applied sciences

MDPI

Article

A Hybrid Retina Net Classifier for Thermal Imaging

Ventrapragada Teju [1], Kambhampati Venkata Sowmya [2], Srinivasa Rao Kandula [1], Anca Stan [3] and Ovidiu Petru Stan [4,*]

[1] Department of Electronics & Communications Engineering, Dhanekula Institute of Engineering and Technology, Ganguru, Vijayawada 521139, India; ventrapragadateju@gmail.com (V.T.); ksrinivas.ece@gmail.com (S.R.K.)
[2] Department of Electronics & Communications Engineering, Koneru Lakshmaiah Education Foundation, Vijayawada 522302, India; sowmyakambhampati@gmail.com
[3] Faculty of Industrial Engineering, Robotics and Production Management, Technical University of Cluj Napoca, 400114 Cluj-Napoca, Romania; anca.stan@muri.utcluj.ro
[4] Faculty of Automation and Computer Science, Technical University of Cluj Napoca, 400114 Cluj-Napoca, Romania
* Correspondence: ovidiu.stan@aut.utcluj.ro

Abstract: Thermal imaging is a cutting-edge technology which has the capability to detect objects in any environmental conditions, such as smoke, fog, smog, etc. This technology finds its importance mainly during nighttime since it does not require light to detect the objects. Applications of this technology span into various sectors, most importantly in border security to detect any incoming hazards. Object detection and classification are generally difficult with thermal imaging. In this paper, a one-stage deep convolution network-based object detection and classification called retina net is introduced. Existing surveys are based on object detection using infrared information obtained from the objects. This research is focused on detecting and identifying objects from thermal images and surveillance data.

Keywords: thermal imaging; CNN; retina net; classification; neural network

Citation: Teju, V.; Sowmya, K.V.; Kandula, S.R.; Stan, A.; Stan, O.P. A Hybrid Retina Net Classifier for Thermal Imaging. *Appl. Sci.* **2023**, *13*, 8525. https://doi.org/10.3390/app13148525

Academic Editors: Milos Manic, Sergiu Dan Stan and Milos Simonovic

Received: 10 July 2023
Revised: 19 July 2023
Accepted: 19 July 2023
Published: 24 July 2023

1. Introduction

Thermal imaging, a technology which enables objects to be identified even in the dark, is a leading-edge technology of great importance. As this technology is mostly used at nighttime in order to increase visibility, thermal images are fused with color images. A technology that augments the features of thermal images with saliency maps is proposed in [1] where the maps are generated using PICA-Net and R3-Net. This is used especially in pedestrian detection during the daytime. To achieve this objective a region-based convolution neural network (CNN) is trained. A CNN-based detection algorithm for pedestrians is proposed, which is a very useful method to detect human movements in video surveillance. The pedestrian detection algorithm, YOLO (You Only Look Once) detector is combined with AMBS saliency feature map in this process [2]. Another method of training the YOLO convolution network with one dataset among the available ones is presented in [3]. This network is trained with a COCO dataset which has outperformed the existing methods.

Comparing the parameters of the proposed model with existing models helps researchers improve their direction of research in a significant way. One such comparison is made in [4], where a model that is custom trained with some images taken from a video is compared with the existing YOLOv3. One can have a comparison of two models here. Counting the number of persons in each frame rather than only detecting the person becomes important in some applications. Thus, algorithms must be developed in such a way that they count the number of people. Gomez et al. proposed such an algorithm based on convolution neural networks [5].

Earlier, the creation of systems related to image processing was particularly restricted to expansion related to interfacing of a user in whichever programs most firms are engaged. These circumstances have been crucially changed with the arrival of Windows OS when most developers converted to resolving issues related to image processing. However, it did not lead to the main improvement in resolving functions in recognizing faces, numbers of cars, road signals, the investigation of images related to remote and medical, etc. All these endless issues are resolved by the trial-and-error method with the help of many engineers and scientists and is very time-consuming. Present-day technical solutions are very costly; thus, a function related to converting designs is based on the tools in software to solve issues of intelligence and prepare them methodically. In the image-processing domain, the necessary kit of tools would support analysis, recognizing an image and its previous information, which is not known all along. The improvement and implementation made by a normal programmer would be significantly improved. It is just like how the Windows toolkit provides support for creating interfaces to solve many issues.

Recognizing an object is to narrate a group of computer vision assignments, which involves actions, such as identification of objects in digital photos. Classification of an image deals with actions, such as prediction of the class of a single object in a particular image. Localization of an object means identifying the location of single or multiple objects in a particular image and drawing an abounding box on all sides. Detecting an object will combine these two functions and will localize and classify a single or multiple objects in a particular image. Generally, "object recognition" means "object detection". It is a challenge for beginners to differentiate among various computer-related vision tasks.

Classification of an image also deals with allocating a label known as class to a particular image, but localization of an object is made by drawing a bounding box on every side of one or more objects in a particular image. Detection of an object is an exciting task which fuses both assignments and will draw a bounding box on all sides of an object in an image and will assign a class label. Finally, all these issues are known as object recognition. Object recognition means a group of connected tasks used to identify an object in a digital input frame. Hybrid classifier is a family of methods that address localization of an object and recognizing tasks. It was invented for the performance of a model. OFSA-OKF [6,7] is implemented to recognize an object which is outlined for speedy and real-time applications.

2. Literature Survey

Thermal imaging finds an extensive application during nighttime, primarily proving its usefulness in securing national borders. Because of the lack of visible light, certain object features become challenging to track. As a solution, a CNN trained on visible images is adapted for thermal image tracking. An ensemble tracker based on correlation filter is proposed in [6], which has the capability of convolutional features extended for multiple layers rather than a single layer.

