Analysis of Multi-modal Imagery for Autonomous Driving

Anshika Sharma,19210993
Department of Computing
Dublin City University, Dublin, Ireland
Email: anshika.sharma27@mail.dcu.ie

Shivani Firke,19211077
Department of Computing
Dublin City University, Dublin, Ireland
Email: shivani.firke2@mail.dcu.ie

***Abstract*—The accurate perception of the environment is fundamental in understanding the autonomous driving system. An autonomous vehicle should precisely detect cars, pedestrians, cyclists, road signs, and other objects in real-time to make the correct control decisions to ensure safety. While several objects are surrounding a vehicle, people are amongst the most important parts of a machine's environment. People are highly mobile, with a high degree of variability in shape, size, and color. Major overlap in the group of people makes it more strenuous for the machine to perceive. In the past few years, the attempt to detect pedestrians has grown swiftly. To continue the rapid rate of experimentation and innovation, we make two major contributions, 1) We perform the object detection on aligned multi-spectral RGB and Thermal images, 2) We propose a new method to enhance the objective of pedestrian detection by proficiently blending the co-aligned images that help us carry out probing and informative comparisons based on the evaluations metrics. The proposed method was experimentally tested on the blended RGB-T images from KAIST pedestrian dataset. Our results show that the Yolov3 and Yolov4 show better pedestrian detection performance on the linearly blended RGB-T images as compared to normal RGB and Thermal images. Furthermore, our research also shows that these blended images show greater performance on images captured during night-time as compared to those captured during the daytime.**

***Index Terms*—Alpha Blending, KAIST, Multi-spectral, Pedestrian detection, RGB, Thermal, YOLO**

I. Introduction

Detecting different categories of objects in an image/video is one of the fundamental tasks in Autonomous Driving. Any vehicle capable enough to analyze its surroundings and make decisions with little or no human input is called an autonomous vehicle. An autonomous vehicle should detect the objects surrounding it with high precision. In the last few years, there has been a lot of work in this area but computer vision and deep learning have produced tremendous results. One of the best performing deep learning algorithms amongst R-CNN, Faster-RCNN, and Single Shot Detector is YOLO - You Only Look Once[1][19].

The two major processing steps in a typical object detection algorithm is feature extraction and detection. In feature extraction, most informative descriptors of the object are obtained followed by the classification and detection performed by scanning and analyzing the images/video frame by frame. YOLO as the name suggests performs object detection as a single regression problem. The image pixels are straight away represented as bounding box coordinates and their associated class probabilities are calculated simultaneously. We have used the latest versions of YOLO, YOLOv3, and YOLOv4[6][21][24][26], which have achieved state-of-the-art performance for object detection on the MS COCO dataset. Yolov4 divides the object detection task into two stages: regression, and classification.

The latest technologies are being used in several images capturing devices that extract necessary and useful information from the image. These images differ from each other depending on the type of the capturing device. One of the factors based on which images can be classified is the mode of capturing, it can be either thermal, RGB, LiDAR, radar, etc. We use the publicly available KAIST[13] multi-spectral pedestrian dataset, classified as a color-thermal dataset providing aligned color and thermal images. A thermal image is nothing but the heat of the object converted into a visible image that represents the distribution of temperature in the image. However, if for certain objects the temperature range is similar, thermal imaging can lead to misleading information and the objects become indistinguishable. Whereas, RGB image is a depiction of every pixel in the image in the form of red, green, and blue color.

Depending upon factors like unsatisfied lighting conditions, the quality of RGB image gets degraded which inflicts various challenges for networks utilizing only RGB images. To come up with robust and accurate pedestrian detection, we propose a concept of leveraging the features of both RGB and Thermal images and blend both the images through alpha blending. Alpha blending is a standard technique of combining two or more input images together to generate a resulting blended image, where each pixel of the blended image is the result of taking the weighted average of the corresponding pixels of the input images[33]. After experimenting and contrasting the performance of the algorithm on RGB, Thermal, and linearly blended images by evaluating the metrics. The main novelty of this research is to generate linearly blended images. The contributions are summarized as follows:

1. We perform an analysis of pedestrian detection using both RGB and Thermal images, individually.
2. We take a subset of both thermal and RGB images from

the same dataset and perform simple alpha linear blending to create the blended RGB-T images.

3. We run the same algorithms on the blended images and eventually compare and contrast the evaluated metrics of the RGB, Thermal, and Blended images.

Apart from the above-mentioned contributions, we understand different methods of fusion and object detection with the advancement in technology during the course of our experiment.

