the individuals in the scene to infer activity. The motion vectors detected by classical image processing techniques such as frame difference or optical flow are analyzed to deduce crowd activity. Information is then gathered by processing the frames for a dynamic crowd. The camera position is required to be fixed to obtain realistic motion features. [15] found that very few studies have focused on panic detection within stationary crowds due to the challenges involved in detecting panic behaviors from small body movements such as expression or posture. However, in both crowd types, the fixed position of the camera is a requirement. These methodologies also require the analysis of a sequence of frames processed together to compute the motion vectors. In both papers, panic behavior is largely based on the rate of displacement computed for extracted features from sequences of frames and requires the camera to be stationary.

The primary focus of our work is object detection of key features like entrance/exit points (doors and windows) and persons needing rescue within an active fire environment. We also add a posture detection framework to assist in prioritization of rescue. Due to the nature of generation of our data (a handheld thermal imagery camera carried by one of the responding firefighters throughout the scene and thus non-stationary) the pose detection is limited to basic poses (sitting, standing, crawling) that are easily distinguishable in individual frame analyses and are thus not dependant upon the surrounding images in a sequence. Furthermore, these largely different poses are not reliant on high levels of detail and thus can be discerned in IR imagery. The application of the algorithms deployed in the above mentioned panic behavior detection research are difficult to deploy on our dataset because the moving camera necessary to complete our research objectives, would produce a much noisier motion vector for the targets and those movements needed to deduce panic state would be indistinguishable from those induced by the camera movements. To minimize such errors induced by camera movements in our dataset, we process one frame at a time and infer activity rather than relying on motion vectors determined from a sequence of frames analyzed together.

Thermal imaging cameras are the most widely used cameras for obtaining footage and circumventing visibility issues within a fireground. Albuquerque Fire Department(AFD) and Santa Fe Fire Department keep at least one camera in a truck at all times. The National Fire Protection Association also supports their usage over RGB-based cameras, which are dependant on light sources for the clear depiction of imagery. In active fire scenes, where electricity may be out and no natural lighting is available, the thermal signature picked up by infrared cameras becomes crucial. Smoke and debris further reduce clarity of view for an RGB-camera but thermal cameras are able to cut through both if a heat source lies beyond the walls of smoke and dust. However, thermal imaging cameras capable of surviving the temperatures found in buildings ablaze, are costly and are not standard issue for every firefighter to carry. Instead, one per fire crew is deployed. Thus, creation of a system that can maximize the thermal imagery available, process it and return accurate, real-time information back to every firefighter on scene would dramatically enhance the combined situational awareness of the responders as a unit. Our research is restricted here to thermal imaging, but it can be extended to RGB or UV imaging as well.

2. System Description

The convolutional neural network presented is based on the structure of the VGG16 neural network presented in [25]. Several different depths of the neural network have been tested, ranging from 1 to 5 convolutional sections as shown in Fig. 1, all of them followed by a fully connected section of two layers. For the depth-1 configuration, the CNNs have an input of dimension 224×224. The corresponding IR images are scaled to this size. They are convolved by 3×3 filters to produce 64 channels of dimension 224×224. The resulting outputs are then passed through a set of ReLU activations. The process is repeated with an identical convolution, ReLU and then pooled to produce 64 channels of dimension 112×112 pixels. Next, the output is passed through a 1×1 convolution to produce two channels of 112×112 pixels. This is then processed through a fully connected layer with 4096 outputs, ReLU activation and dropout, a second identical fully connected layer and then a layer with 5 outputs and soft max activation that gives the classification scores across 5 different classes.

For the depth-2 model, the first convolution layer is identical to the previous model, and after the first pool, two more convolutions are added that produce 128 channels of size 112×112, reduced to 128 channels of dimension 56×56. The subsequent layers have identical structure as in the 1 layer model, where the input to the first fully connected layer has 8 channels of dimension 56×56. The models with 3, 4 and 5 layers are constructed using the same methodology. The architecture for depth-5 is shown in 2.

The networks have been trained and tested in three different modalities. The first one covers object classification, which includes people, ladders, windows, doors and a combination of windows and firefighters (5 classes). The second modality comprises a pose-based classification, which includes standing, sitting and crawling (3 classes). The third modality is binary and includes the presence or absence of fire. For the modalities with 5 and 3 classes, the training was performed using a categorical cross entropy loss, and for the modality with two classes, we used binary cross entropy loss combined with stochastic gradient descent. The learning rate for all trainings was 10^{-4} with a decay of 0.009. Early stopping with cross validation was applied to avoid overfitting.

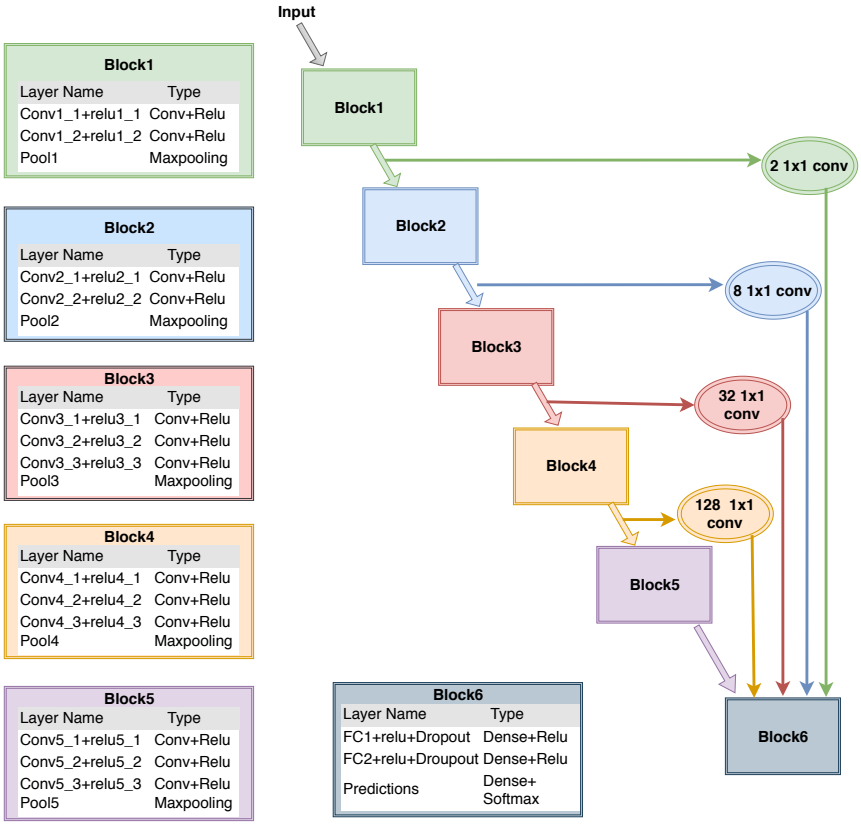

Figure 1: Block diagram of VGG16 at different depth levels.