Long wave infrared sensor (LWIR), as the name itself implies, could sense objects within the range of the invisible IR spectrum. Target detection and false alarm rates are improved using LWIR sensors in the work presented in [7]. The effectiveness of combining a CNN with background modeling for human detection is proposed and demonstrated by Shahid et al. [8]. Improved Gaussian average and human classification using CNN is only performed for foreground objects in real time. The technique to detect in real time humans among images, which are thermal and built by background modeling, CNN, is explained in [8]. Object detection using retina net gives prominently better results compared with the other existing methods. The temporal characteristics of an image add an advantage in extracting the features of a particular image more accurately [9]. When a sequence of images is used, rather than a single image, 21.4% improvement in performance is achieved. When a long-range image is to be detected, the number of pixels on the target image obviously becomes less as the distance increases. Zhang et al. presented a resolution method that was proposed as a solution to address this issue. The method not only increases the resolution of the image but also enhances the baseline quality of inputs for object recognition. The

system was tested using two datasets that included pedestrians and six distinct types of vehicles [10].

A research study [11] that focuses on techniques and systems implemented on Raspberry Pi covers various aspects, including machine learning models, feature extraction techniques, and datasets. The recognition and detection of humans using thermal images, videos, and the utilization of different wavelengths are expanding, creating a demand for research in the field of machine vision, deep learning, and domains, such as infrared. Researchers in thermal imaging commonly show interest in detecting humans and exploring techniques that combine thermal imaging with images captured at different wavelengths [12]. The writers will estimate pixel-level images combining infrared and RGB images to refine the detectors of pedestrians built on CNN, which will work during the day or night and that are useful for video surveillance, autonomous vehicles, and advanced driver-assistance systems (ADAS). In [13], the authors reported a model based on a nine-layer CNN called self-learning soft max which utilizes the near-infrared images to identify the pedestrians. Many samples were collected for testing the CNN-based model and found that this model provides better results in terms of pedestrian recognition and accuracy. Imran et al. [14] introduced a narrative descriptor of saliency awareness known as SSDI, which means stacked saliency difference image, to design spatial–temporal, local, and global movement data to human action recognition (HAR) among infrared images of infrared. Here they used a four-stream deep framework built on CNN and RNN, known as RNN designs, and obtained a result of 83.5% on the dataset of InfAR [15] and a baseline result of 75.17% with the suggested dataset IITR-IAR [14]. The use of CNN to detect humans, their recognition, and classification of action are discussed in [16–20].

The Caltech pedestrian dataset [21,22] introduced as a benchmark for pedestrian detection, surpasses existing datasets in terms of scale and includes various videos, such as low-resolution images and videos captured from moving vehicles. The availability of this dataset opens new research possibilities. Alternatively, Viola et al. [23] introduced a novel framework for object detection that significantly contributes to three key areas: image representation, learning algorithm, and cascading classifiers. Their approach yields superior results compared to other object detection methods. A state-of-the-art classification and detection model based on deep neural networks was proposed by Szegedy [24] which utilizes the computing resources to the core of an inside network, producing better results. In [25,26], the authors introduced the concept of pedestrian detection using deep convolution neural networks, while Teju [6] proposed object detection using OFSA and even object tracking using an optimal Kalman filter [7].

3. Materials and Methods

Classification of an object is a major problem in computer vision [27]. Design is tasked by localizing objects in a thermal image and at the same time dividing them into various groups. Here, we will implement a hybrid retina net, which is a popular detector that is accurate and runs fast. Retina net utilizes a feature called pyramid network to effectively detect objects at various scales and introduces a new loss called focal loss function to reduce the issue of utmost foreground–background class imbalance [28].

Generally, computer vision is a transdisciplinary domain of ML and AI and is concerned with automatic extracting, analyzation, and understanding the important data from an image. With the rapid advancement of technology, there has been a significant increase in the amount of digital information associated with videos and images. In the machine vision domain, precepting and analyzation of images which are thermal poses a crucial challenge for computers compared to humans. Hence, classification of images is done using human intervention. Thermal-imaging information is done solely for the purpose of training and testing. The next images are divided with the help of patterns obtained from earlier stages. The obtained outcomes vary with the found patterns and depend fully on the understanding of the person who will do the classification. The proposed architecture

of deep learning that performs images classification will use multiple layers in a hybrid neural network to obtain the latest features from dataset images.

Figure 1 depicts the proposed structure for the classification of an object. This structure utilizes a grayscale image as an input image with a size of 28 × 28. Layer 1 of the CNN applies 32 filters upon input images in which every size of the image is 3 × 3, generating 32 feature maps which are 26 × 26 in size. Layer 2 applies 64 filters, which are 3 × 3 in size, generating 64 feature maps, which are 24 × 24 in size. Layer 3 is the max pooling layer and is utilized to down sample an image up to 12 × 12 with assistance of the subsampling window, which has the size of 2 × 2. The fourth layer consists of 128 fully connected neurons and utilizes the sigmoid activation function for image classification, producing an image as its output.

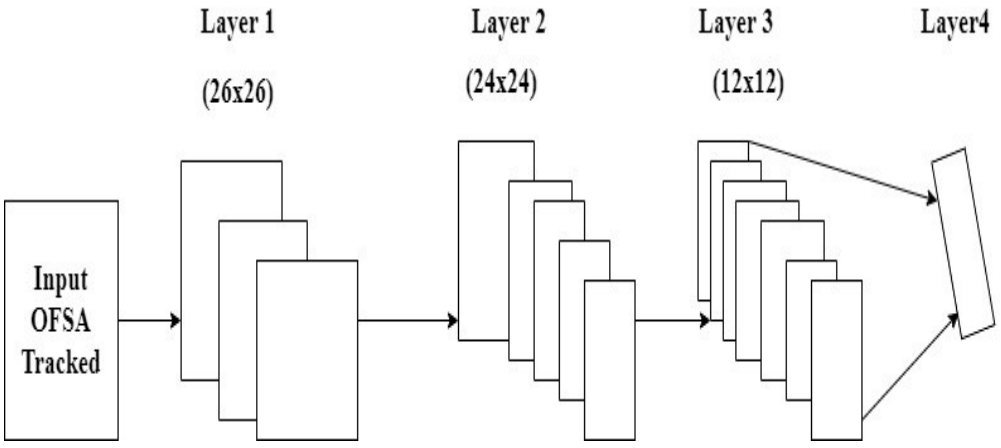


Figure 1. Proposed architecture for object classification.