II. Related Work

Autonomous systems have been a topic of great interest since the beginning of the fourth industrial revolution. We review the contributions in the field of autonomous driving. The first one provides a detailed study of the multi-spectral KAIST Dataset. It comprised a wide collection of images for autonomous driving. The dataset included different perspectives of the world captured at various intervals of time and the drivable range of vehicles, pedestrians from urban to residential, day and night, including sunrise, sunset, afternoon, morning, night, and dawn. Both RGB and thermal cameras were co-aligned to attain the same orientation in both images[13]. A detailed examination of RGB, Thermal, a fusion of RGB and Thermal and LiDAR images along with the appropriate camera specifications is presented. Apart from object detection, the research is based on visual perception tasks like drivable range detection, localization, image enhancement, depth estimation, and colorization using a single/multi-spectral approach and lessons about the progress of capturing datasets. Despite the versatility of the dataset, the quality of the images both in RGB and thermal is low. Given the class *'person?'* it was difficult even for a human eye to detect if an object is a pedestrian or not[2].

Joseph Redmon introduced a single network, object detection algorithm, You Only Look Once. As the name suggests, YOLO frames the bounding boxes by assigning each object a confidence score and classified the object based on how confident it is about the categorization of the object, at once. Santosh Divvala compared the performance of YOLO to other object detection algorithms like DPM and R-CNN[13]. After experimenting with several combinations of different algorithms it was noted that YOLO outperformed both the algorithms individually and other combinations as well[1]. Even though it is a one-time evaluation, it has its drawbacks which can be overcome through its new versions like YOLOv3 and YOLOv4.

Image fusion reads multiple images from the same scene, thus retrieving vital information to put into a single output image which is more informative and suitable for visual perception or computer processing, thereby improving the quality and applicability of the images. This paper also gave a brief description of various image fusion techniques in the spatial and frequency domain, such as averaging, min-max, block replace, HIS, PCA, brovey, pyramid-based techniques, and transform-based techniques, along with various quality metrics for quantitative analysis of these approaches[3]. They provided individual feedback for each approach and suggested various methods for particular applications. The analysis was performed on several different datasets but KAIST generated the best results as it had perfectly co-aligned visible and thermal images, which makes it the best choice for our research.

However, there are few drawbacks of both RGB and Thermal images. Weixun Zuo points out that the primary drawback of RGB semantic segmentation networks is that RGB images are susceptible to degradation with inadequate lighting conditions while thermal cameras generate images using thermal radiation emitted by objects and are independent of lighting conditions. So, they proposed to fuse both the RGB and thermal information in a novel deep neural network[5]. An Encoder-Decoder design concept and ResNet was employed for feature extraction. In most of the cases, the model generated by the authors - RTFNet gave the best results as compared to DUC-HDC and FuseNet. The comparison was also based on day and night time and RTFNet produced the best results. After thorough research, the fusion of the images generated better results in all the cases[4].

In this paper, Volker Fischer took advantage of deep model-based methods of detection that have been effective in the visible domain and extended these approaches to the multi-spectral cases. The first attempt to use deep models to combine the images of a visible and thermal camera was performed. The introduced models and training methods were evaluated on the KAIST multi-spectral pedestrian detection benchmark and were compared to the state-of-the-art approaches[4]. They discovered that late fusion out-performed the state-of-the-art ACF+T+THOG solution whereas early fusion was not even close to the state-of-the-art solution. The late fusion had the minimum miss rate in all three cases, day, night, and all data split. It was noted that RGB images generated better results in the day as compared to thermal, and vice-versa [5].

Since the fusion architecture was generating good results, the challenge was the optimization of the performance of a convolutional neural network within limited resources. The most efficient neural network with good accuracy does not perform efficiently in real-time and requires large GPU's and resources for training massive data. The author was successful in creating CNN that functions efficiently in real-time on a conventional GPU[6]. This was achieved by optimizing parallel computations, designing a fast operating object detector. The state-of-the-art detector YOLOv4 was faster and more accurate than all the other detectors. Both the classifier's and detector's accuracy were improved. After gradual innovations in YOLO, the current best state-of-art object detection algorithm is YOLOv4.[6][26]

Due to low lighting conditions, people face several challenges in the field of fire fighting, rescue, maintenance, and medicine. For challenges like this, the authors introduced an idea of combining video streams both temporally and spatially. Using computer vision, non-uniform RGB, and infrared video streams were merged. Different techniques were implemented

for merging; few of them were, alpha blending RGB and thermal, adaptive per-pixel blending, binary alpha blending with adaptive scaling factor(alpha). The value of alpha was either fixed or varied based on the quality of the image. In the case of per-pixel blending, the value of alpha was decided according to the intensity of each pixel. They found that linearly blending RGB and thermal images generated one of the best results amongst all the methods adopted [7].