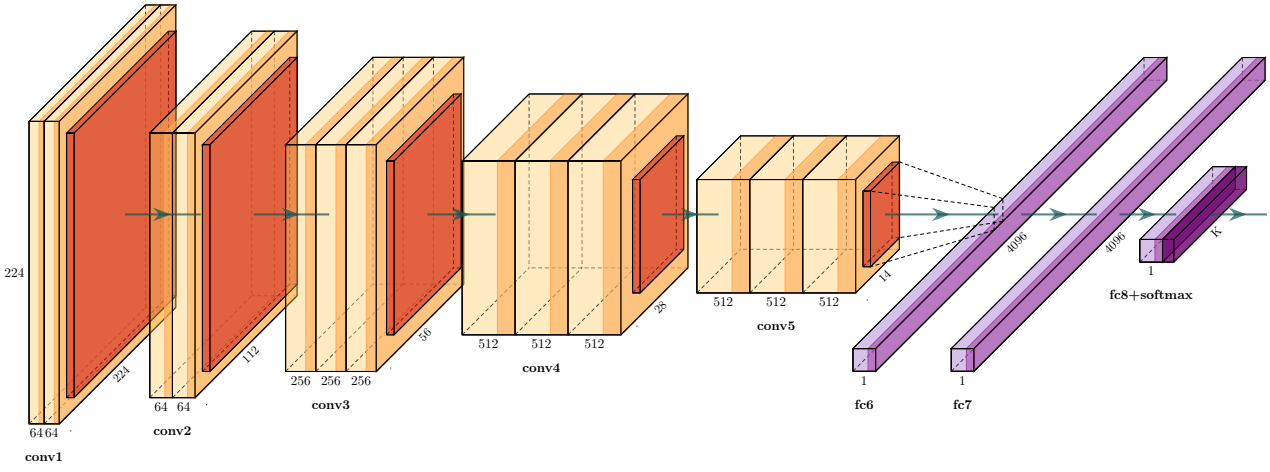

Figure 2: Architechture of depth-5 VGG16.

In order to avoid overfitting between training and test datasets, images taken from different recordings have been used in both processes. The results shown below contain extensive tests using a database of recordings taken by the researchers of this paper. It is described below.

3. Training and Test Datasets

The imagery data set used in this project was recorded at the Santa Fe Firefighting Facility, located in Santa Fe, New Mexico. Extensive video footage was acquired using an IR MSA 5200HD2TIC Camera. This camera is a multipurpose firefighting tool designed to aid search and rescue efforts in structural firefighting environments. It uses an uncooled microbolometer vanadium oxide(Vox) detector which com-

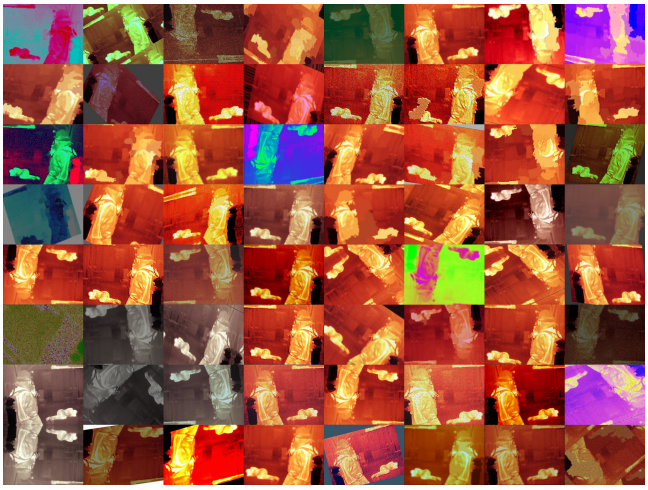

Figure 3: Demonstration of Image augmentation for training set.

prises of 320x240 FPA with the pitch of 38μm and spatial resolution of 7.5 to 13.5μm. This resolution is sufficient to capture necessary features for target detection. It records the image with a 320x240 focal plane array sensor and has the ability to record imagery in two different modes, i.e. low and high sensitivity. This device also features high score imagery, generating 76,000 pixels of image detail in both low and high sensitivity modes. Dense spectral resolution is (7.5 to 13.5μm). The output video is in NTSC format with a frame rate of 30 frames per second. The scene temperature has a maximum operating range of 560 degrees Celsius or 1040 degree Fahrenheit.

Over 6 hours of recorded video in both open and closed environments was acquired. The recording sessions produced more than 150 infrared video files, each one lasting approximately 2 to 3 minutes. All videos contained some combination of sequences involving the desired targets to be detected. In some scenarios, single objects of interest are present in a scene while in others multiple objects of interest are present simultaneously. This variation requires the CNN image classification system to be capable of multiple simultaneous object detections in the same frame. Other objects or postures of interest outside of what we chose can also be detected if those objects or poses occur in sufficient frequency to allow for training the data.

The objects of interest for this research are humans, doors, windows, ladders, and fire. The objective of the above described structure is to detect all objects of interest present in the scene at the same time. Since data sets of sufficient detail are needed to accurately train the neural network, the videos were used to extract a large quantity of images for training and test purposes. The training and test sequences were extracted from different videos in order to avoid overfitting. The video extracted from the camera was produced in grayscale 8 bit format. In order to generate the training data set, the images were pre-processed with data augmentation techniques such as skewing, translation, zooming, cropping and rotation as shown in fig 3 (False color for better visualization).

Object of Interest	number
door	322
firefighter and window	4663
firefighter	15484
ladder	1589
window	1620

Table 1
Data quantification for objects classification task(object set)

Object of Interest	number
fire	7603
No fire	7950

Table 2
Data quantification for fire classification task(fire set)

Object of Interest	number
Crawling	8678
sitting	1803
standing	9928

Table 3
Data quantification for poses classification task(poses set)

Tables 1, 2 and 3 show the total number of images acquired for training and test before the augmentation procedure. Objects of interest contained in images used in the training data set were hand labeled to assign them to a class. The labeled classes were then grouped into 3 sets (objects, fire, human poses). To compensate for the asymmetry of data within the different classes, data augmentation was performed on classes that had lower representation within the original dataset. For example, 7950 images from the original dataset were labeled and added to the "No fire" class within the set "Fire". The "fire class" was augmented to add 347 images to it to bring the total number of "fire" labeled images available for training up to 7950.