Figure 2 illustrates the architecture of the proposed method. In a feed-forward neural network, each hidden layer is composed of neurons that are connected to the previous layer. The final network layer is fully connected and is used to perform image classification. Typically, input image has a size of 28 × 28 × 1 (28 pixels wide, 28 pixels high, and 1 color channel). Consequently, the first hidden layer of the network consists of 784 weights (28 × 28 × 1) corresponding to each input pixel.

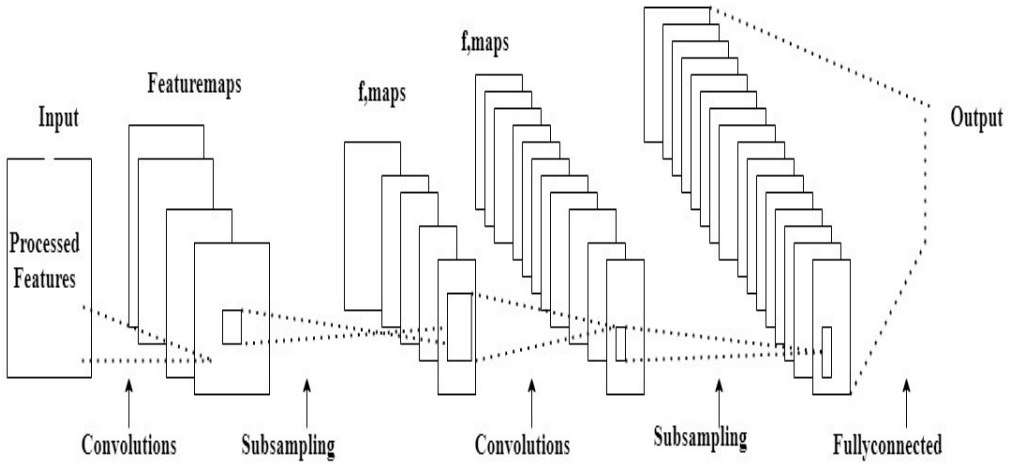


Figure 2. Description of proposed architecture.

Managing a large number of weights becomes challenging as the size of the input images increases. For example, images with dimensions of 400 × 400 × 3 would require 480,000 weights (400 × 400 × 3) to fully connect a layer. In such cases, a fully connected layer may not scale well due to the massive number of weights involved. Alternative

techniques, such as CNNs, are often employed for handling larger image sizes [29]. CNNs utilize weight-sharing and local receptive fields, which significantly reduce the number of parameters compared to fully connected layers, making them more suitable for processing larger images. The architecture of CNN has a different planning compared to the normal neural networks. One of the benefits is it can have as input different sizes [30]. Convnet layers have neurons which have 3 measurements, such as width, height, and depth. Here, depth indicates the third measurement of activation volume and not the depth of a complete neural network, which indicates entire network layers. The input image is $32 \times 32 \times 3$ in size and has volume of dimensions of $32 \times 32 \times 3$, which are width, height, and depth.

Our proposed system uses neural networks for implementation. This is the same as normal neural networks that are built with neurons, which had learnable weights and biases. Each neuron will perform dot product by getting little input and by utilizing bias as it accompanies non-linearity. This entire convent indicates different score function, coming from raw pixels on a single side to another side class score.

They had a loss function called SoftMax on the final layer, which is a completely associated layer. Here, the intake is images to convent, which permits encrypting some features in the design. All the features will perform the forward function more efficiently for performing and mostly decrease the quantity of parameters among a network. The main aim of the classification of an image is to take the features out of rough images.

3.1. Data Collection

The data was collected in many slot periods during wintertime using the FLIR Thermal Cam P10 LWIR camera of thermal imaging and arranged using a tripod with height of 140 cm with a standard $24^\circ \times 18^\circ$ FOV (field of view) lens, also by FLIR 7° FOV Telephoto Lens (P/B series) [31]. Additionally, the sensor in the camera takes thermal resolution, which is 320×240 pixels and is scaled up to 1280×960 pixels with an exterior recorder of video. To measure the distance, we utilized view ranger implementation [32], which is inserted on a CAT S60 [33] GPS-equipped smartphone. Detection using correlation filters is discussed in [34] and Infrared detection on image patch tensor model is explained in [35].

3.2. Algorithm

The below procedure discusses the steps of how to train and test the FLIR dataset to perform image classification. In traditional neural networks, each neuron is connected to all neurons in the preceding layer. However, in real time, this becomes impractical for higher-dimensional inputs, such as images. As an example, input volume is the size of $32 \times 32 \times 3$, and the receptive field is 5×5.

Table 1 discusses Hybrid Neural Network Algorithm. In a convolutional layer, each neuron is connected to a $5 \times 5 \times 3$ region in input volume, resulting in a total of 75 weights (and an additional +1 bias parameter). The number of interconnections along the depth axis is equal to 3, representing the depth of the input volume. To address the computational complexity associated with such large networks, hybrid neural networks employ a technique called parameter sharing. This technique allows the network to share weights across different regions or layers, effectively reducing redundancy and optimizing the overall efficiency of the network.

By sharing parameters, the network can leverage the inherent structure and patterns in the data, leading to improved generalization and reducing the risk of overfitting. This parameter-sharing strategy is particularly beneficial in convolutional neural networks (CNNs) used for image-processing tasks. In the context of image classification, the utilization of parameter sharing and convolutional layers allows CNNs to effectively capture local patterns and spatial hierarchies present in images. This results in more compact and efficient models that can accurately classify images across various classes.

Table 1. HDNN Algorithm.