The human shape is easily distinguishable in color images captured during the daytime because of ample sunlight. Comparatively, a human shape is easily distinguishable during night-time in thermal images due to a significant difference in the temperatures of a human body and surrounding environment[13]. In the daytime, the bright sunlight is responsible for causing background clutters making the human shapes distinguishable from the background in the image domain. These reasons make it evident that using the features highlighted in both color and thermal images can achieve a better performance in pedestrian detection. Since there has been a lot of research in early and late fusion, our aim is to assess the probability of using a simple method of alpha linear blending in autonomous vehicles in order to increase the performance of Advanced Driver Assistance Systems (ADAS) in pedestrian detection.

III. Dataset and Algorithm selection

Dataset

We use the publicly available KAIST Multi-spectral Pedestrian Detection Dataset[13]. The dataset provides a range of traffic scenes in various regions from urban to residential and is not only captured in the rough time slots of both day and night, but also in the fine time slots of dawn, sunrise, morning, afternoon, dusk, sunset and night. It consists of 95,328 color-thermal pairs of images captured using a complex imaging hardware setup on the top of a vehicle. The imaging hardware comprised of a thermal camera, a color camera, a beam splitter, and a three-axis camera jig[13]. The beam splitter allows the optical centers of the two cameras to be aligned.

The captured 95k color-thermal images come with annotations that provide the ground truth for the images/videos, resulting in 103,128 dense annotations and 1,128 unique pedestrians[13]. The objects to be detected have four labels, namely, *person, people, cyclist, and person?* The *person* label was assigned to an individual pedestrian, a group of non-distinguishable pedestrians was labeled as *people*. The label *cyclist* was assigned to an individual riding a two-wheeled vehicle. When it comes to highly crowded scenes, it becomes difficult even for human annotators to determine whether an object that may be human-shaped is a pedestrian or not. Such objects are assigned the label *person?* The significant properties of the KAIST multi-spectral pedestrian dataset are scale, occlusion, position, and appearance change[13].

Hwang et al. divided the dataset into train and test sets, each set consisting of 6 subsets, 3 captured during daytime, and 3 captured during night-time, overall resulting in 12

Fig. 1. Color(Left) and Thermal(Right) images from KAIST dataset

subsets in the entire dataset. Due to limited hardware support, this research is conducted on a subset, precisely one-third of the colossal dataset. We have used 4 subsets, 1 day and 1 night from each train and test set, resulting in approximately 50,000 annotated color-thermal images.

Comparison with existing datasets

Some existing pedestrian datasets have been summarized in figure 2. The datasets have been horizontally classified into; color, thermal, and color-thermal, based on the image types provided in them. Nearly all existing color datasets[8][9][10][11][12] provide sequences of color images captured during the day and under pleasant weather conditions. The most widely used color datasets are Caltech[9] and KITTI[12] consisting of several real traffic scenarios. The Caltech dataset has the highest number of video frames.

	Training		Testing		Properties								
	# pedestrians	# images	# pedestrians	# images	# total frames	occ. labels	color	thermal	moving cam.	video seqs.	temporal corr.	aligned channels	publication
INRIA	1.2k	1.2k	566	741	2.5k		✓						'05
ETHZ	2.4k	499	12k	1.8k			✓			✓			'08
Daimler	15.6k	6.7k	56.4k	21.8k	28.5k		✓		✓	✓			'09
Caltech	192k	128k	155k	121k	250k	✓	✓		✓	✓	✓		'09
KITTI	12k	1.6k	–	–	80k	✓	✓		✓	✓			'12
OSU-T	984	1.9k	–	–	0.2k			✓		✓			'05
LSI	10.2k	6.2k	5.9k	9.1k	15.2k			✓	✓	✓			'13
ASL-TID	–	5.6k	–	1.3k	4.3k			✓		✓			'14
TIV	–	–	–	–	63k			✓		✓			'14
OSU-CT	–	–	–	–	17k		✓	✓		✓		✓	'07
LITIV	–	–	16.1k	5.4k	4.3k		✓	✓		✓		✓	'12
KAIST	41.5k	50.2k	44.7k	45.1k	95k	✓	✓	✓	✓	✓	✓	✓	'15

Fig. 2. Comparison of existing pedestrian datasets based on different properties.[13]

The OSU-T thermal dataset[14] was created for benchmarking tracking algorithms, while other datasets like ASL-TID, LSI, and TIV[15][16][17] provide a trajectory instead of a bounding box. The benchmark TIV dataset[17] provides the

largest number of multi-resolution image sequences, having annotated labels such as pedestrians, cars, bicycles, etc.