The primary training was performed using an Alienware Aurora R6 Desktop computer configured with 32GB RAM memory, and a Dual GTX 1080 with 16GB GPU memory. The cross validation and hyper parameter tuning portion of the research was performed on the high performance computer, Xena housed at the UNM Center for Advanced Research Computing(CARC). The machine has 24 single GPU nodes and 4 dual GPU nodes. These dual GPU nodes have 2 NVIDIA Tesla K40m GPUs with GPU memory of 11GB each and 64GB RAM memory per node. The computations for cross validation and hyper parameter tuning utilized the 4 dual GPU nodes.

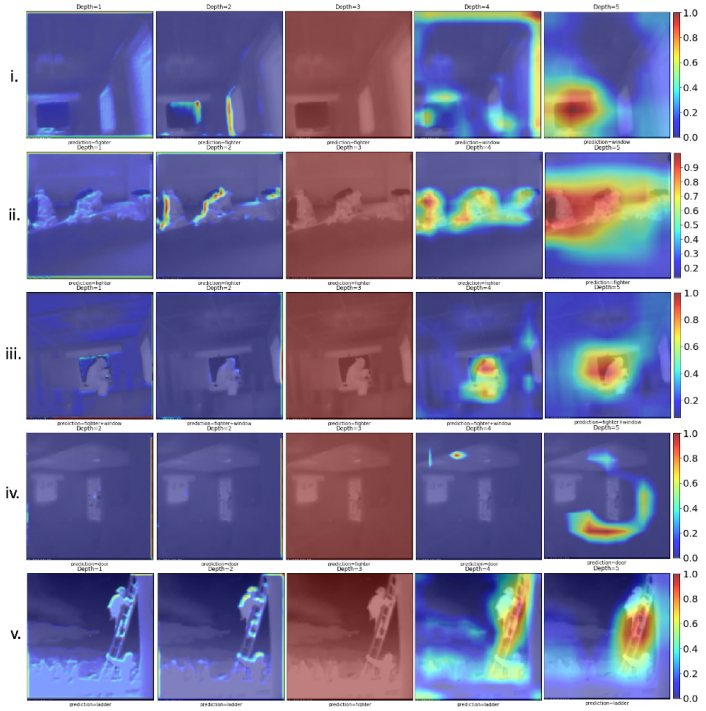

Figure 4: Visualization of the features obtained at the last Convolutional layer for Object Classification for different depths for respective classes **i.** Window, **ii.** Fighter, **iii.** Fighter+Window, **iv.** Door and **v.** Ladder

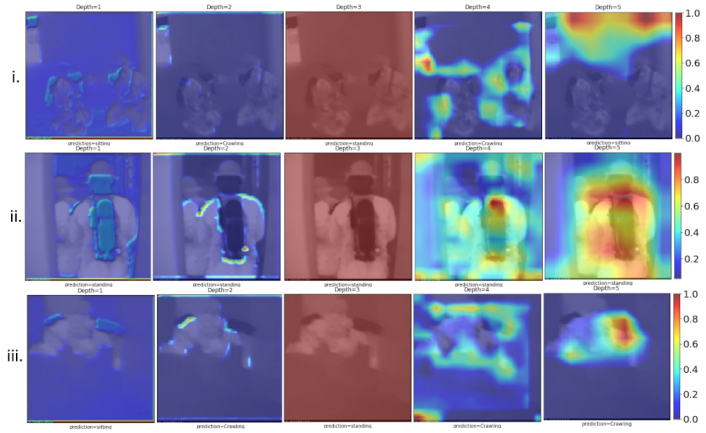

Figure 5: Visualization of the features obtained at the last Convolutional layer for pose Classification for different depths for respective classes **i.** Sitting, **ii.** Standing and **iii.** Crawling

The test data consists of 1/10 of the data set. The rest of the data has been used for training and validation purposes. A validation has been performed with a 9-stratified fold procedure with the remaining 9/10ths of data.

4. Results

Our network is trained to detect objects including ladders, doors, windows, people, and fire. The network is also trained to detect and classify different body positions of the detected people, civilian and firefighter alike. We classify three different poses corresponding to standing, sitting and crawling, which cover the variation found in the videos. Other important positions can be detected if they are sufficiently available in the image dataset used for training. The following results present the visualization of the features extracted by the convolutional section of the network, the F1 scores and achieved accuracy of the network, confusion matrices and ROC curves to estimate the false alarm versus the detection probability of the network. For this experiment, we change the detection probability by sweeping the detection threshold of the network.

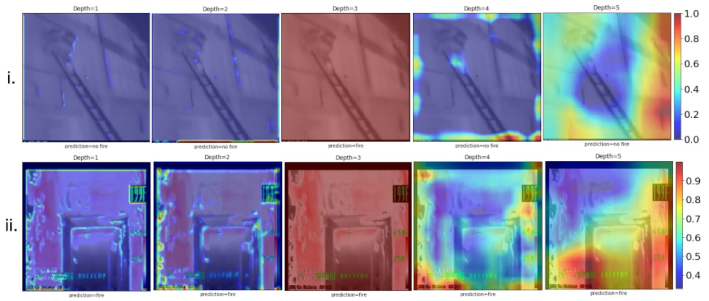

Figure 6: Visualization of the features obtained at the last Convolutional layer for fire Classification for different depths for respective classes **i.** fire and **ii.** No Fire

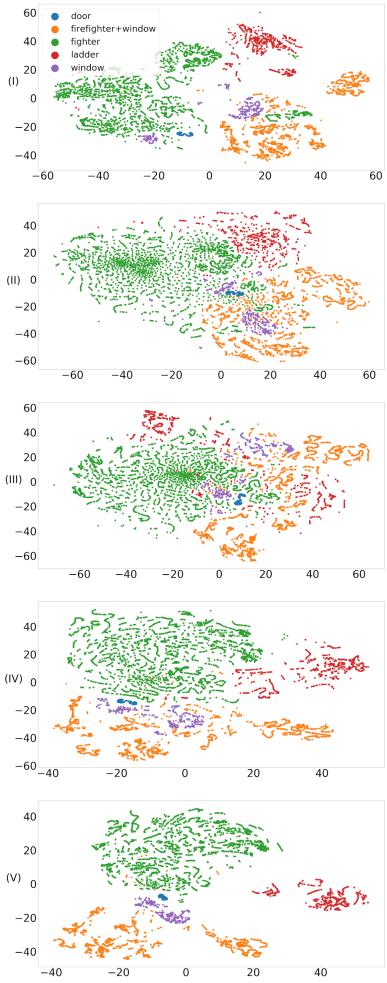

Figure 7: t-SNE of the CNN when trained to detect objects. The number of convolutional layers was respectively I) depth-1, II) depth-2, III) depth-3, IV) depth-4 and V) depth-5.