Require	Size of the Batch Number of Classes Number of Epochs	128 2 5
Require	Input Image Dimension	28 × 28
Step #1	Load images of input from FLIR dataset with optimal feature vector	
Step #2	Take variable exploration	X = test data set (100, 28, 28, 1) Train dataset (600, 28, 28, 1)
Step #3	Create and compile the network design	
Step #4	Train the network using prepared dataset	

3.3. Classification

Retina net is a type of single-stage detector that addresses the challenge of class imbalance between the foreground and background in object detection. There are 2 techniques which retina net utilizes. The first one is the use of a feature pyramid network (FPN) backbone, which is built on top of a CNN. The FPN is responsible for extracting convolutional feature maps from the entire input image, allowing for multi-scale feature representation. The second technique is focal loss, which serves as a specialized loss function. Focal loss is designed to effectively handle the class imbalance problem by assigning higher weights to challenging examples and reducing the impact of easily classified examples. This helps to improve the overall performance of the detector, particularly for objects that are rare or difficult to detect. FPN is constructed on the uppermost CNN and is in charge of extracting convolutional feature maps from the complete image. By utilizing (focal loss/retina net changes weights with the loss function/focus on difficult/misclassified illustrations), which refines the accuracy of prediction. ByResNet (FPN) as the foundation to extract features, 2 subnetworks to classification, bounding-box regression, retina net has attained this recent stage in technological development and obtained the best performance.

Consider a building block where the output vector y is calculated as the function F and applied to the input vector x with Kalman weighted factors $\{KWi\}$ and then added to x:

$$y = F(x, \{KWi\}) + x \tag{1}$$

where:

- x and y represent the input and output vectors of the layers.
- The function $F(x, \{KWi\})$ represents the residual map that needs to be learned.

$$F = W_2\, \sigma(KW_1 x) \tag{2}$$

- The function $F + x$ is implemented using a shortcut connection, which involves element-wise addition.

Note that the parameters W_1 and W_2 are weight matrices, and K represents the Kalman weighted factors. The expression KW_1x indicates the intermediate result obtained by applying the weight matrix W_1 to the input vector x with the Kalman weighted factors. The resulting vector is then passed through the ReLU activation function and further transformed by the weight matrix W_2. The inclusion of the element-wise addition with the input vector x allows for the residual network to learn the residual mapping, facilitating the optimization process and enabling the network to effectively handle degradation issues. This approach provides flexibility in the structure of the residual function F, allowing for the incorporation of multiple layers and enhancing the expressiveness of the network. The utilization of residual connections and the element-wise addition operation on the feature

maps, channel by channel, further enhances the capabilities of convolutional layers within the network. This framework offers a powerful tool for learning complex features and addressing various challenges in deep learning tasks. The ReLU activation function (σ) is applied to the output of F, and biases are excluded to simplify the code.

Equation (1) is important in comparing plain and residual networks as it indicates that there are no additional parameters or computation complexity. This allows for a fair comparison between plain and residual networks that have the same specifications, such as depth, width, and computational cost (excluding the negligible element-wise addition). To ensure a fair comparison, the measurements of both the input vector x and the function F should be the same as in Equation (1). However, if the measurements are not the same (e.g., when changing input or output channels), a linear projection W_s can be accomplished through shortcut connections to match the measurements.

$$y = F(x, \{KW_i\}) + W_s x \tag{3}$$

This additional step ensures that the measurements of the input and the residual function match, allowing for a fair comparison between plain and residual networks. Equation (2) represents the inclusion of the linear projection W_s through shortcut connections.

In Equation (1), we utilize a square matrix W_s. However, extensive investigations have shown that using identity mapping is sufficient for addressing degradation issues and is computationally inexpensive. The matrix W_s is only utilized for matching measurements when necessary. The structure of the residual function F provides flexibility. In the present research, F consists of 2 or 3 layers, but it is possible to have additional layers. If F has only one layer, Equation (1) becomes a linear layer: $y = KW_{1x} + x$, which does not offer noticeable advantages (Equation (3)).

All the elements mentioned above are completely connected layers to simplify the implementation, and this has applications in convolutional layers. The function $F(x, \{KW_i\})$ indicates different convolutional layers, and the element-wise addition is performed on two feature maps, channel by channel.

This approach allows for the combination of multiple convolutional layers in F and the fusion of their results using element-wise addition, enabling the learning of more complex features and enhancing the expressiveness of the network.

The features for retina net classifier are obtained from OFSA—optimized feature selection algorithm [6], which achieved the best run time compared to the existing models.

Advantages of this retina net classifier are best feature selection, best performance is achieved in terms of accuracy, and pruning is more efficient. Pruning is done using weighted functions to achieve higher accuracy. Retina net utilizes a focal loss function to address class imbalance during training. We used the behavior of retina net, but all the features are hybrid.

4. Results

Figure 3 depicts the programmatically designed structure of the proposed neural network, illustrating the architecture and the connections between the layers.

The classification performed here is binary classification with two classes. Some datasets are used for testing, and some are used for training. Based on the iterations, the outputs are as shown.

Figure 4 showcases the accuracy achieved by the proposed network during the training and testing phases. It provides insights into the network's performance in correctly classifying the FLIR dataset. X-axis represents the number of epochs and Y-axis represents the accuracy.