The KAIST multi-spectral pedestrian dataset is classified as a color-thermal dataset providing aligned color and thermal images. Compared to other color-thermal datasets OSU-CT[31] and LITIV[32], KAIST has a centrally aligned moving view of the traffic scene[13][18]. It has a sufficiently large number of annotated frames in comparison to the other datasets. KAIST also provides temporal correspondences along with occlusion labels, that play an important role in pedestrian detection, identification, and tracking[13].

You Only Look Once (YOLO)

In 2016, the first version of YOLO came out and set the cores of the algorithm. Redmond et al.[1] presented a new approach to object detection where they framed the object detection as a regression problem, as compared to prior approaches that repurposed the classifiers or localizers to perform detections[1][19]. They put forth an extremely fast unified architecture for real-time object detection using a base YOLO model that processes images at 45 frames per second, achieving twice the mean average precision than other leading-edge detection systems. The image pixels were represented as bounding box coordinates and their associated class probabilities[1].

YOLO is a single convolutional neural network that is capable of simultaneously predicting multiple bounding boxes and the associated class probabilities for these boxes. It trains on complete images while directly optimizing the detection performance[1]. The architecture of YOLO is inspired by the network of GoogleNet[20]. It consists of 24 convolutional layers that provide the functionality of feature extraction and 2 fully connected dense layers for making the predictions. However, unlike GoogleNet which uses inception modules, YOLO architecture simply makes use of 1x1 reduction layers and 3x3 convolutional layers, which reduces the feature space from the previous layer[19][20].

The algorithm works by subdividing an image into a grid of cells. For each cell, the algorithm then predicts the bounding boxes and confidence scores as well as the class probabilities[1][27]. Confidence scores are expressed in terms of Intersection over Union (IoU), a metric that measures how overlapped (intersection) the predicted bounding boxes and ground truth boxes are as a fraction of the total area (union) covered by the two together[27]. The prediction of locations of the bounding boxes, their sizes, and the confidence scores are taken into account by the loss that the algorithm minimizes, for the predictions and predicted classes[1][19][21][27].

Why YOLO?

The reasons for selecting YOLO for this research are as follows:

- YOLO is capable of detecting multiple objects in an image in one go. Since the detection is framed as a regression problem, there is no complex pipeline, which in turn makes the algorithm extremely fast and performant[1][19].
- As compared to sliding window and region proposal-based techniques, YOLO sees the entire image during training and testing, implicitly encoding contextual information about the class of the objects and their position[1][19][20]. Contrastingly, as Faster R-CNN [22] fails to see the larger context, it makes more background errors by predicting background patches as objects[19].
- YOLO, being highly generalizable, outperforms the leading detection methods like DPM[23] and R-CNN[22] by a wide margin as it learns the general representation of the object. When applied to any unexpected inputs, YOLO is less likely to crash[1][19].

YOLOv3

Yolov3, the third version of YOLO, was published in the year 2018 and had significant improvements over the previous two versions. Yolov3 could predict bounding boxes at different scales. Darknet, an open-source convolutional neural network framework written in C and CUDA, features as a basis for YOLO[21]. Darknet is immensely flexible and advances state-of-the-art object detection results. Darknet53, a variant of Darknet, serves as a backbone to Yolov3 resulting in a 106 layer fully convolutional underlying architecture for Yolov3. As the name suggests, the network is 53 layers deep and is a pre-trained version of a network that is trained on millions of images from the ImageNet database[21][24].

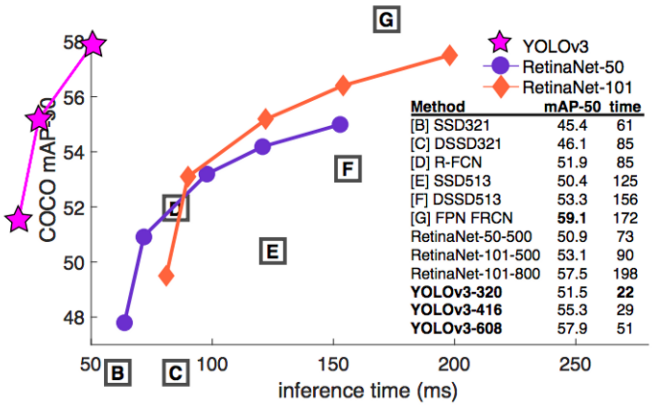

Fig. 3. Comparison of performance of Yolov3 and other state-of-the-art detectors[19][21]

The performance of Yolov3 has been at par with other cutting edge detectors like RetinaNet, it performs significantly faster at the COCO mAP 50 benchmarks, as seen in figure 2. The 50 in COCO 50 benchmark represents IoU, which is a measure of how well the predicted boxes and ground truth boxes align with each other. When this number increases for

say COCO 75 (IoU 0.75), RetinaNet outperforms Yolov3 as the predicted boxes of RetinaNet are well aligned compared to Yolov3[19][21].