4.1. Visualization of the extracted features

All the configurations were tested against the available data. We have used grad-CAM(gradient weighted Class Activation Maps) [24] for visualizing attention over input. Grad-CAM uses the Convolutional layer before the dense layer in order to utilize spatial information for displaying the saliency maps corresponding to each predicted class. The activations correspond to the class with the highest detection score. We can see the classification scores for different classes corresponding to different images at different depths in table 4, 5 and 6. Based on the classification score of the table, we visualize the activations corresponding to classes with highest score. Figure 4,5 and 6 show grad-CAM for different depths for each classification task. Each of these figure comprises of subfigures corresponding to each classes depicting the grad-CAM for different depths from left to right. For most of these maps for depth-1 and depth-2 show that activations come from either edges or small regions from the objects of interest. With depths equal to 3, the network is unable to produce feature extraction, showing that this configurations is not useful for classification. Nevertheless, at higher levels of abstraction (depths equal to 4 and 5),features are extracted that produce exemplary results.

Figure 7 shows the visualization of the output of the neural networks using the t-distributed stochastic neighbor embedding technique presented in [18]. This technique allows the user to obtain a low dimensional representation of the data to better understand the data's distribution and separability. 7 (I) shows the distribution of the 5-dimensional output of the network when the depth is 1. In this configuration, the separability of the data is intuitively fair. 7 (II) and (III) show the distribution for depth 2 and 3. These representations show a high level of overlapping of the different classes, which is consistent with the poor extraction of features, as shown in figure4,5 and 6. The representation for depth of 4 and 5 as shown in 7 (IV) and (V) is highly improved, as the overlapping is dramatically decreased compared to the 2 and 3 depth networks.

The green dots correspond to images of firefighters. Orange dots are images containing firefighters and windows. These two classes comprise the majority of the data set but they show a low overlap between them indicating high accuracy in the classifier's ability to distinguish between firefighters in the presence or absence of a window. The blue dots correspond to images with doors. Door shapes vary due to the angle of the camera. However, the classifier is able to extract all door features regardless of these differences in

Table 4
Prediction score for object classification for different depths (green and yellow cells are the top 2 prediction scores)

Figure	Depth	Door	F/W	Ladder	Window	Fighter
4 i.	1	0.003	0.053	0	0.408	0.536
	2	0	0.43	0	0.01	0.56
	3	0	0	0	0	1
	4	0.002	0.037	0	0.961	0
	5	0.001	0.004	0	0.996	0
4 ii.	1	0.006	0	0.001	0.001	0.992
	2	0	0	0	0	1
	3	0	0	0	0	1
	4	0	0	0	0	1
	5	0	0	0.001	0	0.9999
4 iii.	1	0	0.997	0	0.003	0
	2	0	1	0	0	0
	3	0	0	0	0	1
	4	0	0.974	0	0.025	0
	5	0	0.999	0	0.001	0
4 iv.	1	0.988	0.001	0	0.001	0.01
	2	0.996	0	0	0.001	0.003
	3	0	0	0	0	1
	4	0.95	0.029	0	0.02	0.001
	5	0.805	0.01	0	0.181	0.004
4 v.	1	0.007	0.001	0.976	0.001	0.015
	2	0	0	0.999	0	0.001
	3	0	0	0	0	1
	4	0	0	0.998	0	0.002
	5	0	0	0.9	0	0.1

perspective-induced shape, and all features appear clustered in a small area. They are highly overlapped with the images containing firefighters at lower depths. The overlap decreases with higher level of abstraction (4 and 5 depths). The violet spots represent windows, which appear to be overlapping with the images of firefighters and windows. The accuracy and classification rates are in high agreement with this visualization.

4.2. Accuracy and precision

The accuracy of the neural network for different depths is consistent with the visualization of the features and the t-SNE. Figures 11 and 13 show the F1 scores and the precision of the network at depths 1 to 5. For the purposes of this research, the networks have been trained to detect objects pertinent to fire navigation and rescue including doors, people, ladders, windows and combination of firefighters and windows. The network with one layer(depth 1) has a reasonable accuracy close to 75% and shows a high variance in the results. The depth 1 framework requires a low computational burden in training and test. Computational burden increases with the growth in computational complexity. Each additional depth increases computational complexity due to the additional number of Floating Points(FLOPs) operations. Interestingly, the use of 2 and 3 depths produces an unacceptable performance, but dramatically improves when using 4 and 5 depths. The use of depth-5 is not necessary since its performance is almost identical to the run with 4 layers but the computational burden is significantly higher.

Table 5
Prediction score for Pose classification for different depths

Figure	Depth	Crawling	Standing	Sitting
5 i.	1	0.33	0	0.67
	2	0.83	0.001	0.169
	3	0	1	0
	4	0.63	0	0.37
	5	0.20	0	0.80
5 ii.	1	0	1	0
	2	0.01	0.99	0.001
	3	0	1	0
	4	0	1	0
	5	0	1	0
5 iii.	1	0.33	0.02	0.65
	2	0.536	0.311	0.153
	3	0	1	0
	4	0.62	0.097	0.283
	5	.8	0	0.2

Figure 14 shows the accuracy of the network when detecting fire. In this case it is only necessary to run the network with one layer. As in the previous experiments, the network with three layers produces a poor performance, while the other depth tests show an accuracy that is almost iden-

Table 6
Prediction score for Fire classification for different depths

Figure	Depth	Fire	No Fire
6 i.	1	0	1
	2	0.001	0.999
	3	1	0
	4	0	1
	5	0	1
6 ii.	1	1	0
	2	1	0
	3	1	0
	4	1	0
	5	0.999	0.001

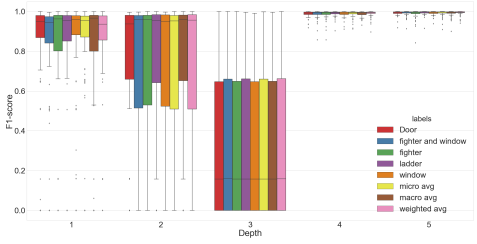

Figure 8: F1 scores for the classification of objects with CNNs of different depths.