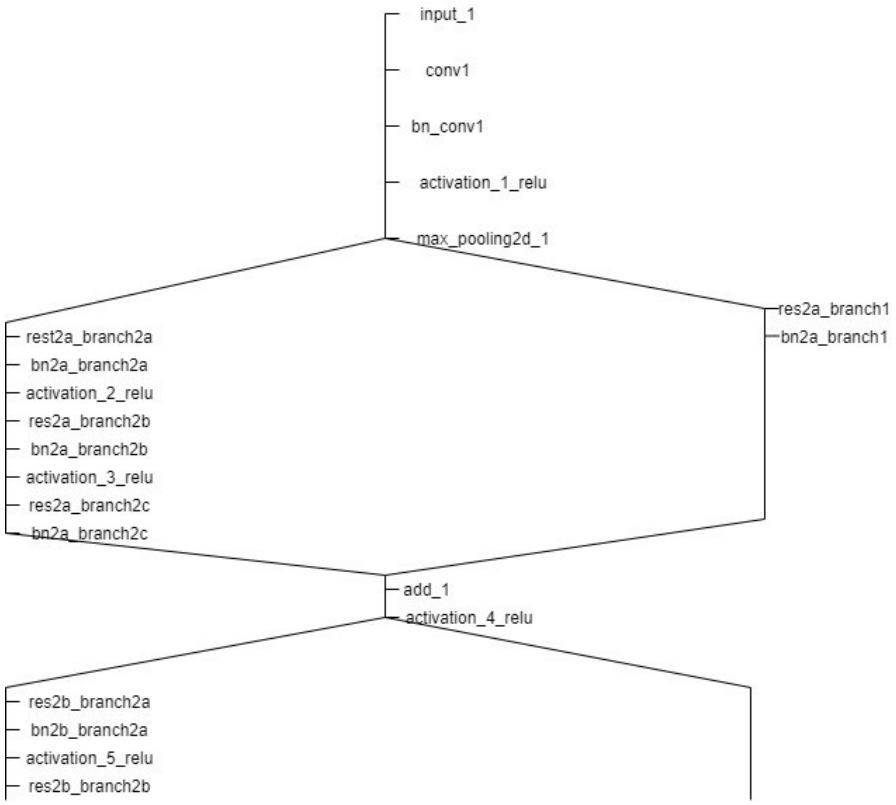


Figure 3. Proposed neural network structure programmatically.

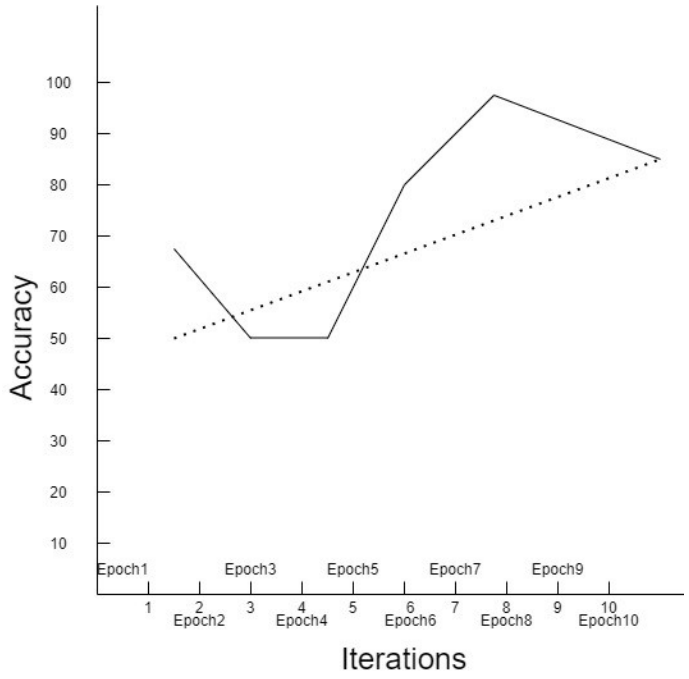


Figure 4. Accuracy for the proposed network.

Figure 5 presents the loss function and the number of iterations during the training process. It demonstrates the convergence of the network and the reduction in the loss function over time. X-axis represents the number of iterations and Y-axis represents the amount of loss.

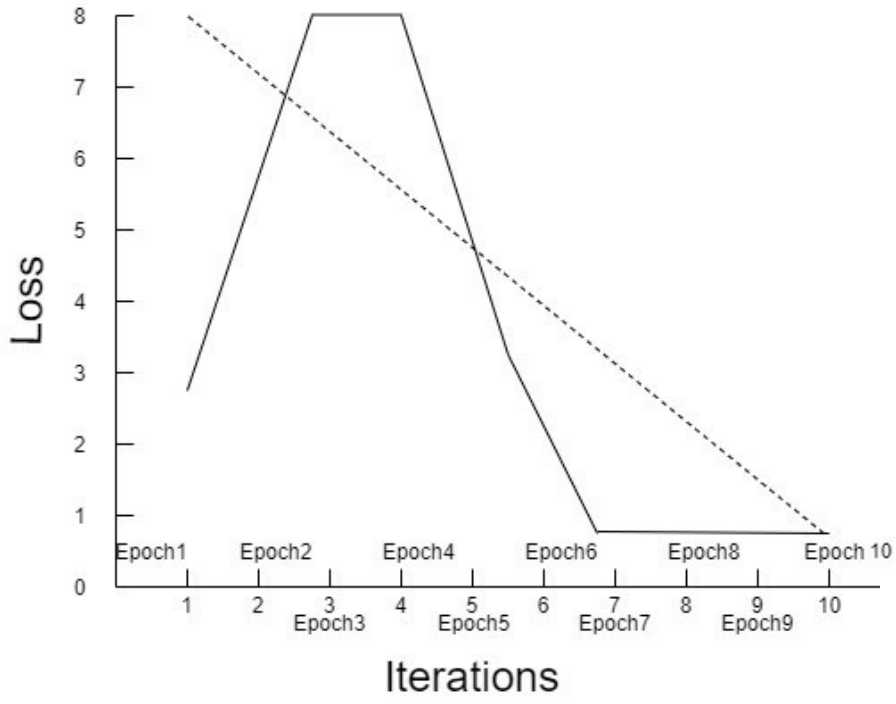


Figure 5. Accuracy of the network with loss and iterations.

The decreasing trend of the loss function in Figure 5 signifies that the network is effectively optimizing its parameters to better fit the training data. It indicates that the network is improving its ability to make accurate predictions and reduce errors during the training process.