YOLOv4

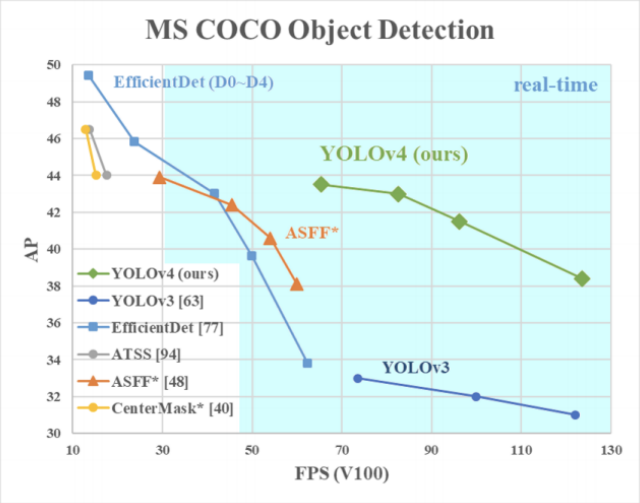

Fig. 4. Comparison of Yolov4 and other leading edge object detectors[6][26]

Progress in the YOLO family continues with the recent release of Yolov4, the fourth version of YOLO[26] It was published in April 2020 and has achieved state-of-the-art performance for object detection on the MS COCO dataset. Yolov4 divides the object detection task into two stages, regression, and classification. The first stage of regression is used to generate bounding boxes to identify the position of the object, while the second stage of classification determines the class of the object[6]. The architecture of Yolov4 consists of the (Cross Stage Partial Network connections) CSPDarknet53 neural network as a backbone, Spatial pyramid pooling block, and PANet path-aggregation as the neck and a Yolov3 head for prediction[6][26].

The key improvements in Yolov4 include an efficient, much faster, accurate, and powerful object detection model suitable for single GPU training[26]. As seen in figure 3, Yolov4 outperforms EfficientDet by running twice as fast with significantly comparable performance. It also shows improvement over Yolov3's AP by 10% and FPS by 12%[6].

IV. Methodology

In this research, we have performed an analysis of multi-modal imagery such as color, thermal and linearly blended color-thermal images to evaluate the performance of a benchmark object detection algorithm, while assessing the possibility of using linearly blended color-thermal imagery mode for autonomous driving systems. The hardware specifications used in the initial stage of the project include a 4GB NVIDIA GeForce 940MX GPU and 8GB CPU RAM, and a windows 10 operating system. At first, we started off with the huge KAIST Multi-spectral Pedestrian Benchmark dataset[2][13].

The annotations provided in the dataset had the below format:

```
% bbGt version=3
person 42 208 30 69 0 0 0 0 0 0 0
```

This format was inconsistent with the format required for our algorithm YOLO, which is

```
<object-class> <x_center> <y_center> <width> <height>
```

where object-class represents the class of the object to be detected, *x_center* represents the x centre coordinate of the object's bounding box, *y_center* represents the y center coordinate of the object's bounding box, *width* represents the width of the object's bounding box, and *height* represents the height of the object's bounding box. The values *x_center*, *y_center*, *width*, and *height* are decimal values relative to the width and height of the image, and lie within the range (0.0 to 1.0). The original annotations were converted into YOLO format using the below formulas:

```
<x> = <absolute_x> / <image_width>
<y> = <absolute_y> / <image_height>
<width> = <absolute_width> / <image_width>
<height> = <absolute_width> / <image_height>
```

The label *person* was mapped to object class 0, the *people* label was mapped to object class 1, *cyclist* was mapped to object class 2, and *person?* was mapped to object class 3. The final annotations had the following format:

```
0 0.0890625 0.47265625 0.046875 0.134765625
```

We made use of the darknet repository created by Bochkovskiy et al. [28] to train Yolov3 on our custom dataset for pedestrian detection. First, we created an obj.names file consisting of our four class labels, namely *person, people, cyclist, and person?* Next, we created an obj.data file consisting of a number of classes, absolute path to our training set and testing sets, the absolute path to our obj.names file and absolute path to a backup folder to save the weights obtained during training. In addition to that, we used the pre-trained weights darknet53.conv.74[28] for the convolutional layers. The pre-trained model has been trained for a very high number of classes and for a different size of the dataset, however, to use it for our custom dataset, we created our own custom configuration file.