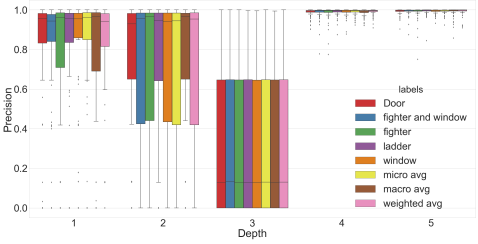

Figure 9: Precision curves for object detection with CNNs of different depths.

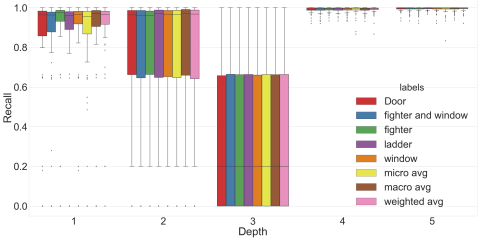

Figure 10: Recall curves for object detection with CNNs of different depths.

tical. The average test accuracy for all objects is depicted in figure 15. From this graph, we can conclude that a good trade off between computational burden and accuracy is obtained with a network of only one convolutional layer, while the highest level of accuracy is obtained using a network with 4 layers, which can achieve an accuracy of more than 97%.

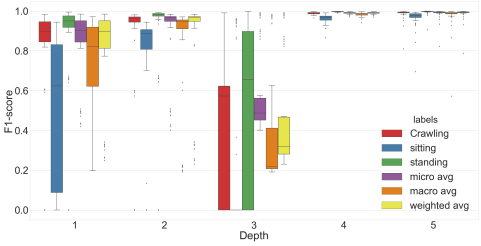

Figure 11: F1 scores for the classification of poses with CNNs of different depths.

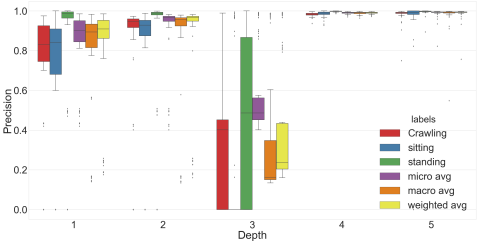

Figure 12: Precision curves for poses detection with CNNs of different depths.

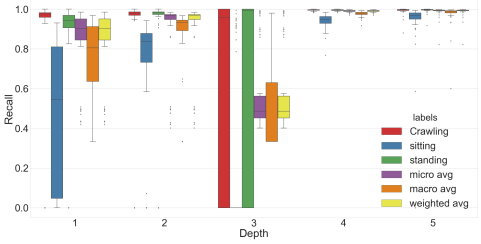

Figure 13: Recall curves for poses detection with CNNs of different depths.

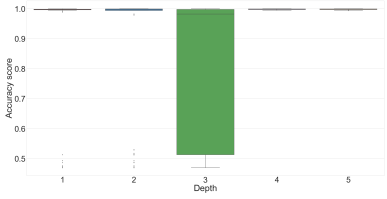

Figure 14: Test accuracy in fire detection with CNNs of different depths.

The network's ability to accurately distinguish between differing human positions and classify them accordingly presents new and interesting opportunities. With pose recognition, body position can be used to assist in making health inferences. For example, the presence of a person laying down is very important in a fire scenario, as it may be a significant indicator of a person who has succumbed to smoke inhalation and is in desperate need of rescue. The results can aid in alerting rescue teams to the possible health condition of

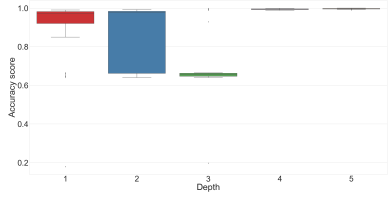

Figure 15: Test accuracy in object detection with CNNs of different depths.

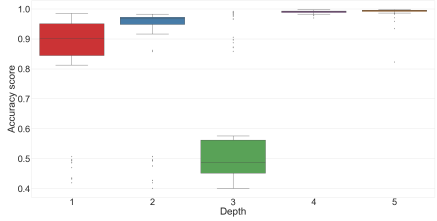

Figure 16: Test accuracy in pose detection with CNNs of different depths.

victims and prioritize evacuation. In our tests, we trained the network to detect persons standing, sitting or crawling from labelled images. The network shows a similar performance with accuracy of 95% on average.

4.3. Confusion matrices

The confusion matrices for all detection modalities have been computed. Tables 7 to 10 shows the confusion matrices for the task of object and human detection. As stated before, the network with just one depth of convolutional blocks performs reasonably well without excess computational burden. The detection probabilities are around 75%. We note however that at lower depths(i.e 1 and 2) the network is challenged to make clear distinctions between classes that tend to occur together. For example, both windows and ladders tend to occur in a scene with one or more firefighters. The network has a high confusion rate in determining windows and ladders as individual classes, as there is a confusion of about 19% with firefighters in the classification of these elements. This is likely due to a significantly higher number of unaugmented images in the firefighter class when compared to the others in the "object of interest" set. When the number of depths is increased to 2, the confusion is even worse, reaching 35%. However, accuracy of detection of firefighters as a class is between 97 and 99%. If the number of layers is increased to 4, the confusion matrices show a confusion between 0 and 1.7% for all objects being classified. As stated above, there are no significant differences between 4 and 5 depths. Note that the classification performance with these number of depths ranges between 97 and 99.8%.

A similar trend in detection accuracy related to the number of depths can be seen in the detection of poses. Results at depths 1 and 2 perform poorly compared to 4 and 5 as shown in Tables 11, 12, 13 and 14 . At depths 1 and 2, the network was mainly challenged in distinguishing the differences between sitting and crawling. This is possibly due to the fact that the relative positions of the firefighters in these two poses are similar but simply rotated. The accuracy increases significantly with 4 depths (Table 13), where the confusion decreases to a range between 0 and 0.4 in all cases except for the detection of the sitting pose, which stands at a confusion rate of 5.7 (in the network with 4 depths) and 5.1% with the crawling pose with 5 depths (Table 14). Again we found no significant differences between the results of these two network configurations.