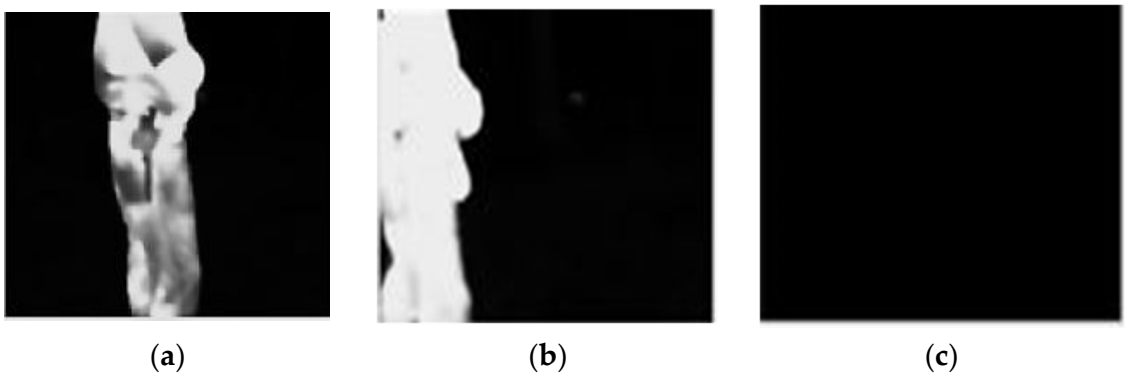
Monitoring the loss function during training is crucial as it helps in assessing the network's progress and determining if further training iterations are needed. It also aids in identifying potential issues, such as overfitting or underfitting, and guiding adjustments to the network architecture or training strategies if necessary.

Figure 6 shows examples of positive cases where the input image contains a detected weapon. It visualizes the input image, the detected weapon, and the corresponding highlighted area.

(**a**) (**b**) (**c**)

Figure 6. Weapon region detection from the input frame positive case:(**a**) represents input image; (**b**) represents detected weapon, i.e., positive case; (**c**) represents detected area.

Figure 7 shows examples of negative cases where the input image contains a detected weapon. It visualizes the input image, the detected weapon, and the corresponding highlighted area.



(**a**) (**b**) (**c**)

Figure 7. Weapon region detection from the input frame negative case:(**a**) represents input image; (**b**) represents not detected weapon, i.e., negative case; (**c**) represents non detected area.

Figure 8 highlights the performance metrics of proposed HDNN. (Hybrid Neural Network), for a specific dataset. The metrics evaluated include total frames, true positive (TP), false positive (FP), false negative (FN), sensitivity, positive predictive value (PPV), false alarm rate, and accuracy.

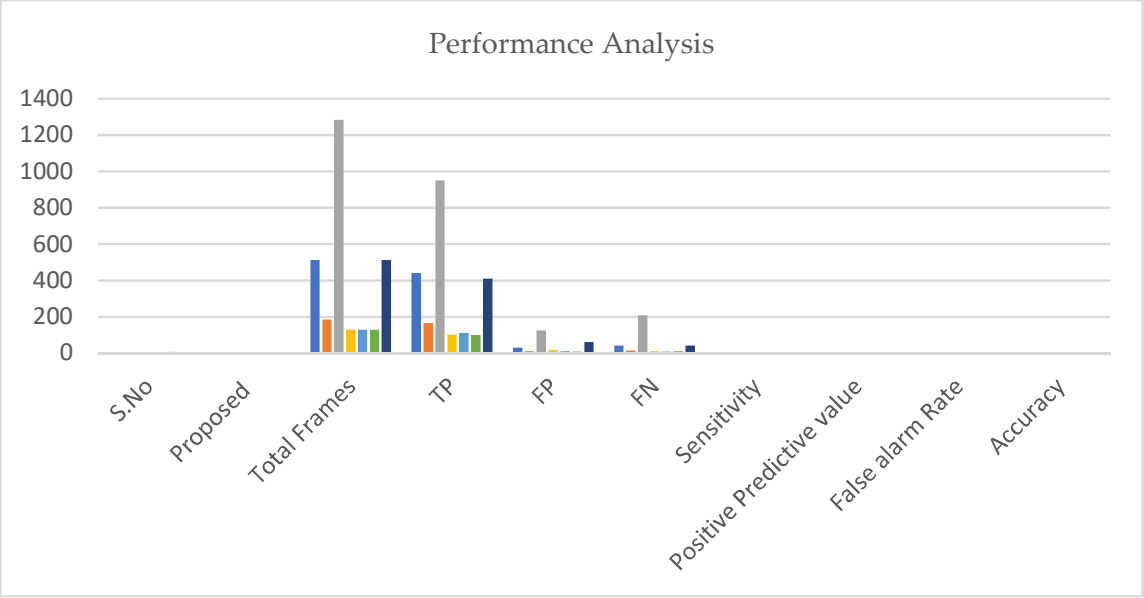


Figure 8. Performance analysis of the proposed method.

Table 2 represents the comparison of the proposed hybrid neural network (HDNN) with existing models. The proposed HDNN is compared with the DAG model, Haar transform + LBP, Kalman + Haar + LMP, Gaussian mixture models (GMM), and the existing convolutional neural network (CNN). Compared with the existing models, our proposed HDNN outweighed all of them in terms of accuracy.

The proposed HDNN is implemented with 512 frames, which achieved accuracy of 0.99. In the existing methods, such as GMM, Kalman Filter (HaaR + LBP), Kalman + HaaR + LBP, detection using aerial thermal views, detection and tracking in thermal videos using direct acyclic graph, accuracy is low. In parameters, such as TP, FP, FN, sensitivity, PTV, false alarm rate, the proposed HDNN gives better results.

The percentage of accurately classified instances is referred to as accuracy. The proportion of true positives that are successfully identified is measured by sensitivity or recall. The F-measure is a test accuracy metric. Precision, also known as positive predictive value,

is a measure of the number of relevant instances found among the retrieved instances. MCC assesses the accuracy of binary classification. The false alarm rate is calculated as the number of false alerts divided by the total number of non-events. True positive is defined as the positive class that was correctly predicted. True negative is defined as the negative class that was correctly predicted. False positive is an outcome in which the model forecasts the positive class inaccurately. False negative is an outcome in which the model forecasts the negative class wrongly.