This configuration file included details of all the convolutional layers required for training. The most important parameters in the file are batch=64, represents the number of images

chosen for each batch, classes=4, represent the number of classes, subdivision=16, represent the size into which a batch should be subdivided, and width=608, height=608, represent the input resolution for the images to be trained. Finally, we started the training on our custom dataset. However, due to hardware support limitations, we ran into a few speed bumps. The training was taking days to run and we had to migrate from the local system to Google colab.

Nonetheless, due to limited storage and expensive read-write operations on google drive, we decided to use a subset of the gigantic 36.5GB dataset. We have used precisely one-third of the dataset, that is 4 sets out of 12 sets. This criterion followed for choosing the subset is as follows:

- Two sets were selected from train and test sets each.
- Two sets were captured during the daytime, while 2 were captured during night-time. Contrastingly, as Faster R-CNN [22] fails to see the larger context, it makes more background errors by predicting background patches as objects[19].
- The four sets consisted of an equivalent number of images resulting in a balanced subset of approximately 50,000 color-thermal images.

With improved hardware specifications and comparatively lesser dataset to process, training process certainly speeded up. We planned to train the Yolov3 algorithm separately on color and thermal images in order to evaluate the performance of the algorithm on the two different types of images. Meanwhile, in April 2020, the fourth version of YOLO - Yolov4[6][28] was published, which showed immense improvements over Yolov3, increasing the average precision by 10% and frames per second by 12%[6]. Bearing in mind the recent improvements in the YOLO algorithm family, we decided to expand the scope of our project by training Yolov4 on the KAIST dataset to detect pedestrians. We used the new Yolov4 darknet repository and pre-trained weights yolov4.conv.137 provided by Bochkovskiy et al.[28] for custom training Yolov4 on KAIST. Similar steps mentioned for Yolov3 were performed to train Yolov4.

With the training in progress, we then moved on to the next stage of our project - combining the features from color and thermal images to evaluate the performance of the algorithm and assess the probability of using blended RGB-T images in autonomous driving. Motivated by the work on 'Saliency Guided 2D-object Annotation for Instrumented Vehicles' by Venkatesh et al.[29], we decided to explore the option of binary masking in order to highlight the important aspects of a thermal image. We used both color and thermal images in the KAIST dataset for creating masked RGB-T images. During the research, we came across the 'plantcv' image processing library[30] available in python which provides the functionality of creating masks and applying it to other images. It also provides the functionality of adaptive thresholding, binary thresholding, analysis of different workflows such as thermal, visible, PSII, dual image, etc[30].

Being keen on exploring the option of binary thresholding, we first rescaled the thermal image data to transform it into a color range between 0-255. Next, we converted the 3 channel

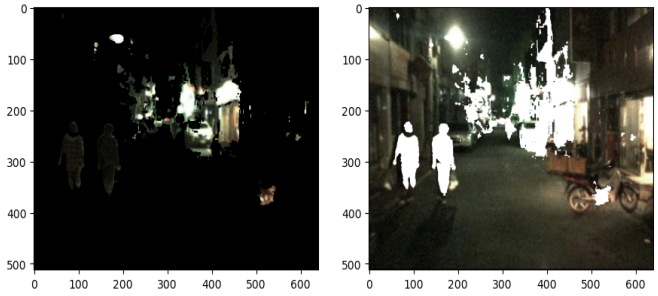

Fig. 5. Masked images created using binary thresholding

thermal input to a 2 channel grayscale image and used this grayscale image for binary thresholding. The parameters for creating a binary mask included input image, threshold, and object type[30]. The input image was the grayscale scaled thermal image and the threshold was experimentally determined by trying out different threshold values between 0-255. The object type was light or dark based on whether the object was lighter or darker than the background. We then used this binary mask and applied it to the RGB image. The resulting masked images were as shown in figure 5. The thermal images from the dataset showcased ambient amounts of heat which resulted in masked images that were not suitable for our research.

Finally, we adopted the method of simple alpha linear blending for blending the two modes of imagery highlighting the important features required for pedestrian detection. The images were linearly blended using the formula:

$$\text{Linearly Blended Image} = \alpha \times \text{Thermal} + \beta \times \text{RGB}$$

Fig. 6. Linearly Blended Color-Thermal Image

The value of *alpha* was chosen experimentally by using different values between 0 and 1. The optimal linearly blended output image was obtained by the value *alpha* = 0.75. The resulting image was as shown in figure 6. These linearly blended images were then fed to Yolov3 and Yolov4 algorithms to be

trained for pedestrian detection. The overall results have been summarized in the next section.