Table 7
Confusion matrix for object detection, depth=1

Door	73.2	1.1	23.5	0.0	2.2
F/W	0.0	77.7	19.0	0.0	3.3
Fighter	0.3	1.8	97.0	0.1	0.8
Ladder	0.0	1.2	19.1	79.6	0.1
Window	0.3	4.3	19.9	0.0	75.5
	Door	F/W	Fighter	Ladder	Window

Table 8
Confusion matrix for object detection, depth=2

Door	61.3	0.0	38.2	0.0	0.5
F/W	0.0	62.4	36.0	0.0	1.7
Fighter	0.1	0.2	99.4	0.1	0.2
Ladder	0.0	0.0	34.3	65.7	0
Window	0.1	1.0	35.9	0	63.0
	Door	F/W	Fighter	Ladder	Window

Table 9
Confusion matrix for object detection, depth=4

Door	97.1	0.2	0.9	0.0	1.7
F/W	0.0	98.3	1.0	0.0	0.7
Fighter	0.1	0.1	99.8	0.0	0.0
Ladder	0.0	0.0	0.2	99.8	0.0
Window	0.3	0.5	0.3	0.0	98.9
	Door	F/W	Fighter	Ladder	Window

Table 10
Confusion matrix for object detection, depth=5

Door	98.0	0.0	0.7	0.0	1.3
F/W	0.0	99.2	0.5	0.0	0.3
Fighter	0.0	0.1	99.9	0.0	0.0
Ladder	0.0	0.0	0.1	99.9	0.0
Window	0.3	0.6	0.2	0.0	98.9
	Door	F/W	Fighter	Ladder	Window

4.4. Detection versus false alarm probabilities

Figures 17, 18 and 19 show the probability of detection versus false alarm for the classification of objects and humans, poses and fire. The graphs have been obtained by sweeping the detection threshold from 0 to 1. In all cases, it is observed that the probability of false alarm is negligible when the networks have a depth of 4 or 5. As expected given the poor performance of the classifier at depths 2 and 3, the results showed high false alarm probabilities. In the case of

Table 11
Confusion matrix for pose detection, depth=1

Crawling	85.6	1.2	13.2
Sitting	41.4	45.9	12.7
Standing	10.8	0.4	88.8
	Crawling	Sitting	Standing

Table 12
Confusion matrix for pose detection, depth=2

Crawling	86.3	0.8	12.9
Sitting	19.3	67.8	12.9
Standing	7.5	0.2	92.3
	Crawling	Sitting	Standing

Table 13
Confusion matrix for pose detection, depth=4

Crawling	99.5	0.2	0.3
Sitting	5.7	94.1	0.2
Standing	0.5	0.0	99.5
	Crawling	Sitting	Standing

Table 14
Confusion matrix for pose detection, depth=5

Crawling	99.4	0.2	0.4
Sitting	5.1	94.7	0.2
Standing	0.5	0.0	99.5
	Crawling	Sitting	Standing

the depth 1 network, the results show flaws in the probability of detection versus probability of false alarm. A detection rate of about 82% is achieved with a false alarm rate of about 18% (17)for object detection as per the micro/macro average ROC area. The optimal false alarm rates in individual class object detection range from 15-20%, in particular, in the detection of humans. Similar results are obtained in pose detection. Therefore, although depth 1 networks can be used with lower computational burden, they should only be used in cases where such false alarm rates can be tolerated. False alarm probability can be decreased if the detection is performed on a sequence of images chronologically, and then a voting procedure is applied to all detections. This would be done at the cost of additional computational time in the object or pose detection procedures.

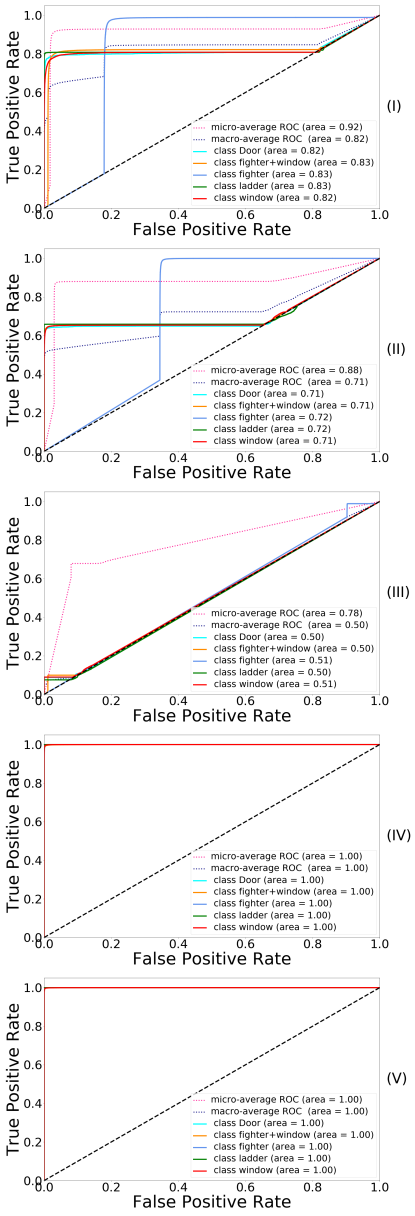

Figure 17: ROC curve for object detection for I) Depth-1, II) Depth-2, III) Depth-3, IV) Depth-4 and V) Depth-5 architectures.

5. Conclusion

The scene of a structural fire is chaotic, dangerous, and disorienting. Heavy smoke, near zero visibility, extreme heat and active flames create a perfect storm for stress induced misjudgments that can affect even seasoned fire fighters. These are prime conditions for computer aided assistance and artificially intelligent solutions. Our research provides a mechanism that can supplement firefighters with real time information and offer guidance by automatically interpreting the fireground from the information provided by the hand held thermal cameras already in use by firefighters.