Table 2. Comparison of Proposed and Existing Model.

#	Algorithms	Total Frames	TP	FP	FN	Sensitivity	Positive Predictive Value	False Alarm Rate	Accuracy
1	Proposed HDNN	512	440	30	42	0.99	0.966	0.036	0.99
2	Object detection and tracking in thermal video using DAG- Directed Acyclic Graph	184	165	10	14	0.977	0.88	0.054	0.98
3	People detection and tracking from aerial thermal views	1282	950	124	208	0.97	0.85	0.044	0.95
4	Kalman Filter (Haar + LBP)	128	102	16	10	0.95	0.89	0.032	0.92
5	Kalman + Haar + LBP	128	110	10	8	0.95	0.87	0.032	0.93
6	GMM	128	100	8	10	0.95	0.87	0.032	0.94
7	Existing CNN	512	410	60	42	0.95	0.912	0.032	0.95

5. Conclusions

In this research, we utilized a hybrid retina net for the classification of images using thermal images from the FLIR dataset. The proposed model was trained and tested using these thermal images. Impressively, the model achieved an accuracy of 99.5% in classifying the objects within the thermal images.

It is worth noting that processing thermal images computationally requires more time compared to regular JPEG images. However, by augmenting the network with additional layers, incorporating more training data, and leveraging the power of multiple GPUs, more accurate results can be obtained for object classification in thermal imaging.

By stacking additional layers, the network can capture more intricate features and patterns present in the thermal images, leading to enhanced classification performance. Moreover, training the network with a larger and more diverse dataset can further improve the model's ability to generalize and accurately classify objects in thermal images. Furthermore, leveraging the computational capabilities of groups of GPUs can significantly expedite the processing time for these thermal images, allowing for faster inference and analysis. This can be particularly advantageous in real-time applications or scenarios where quick decision-making based on thermal imaging data is crucial.

This research proposes a groundbreaking approach by utilizing a hybrid neural network for detection and classification tasks using thermal imaging. The proposed technique has demonstrated exceptional accuracy even under challenging conditions, such as extreme changes in illumination, occlusion, and longer distances.

One notable advantage of the suggested system is that it eliminates the need for pre-processing steps, such as image rectification and enhancement in thermal imaging. These pre-processing steps are typically employed to improve the quality of the image before detection and classification tasks. However, the proposed approach leverages the optimal features inherent in the thermal images themselves, allowing for more accurate tracking of objects without the need for extensive pre-processing. Additionally, the research highlights that earlier feature selection methods can have a detrimental impact on the training of CNN designs. By adopting the hybrid neural network approach, the proposed

system overcomes this limitation and achieves superior performance in object detection and classification tasks.

The ability to detect and classify objects accurately in thermal imaging without relying heavily on pre-processing or post-processing techniques is a significant advancement. It enables the system to effectively handle challenging scenarios where traditional methods may struggle.

Overall, the utilization of a hybrid retina net along with additional layers, more extensive training data, and the utilization of multiple GPUs offers promising avenues for achieving precise and efficient object classification in thermal-imaging applications.

Author Contributions: Conceptualization, V.T. and K.V.S.; methodology, K.V.S. and S.R.K.; software, K.V.S.; validation, V.T., K.V.S., S.R.K., A.S. and O.P.S.; formal analysis, S.R.K., K.V.S. and O.P.S.; investigation, K.V.S.; resources, V.T., K.V.S. and S.R.K.; data curation, A.S. and O.P.S.; writing—original draft preparation, V.T., K.V.S., S.R.K., A.S. and O.P.S.; writing—review and editing, K.V.S., A.S. and O.P.S.; visualization, K.V.S.; supervision, V.T. and K.V.S.; funding acquisition, A.S. and O.P.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by the project 38 PFE in the frame of the program PDI-PFE-CDI 2021.