V. Experimental Results

In this section, we discuss the results obtained in this research. Overall, we trained our algorithms Yolov3 and Yolov4 on three modes of images each, color, thermal, and color-thermal of the KAIST Multi-spectral dataset. We have used the metrics mean Average Precision (mAP)[25] for evaluating the performance of our algorithms. In spite of hardware limitations, we obtained some interesting results and they have been divided into three categories:

A. Comparison of different modes of imagery

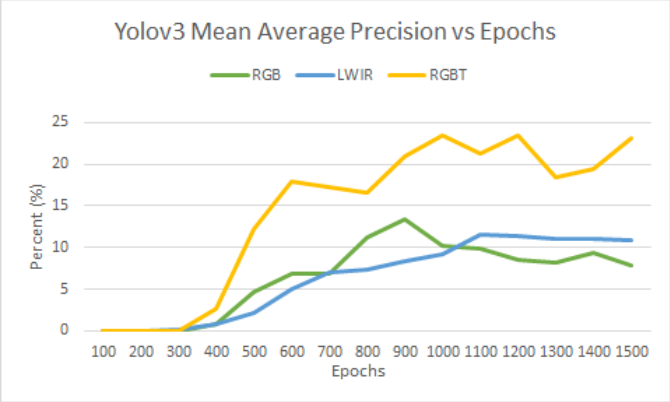

Fig. 7. Comparison of RGB, Thermal and Linearly Blended RGB-T images using Yolov3

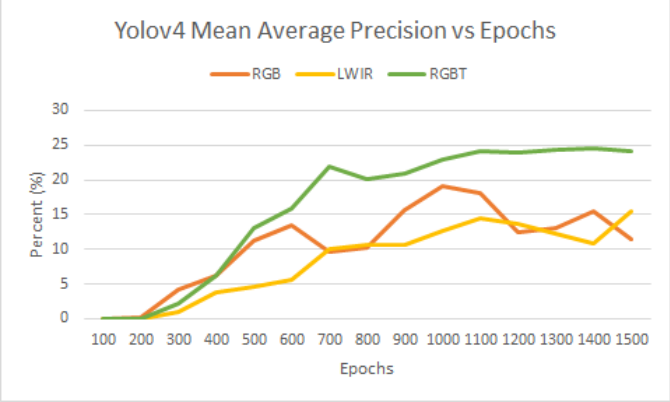

Fig. 8. Comparison of RGB, Thermal and Linearly Blended RGB-T images using Yolov4

We trained our algorithms on 3 types of images, namely RGB, LWIR/Thermal, and RGB-T while running 1500 training epochs for each category using both algorithms Yolov3 and Yolov4. Research over the past few years has shown that RGB or visible images are useful to perform better pedestrian detection during the daytime, while thermal images are useful in better pedestrian detection during night-time. Each mode of imagery having its own limitation, a combination of features from both image modes can be used for better and comprehensive pedestrian detection. The overall results obtained in our research show that RGB-T images, created using simple alpha linear blending, gave better and fast results as compared to RGB and Thermal images separately. This is visualized in the figures above.

As seen in the graphs, the mean Average Precision for RGB and Thermal images is comparable for both Yolov3 and Yolov4. Contrastingly, the mean Average Precision (mAP) values for linearly blended RGB-T are high as compared to both RGB and thermal imagery. We obtained the highest mAP of 24.5% using Yolov4, and by using Yolov3 it was 23.52% on the RGB-T blended images.

B. Comparison of the Yolov3 and Yolov4 for person detection

It has already been established in recent research that the performance of Yolov4 is much better as compared to the performance of Yolov3. We have compared the two algorithms on the average precision obtained for detecting the class *person* in a color, thermal and color-thermal images.

Fig. 9. Pedestrian Detection using Yolov3 in RGB (top left), Thermal (top right) and RGB-T (bottom) modes of images

Comparing the results of Yolov3 and Yolov4 on RGB images, the best-obtained mAP was 13.37% for Yolov3 and 19.11% for Yolov4. The highest average precision obtained for *person* detection in RGB images using Yolov3 was 46.89% while that for Yolov4 was 61.06%. Similarly, the comparison of Yolov3 and Yolov4 on thermal images yielded an mAP of 11.62% using Yolov3 and using Yolov4. Since the subset of 95k images was selected, amongst the 50k images there were very few scenarios in which cyclists and people were present, which eventually affected the collective result. However, the highest average precision for *person* detection using Yolov3 was 45.44% while that using Yolov4 was 53.21%.