We present a deep learning based technology that is capable of accurate automated detection of objects of interest utilizing thermal imagery being recorded on the scene. Convolutional neural networks have demonstrated outstanding performance in object detection on RGB imagery. However, the zero light, heavy smoke conditions typical of structural fires render RGB cameras useless. The CNN developed for this application is unique in that it achieves high detection accuracy applied to thermal imagery and is capable of processing, analysis and result generation in real time if a camera is connected to a simple single board commercial computer endowed with GPU capabilities (for example, an NVIDIA

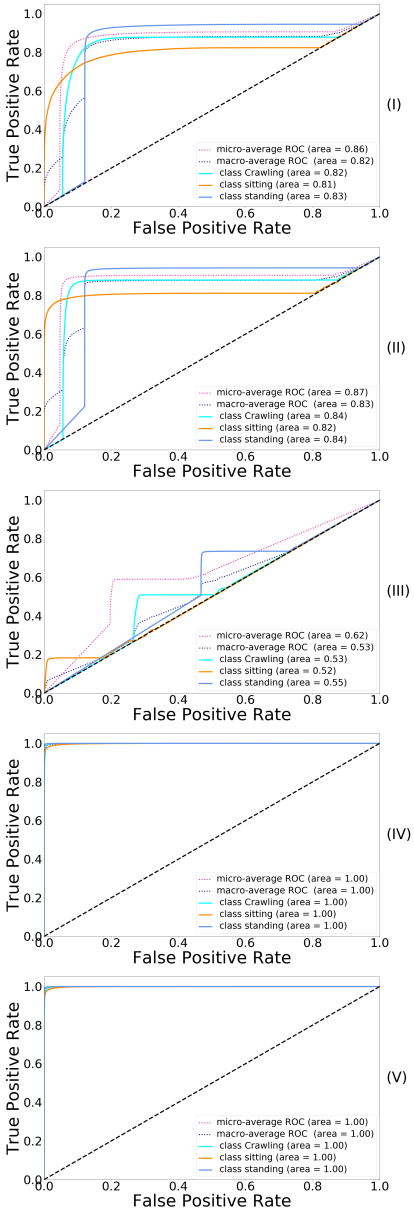

Figure 18: ROC curve for pose detection for I) Depth-1, II) Depth-2, III) Depth-3, IV) Depth-4 and V) Depth-5 architectures .

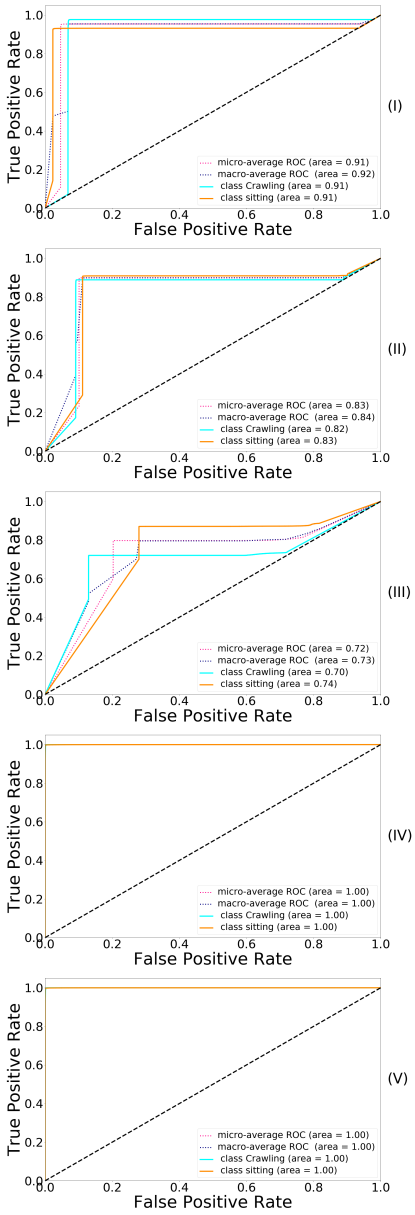

Figure 19: ROC curve for fire detection for I) Depth-1, II) Depth-2, III) Depth-3, IV) Depth-4 and V) Depth-5 architectures .

JETSON). Our model is able to accurately detect objects of interest such as doors, windows, and ladders vital to evacuation. It is also able to detect people and differentiate postures. Posture detection may assist in prioritization of rescue as a rough estimate of the health status of a victim (for instance a prone posture may indicate a person who has succumbed to the effects of smoke inhalation and require immediate evacuation and paramedic assistance). Our framework is capable of performing with greater than 95% accuracy on the detection of these objects of interest and more then 90% accuracy on posture identification. We also present an evaluation of computational time to accuracy achievement trade-off and show that the model performs above 70% even at the lowest convolutional depth, making it highly adept to usage in the field where computational power may be a concern.

This work lays a foundation to the development of a real time situational map of the structural configuration of a building that is actively built and updated via the live thermal imagery being recorded by firefighters moving through the scene. This map, which is updated in real time, could be used by firefighters to assist them in safely navigating the burning structure and improve the situational awareness necessary in decision making by tracking exits that may become blocked and finding alternatives. Utilizing the fea-

tures detected via the approach presented in this paper, a robust localization and tracking system to track objects of interest in sequences of frames can be built. The visual features from this framework have also be coupled with a Natural Language Processing(NLP) system for scene description and allow the framework to autonomously make human understandable descriptions of the environment to aid firefighters to improve their understanding of the immediate surroundings and assist them when anxiety levels are heightened. Our future work seeks to join these two components with a reinforcement learning(Q-learning) algorithm that utilizes the continuously updated state map and reinforcement learning techniques that assist in path planning and can be vocalized through the NLP system. The deep Q-learning based approach provides a navigation system that actively avoids hazardous paths. These three components, built on the backbone of the research presented here, can be fused to accomplish the ultimate goal of providing an artificially intelligent solution capable of guiding firefighters to safety in worst case scenarios.

References

[1] Ajith, M., Martínez-Ramón, M., 2019. Unsupervised segmentation of fire and smoke from infra-red videos. IEEE Access Submitted.

[2] Celik, T., Demirel, H., Ozkaramanli, H., Uyguroglu, M., 2007. Fire detection using statistical color model in video sequences. J. Vis. Comun. Image Represent . 18, 176–185.

[3] Chacon-Murguia, M.I., Perez-Vargas, F.J., 2011. Thermal video analysis for fire detection using shape regularity and intensity saturation features, in: Mexican Conference on Pattern Recognition, Springer. pp. 118–126.

[4] Chevalier, M., Thome, N., Cord, M., Fournier, J., Henaff, G., Dusch, E., 2016. Low resolution convolutional neural network for automatic target recognition, in: 7th International Symposium on Optronics in Defence and Security.

[5] Cho, Y., Bianchi-Berthouze, N., Marquardt, N., Julier, S.J., 2018. Deep thermal imaging: Proximate material type recognition in the wild through deep learning of spatial surface temperature patterns, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, ACM. pp. 2:1–2:13.

[6] Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01, IEEE Computer Society, Washington, DC, USA. pp. 886–893.