Data Availability Statement: FLIR Dataset—https://www.flir.in/oem/adas/adas-dataset-form/.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ghose, D.; Desai, S.M.; Bhattacharya, S.; Chakraborty, D.; Fiterau, M.; Rahman, T. Pedestrian detection in thermal images using saliency maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019.
2. Heo, D.; Lee, E.; Ko, B.C. Pedestrian detection at night using deep neural networks and saliency maps. *Electron. Imag.* **2018**, *17*, 060403-1–060403-9. [CrossRef]
3. Ivasic-Kos, M.; Kristo, M.; Pobar, M. Human detection in thermal imaging using YOLO. In Proceedings of the 5th International Conference on Computer and Technology Applications (ICCTA), New York, NY, USA, 16–17 April 2019.
4. Ivasic-Kos, M.; Kristo, M.; Pobar, M. Person Detection in Thermal Videos Using YOLO. In *Intelligent Systems and Applications*; Bi, Y., Bhatia, R., Kapoor, S., Eds.; IntelliSys, Advances in Intelligent Systems and Computing; Springer: Cham, Switzerland, 2019; Volume 1038.
5. Gomez, A.; Conti, F.; Benini, L. Thermal image-based CNN's for ultra-low power people recognition. In Proceedings of the 15th ACM International Conference on Computing Frontiers, New York, NY, USA, 8–10 May 2018; pp. 326–331.
6. Teju, V.; Bhavana, D. An efficient object detection using OFSA for thermal imaging. *Int. J. Electr. Eng. Educ.* **2020**. [CrossRef]
7. Teju, V.; Bhavana, D. An efficient object tracking using optimamalkalman filter. *Int. J. Eng. Trends Technol.* **2021**, *69*, 197–202.
8. Liu, Q.; Lu, X.; He, Z.; Zhang, C.; Chen, W.S. Deep convolutional neural networks for thermal infrared object tracking. *Knowl.-Based Syst.* **2017**, *134*, 189–198. [CrossRef]
9. Rodger, I.; Connor, B.; Robertson, N.M. Classifying objects in LWIR imagery via CNNs. In Proceedings of the Electro-Optical and Infrared Systems: Technology and Applications, Edinburgh, UK, 21 October 2016; Huckridge, D.A., Reinharrd, E., Stephen, L., Eds.; SPIE: Bellingham, WA, USA, 2016; Volume 9987.
10. Shahid, N.; Yu, G.H.; Trinh, T.D.; Sin, D.S.; Kim, J.Y. Real-time implementation of human detection in thermal imagery based on CNN. *J. Korean Inst. Inf. Technol.* **2019**, *17*, 107–121. [CrossRef]
11. Wang, X.; Hosseinyalamdary, S. Human detection based on a sequence of thermal images using deep learning. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **2019**, *42*, 127–132. [CrossRef]
12. Zhang, H.; Luo, C.; Wang, Q.; Kitchin, M.; Parmley, A.; Monge-Alvarez, J.; Casaseca-de-la-Higuera, P. A novel infrared video surveillance system using deep learning based techniques. *Multimedia Tools Appl.* **2018**, *77*, 26657–26676. [CrossRef]
13. Farouk-Khalifa, A.; Badr, E.; Elmahdy, H.N. A survey on human detection surveillance systems for raspberry pi. *Image Vis. Comput.* **2019**, *85*, 1–13. [CrossRef]
14. Hou, Y.L.; Song, Y.; Hao, X.; Shen, Y.; Qian, M.; Chen, H. Multispectral pedestrian detection based on deep convolutional neural networks. *Infrared Phys. Technol.* **2018**, *94*, 69–77. [CrossRef]
15. Dai, X.; Duan, Y.; Hu, J.; Liu, S.; Hu, C.; He, Y.; Chen, D.; Luo, C.; Meng, J. Near infrared nighttime road pedestrians recognition based on convolutional neural network. *Infrared Phys. Technol.* **2019**, *97*, 25–32. [CrossRef]
16. Imran, J.; Raman, B. Deep residual infrared action recognition by integrating local and global spatio-temporal cues. *Infrared Phys. Technol.* **2019**, *102*, 103014. [CrossRef]
17. Gao, C.; Du, Y.; Liu, J.; Lv, J.; Yang, L.; Meng, D.; Hauptmann, A.G. InfAR dataset: Infrared action recognition at different times. *Neurocomputing* **2016**, *212*, 36–47. [CrossRef]

18. Lee, E.J.; Ko, B.C.; Nam, J.Y. Recognizing pedestrian's unsafe behaviors in far-infrared imagery at night. *Infrared Phys. Technol.* **2016**, *76*, 261–270. [CrossRef]
19. Lakshmi, A.; Faheema, A.G.J.; Deodhare, D. Pedestrian detection in thermal images: An automated scale-based region extraction with curvelet space validation. *Infrared Phys. Technol.* **2016**, *76*, 421–438. [CrossRef]
20. Qi, W.; Han, J.; Zhang, Y.; Bai, L.F. Infrared object detection using global and local cues based on LARK. *Infrared Phys. Technol.* **2016**, *76*, 206–216. [CrossRef]
21. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian Detection: A Benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 18 August 2009; pp. 304–311.
22. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 25 July 2005.
23. Viola, P.; Jones, M. Robust Real-time Object Detection. In Proceedings of the 2nd International Workshop on Statistical and Computational Theories of Vision, Modeling, Learning, Computing, and Sampling, Vancouver, BC, Canada, 13 July 2001.
24. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
25. Tome, D.; Monti, F.; Baroffio, L.; Bondi, L.; Tagliasacchi, M.; Tubaro, S. Deep convoluted neural networks for pedestrian detection. *Signal Process. Image Commun.* **2016**, *47*, 482–489. [CrossRef]
26. Angelova, A.; Krizhevsky, A.; Vanhoucke, V.; Ogale, A.; Ferguson, D. Real-Time Pedestrian Detection with Deep Network Cascades. In Proceedings of the British Machine Vision Conference, Swansea, UK, 7–10 September 2015.
27. Petrović, A.D.; Banić, M.; Simonović, M.; Stamenković, D.; Miltenović, A.; Adamović, G.; Rangelov, D. Integration of Computer Vision and Convolutional Neural Networks in the System for Detection of Rail Track and Signals on the Railway. *Appl. Sci.* **2022**, *12*, 6045. [CrossRef]
28. Sharma, M.; Lim, J.; Lee, H. The Amalgamation of the Object Detection and Semantic Segmentation for Steel Surface Defect Detection. *Appl. Sci.* **2022**, *12*, 6004. [CrossRef]
29. Zhang, H.; Wang, P.; Zhang, C.; Jiang, Z. A Comparable Study of CNN-Based Single Image Super-Resolution for Space-Based Imaging Sensors. *Sensors* **2019**, *19*, 3234. [CrossRef]
30. Coleman, S.; Kerr, D.; Zhang, Y. Image Sensing and Processing with Convolutional Neural Networks. *Sensors* **2022**, *22*, 3612. [CrossRef]
31. The Thermal Camera Lens. Available online: https://www.pass-thermal.co.uk/flir-131-mm-7-degree-telephoto-pb-series-lens (accessed on 15 December 2019).
32. Augmentra Ltd. ViewRanger: Trail Maps for Hiking, Biking, Skiing. 2018. Available online: https://play.google.com/store/apps/details?id=com.augmentra.viewranger.android (accessed on 10 May 2018).
33. Gsmarena.com. CAT S60—Full Phone Specifications. Available online: https://www.gsmarena.com/cat_s60-7928.php (accessed on 10 May 2018).
34. Yu, T.; Mo, B.; Liu, F.; Qi, H.; Liu, Y. Robust thermal infrared object tracking with continuous correlation filters and adaptive feature fusion. *Infrared Phys. Technol.* **2019**, *98*, 69–81. [CrossRef]
35. Zhang, X.; Ding, Q.; Luo, H.; Hui, B.; Chang, Z.; Zhang, J. Infrared small target detection based on an image-patch tensor model. *Infrared Phys. Technol.* **2019**, *99*, 55–63. [CrossRef]