[7] De Oliveira, D.C., Wehrmeister, M.A., 2018. Using deep learning and low-cost RGB and thermal cameras to detect pedestrians in aerial images captured by multirotor UAV. Sensors 18.

[8] Ding, Z., Nasrabadi, N., Fu, Y., 2016. Deep transfer learning for automatic target classification: MWIR to LWIR, in: Automatic Target Recognition XXVI, International Society for Optics and Photonics. p. 984408.

[9] Elangovan, K., Seenivasan, V., Anand, K., raj, C., 2014. Human detection in hours of darkness using Gaussian mixture model algorithm. International Journal of Information Sciences and Techniques 4, 83–89. doi:10.5121/ijist.2014.4311.

[10] Ge, J., Luo, Y., Tei, G., 2009a. Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. IEEE Transactions on Intelligent Transportation Systems 10, 283–298.

[11] Ge, J., Luo, Y., Tei, G., 2009b. Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. IEEE Transactions on Intelligent Transportation Systems 10, 283–298.

[12] Jackson, M., Robins, I., 1994. Gas sensing for fire detection: Measurements of CO, CO2, H2, O2, and smoke density in European standard fire tests. Fire Safety Journal 22, 181 – 205.

[13] Kim, J.H., Jo, S., Lattimer, B.Y., 2016a. Feature selection for intelligent firefighting robot classification of fire, smoke, and thermal reflections using thermal infrared images. Journal of Sensors 2016.

[14] Kim, J.H., Jo, S., Lattimer, B.Y., 2016b. Feature selection for intelligent firefighting robot classification of fire, smoke, and thermal reflections using thermal infrared images. Journal of Sensors 2016.

[15] Kok, V.J., Lim, M.K., Chan, C.S., 2016. Crowd behavior analysis: A review where physics meets biology. Neurocomputing 177, 342–362.

[16] Lee, E.J., Ko, B.C., Nam, J.Y., 2016. Recognizing pedestrianâĂŹs unsafe behaviors in far-infrared imagery at night. Infrared Physics & Technology 76, 261–270.

[17] Luo, R.C., Su, K.L., 2007. Autonomous fire-detection system using adaptive sensory fusion for intelligent security robot. IEEE/ASME Transactions on Mechatronics 12, 274–281.

[18] Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-SNE. Journal of machine learning research 9, 2579–2605.

[19] Marbach, G., Loepfe, M., Brupbacher, T., 2006a. An image processing technique for fire detection in video images. Fire safety journal 41, 285–289.

[20] Marbach, G., Loepfe, M., Brupbacher, T., 2006b. An image processing technique for fire detection in video images. Fire safety journal 41, 285–289.

[21] Merino, L., Caballero, F., Martínez-De-Dios, J.R., Maza, I., Ollero, A., 2012. An unmanned aircraft system for automatic forest fire monitoring and measurement. Journal of Intelligent & Robotic Systems 65, 533–548.

[22] Pawlowski, P., Piniarski, K., Dabrowski, A., 2015. Pedestrian detection in low resolution night vision images, in: 2015 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), IEEE. pp. 185–190.

[23] Rodin, C.D., de Lima, L.N., de Alcantara Andrade, F.A., Haddad, D.B., Johansen1, T.A., Storvold, R., 2018. Object classification in thermal images using convolutional neural networks for search and rescue missions with unmanned aerial systems, in: 2018 International Joint Conference on Neural Networks (IJCNN).

[24] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, pp. 618–626.

[25] Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

[26] Stone, K., Keller, J., 2014. Convolutional neural network approach for buried target recognition in FL-LWIR imagery, in: Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XIX, International Society for Optics and Photonics. p. 907219.

[27] TÃűreyin, B.U., Cinbis, R.G., Dedeoglu, Y., Cetin, A.E., 2007. Fire detection in infrared video using wavelet analysis. Optical Engineering 46, 1 – 9.

[28] Valldor, E., 2014. Person Detection in Thermal Images using Deep Learning. Ph.D. thesis. University of Uppsala.

[29] Vandecasteele, F., Merci, B., Jalalvand, A., Verstockt, S., 2017. Object localization in handheld thermal images for fireground understanding, in: Proc. SPIE 10214, Thermosense: Thermal Infrared Applications XXXIX, 1021405 (5 May 2017).

[30] Wang, H., Cai, Y., Chen, X., Chen, L., 2016. Night-time vehicle sensing in far infrared image with deep learning. Journal of Sensors 2016.

[31] Wang, Y., Chua, T.W., Chang, R., Pham, N.T., 2012. Real-time smoke detection using texture and color features, in: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), IEEE. pp. 1727–1730.

[32] Xu, F., Liu, X., Fujimura, K., 2005. Pedestrian detection and tracking with night vision. IEEE Transactions on Intelligent Transportation Systems 6, 63–71.

[33] Yun, K., Huyen, A., Lu, T., 2018. Deep neural networks for pattern

recognition. CoRR abs/1809.09645. URL: http://arxiv.org/abs/1809.09645.

[34] Zhang, H., Luo, C., Wang, Q., Kitchin, M., Parmley, A., Monge-Alvarez, J., Casaseca-de-la Higuera, P., 2018a. A novel infrared video surveillance system using deep learning based techniques. Multimedia Tools and Applications 77, 26657–26676.

[35] Zhang, X., Yu, Q., Yu, H., 2018b. Physics inspired methods for crowd video surveillance and analysis: a survey. IEEE Access 6, 66816–66830.

[36] Zhao, Y., Ma, J., Li, X., Zhang, J., 2018. Saliency detection and deep learning-based wildfire identification in UAV imagery. Sensors 18.

Manish Bhattarai received the Masters degree in Electrical Engineering with specialization in signal processing from the University of New Mexico in 2017.He is currently working towards his Ph.D. degree in Electrical Engineering from the same school as well as employed as a full time research assistant at the Los Alamos National Laboratory. His research interests are Machine Learning, deep learning ,Computer Vision, AI and HPC.

Manel Martínez-Ramón is a professor with the ECE department of The University of New Mexico. He holds the King Felipe VI Endowed Chair of the University of New Mexico, a chair sponsored by the Household of the King of Spain. He is a Telecommunications Engineer (Universitat Politecnica de Catalunya, Spain, 1996) and PhD in Communications Technologies (Universidad Carlos III de Madrid, Spain, 1999). His research interests are in Machine Learning applications to smart antennas, neuroimage, first responders and other cyber-human systems, smart grid and others. His last work is the monographic book "Signal Processing with Kernel Methods", Wiley, 2018.