Finally, we obtained outstanding mAP results for the linearly blended RGB-T images as discussed in the previous section. The highest average precision obtained for *person*

Fig. 10. Pedestrian Detection using Yolov4 in RGB (top left), Thermal (top right) and RGB-T (bottom) modes of images

detection in RGB-T images using Yolov3 was 94.09% and that using Yolov4 was 98.01%.

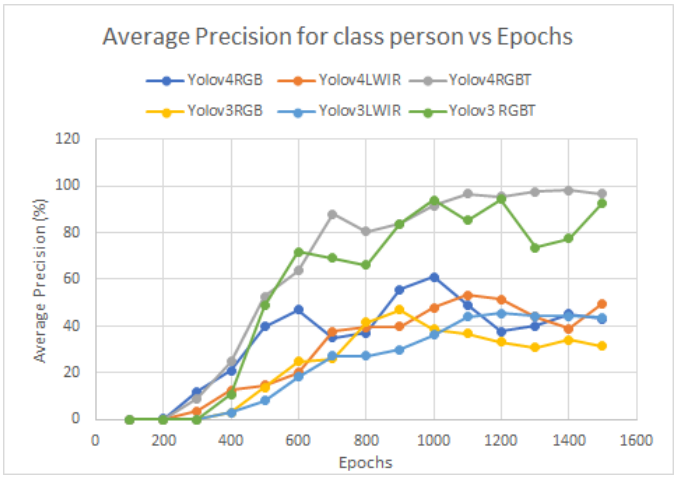

Fig. 11. Comparison of average precision for detection of the class person using Yolov3 and Yolov4 algorithms trained on RGB, Thermal and RGB-T.

C. Comparison of daytime/night-time pedestrian detection using blended images

Lastly, we have analyzed the performance of Yolov3 and Yolov4 trained on linearly blended color-thermal images by testing the best weights on two categories of images: those captured during daytime and night-time.

The results have been visualized in figure 10. The performance of RGB-T trained Yolov3 yielded the mAP of 19.15% for pedestrian detection in images captured during daytime while the mAP of 24.85% for pedestrian detection in images captured during night-time. Comparatively, the RGB-T trained Yolov4 produced better results with the mAP of 23.57% for pedestrian detection in daytime images and the mAP of 24.98% for pedestrian detection in night-time images. It can

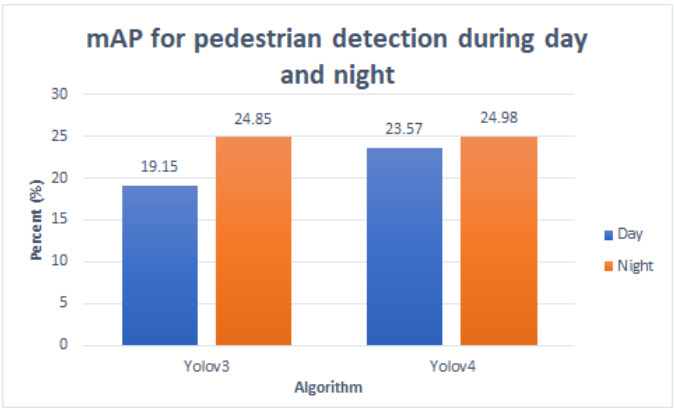

Fig. 12. Comparison of RGB-T trained Yolov3 and Yolov4 for pedestrian detection during daytime and night-time

be concluded from the results that simple linear blending can be used to combine the features of RGB and Thermal imagery modes, and this blended imagery can be used for pedestrian detection in autonomous vehicles during day and night time.

Fig. 13. Yolov3 RGBT Daytime detection (top left), Yolov4 RGBT Daytime Detection (top right), Yolov3 RGBT night-time Detection (bottom left) and Yolov4 RGBT night-time Detection (bottom right)

VI. Limitations

The main focus of this research was the analysis of multi-modal imagery such as visible(RGB), long-wavelength infrared (thermal) and color-thermal in pedestrian detection. Initially, we chose the FLIR[34] dataset for this research, however, it had its own limitations. FLIR is primarily a thermal images dataset captured using the FLIR automotive-qualified thermal sensor. Along with thermal images, FLIR provides a set of visible, LiDAR, and radar images. However, the dataset had its own limitations, the thermal and visible images, having been captured by two different imaging hardware, had different optical foci, resulting in non-aligned pair of images. Moreover, the visible images provided were non-annotated, and creating manual annotations for a huge dataset was not feasible[34].

Further research led us to the KAIST Multi-spectral Pedestrian Benchmark dataset which was perfectly tailored for our requirements. The color-thermal images, captured using imaging hardware consisting of color-thermal cameras separated by a beam splitter, were not only optically aligned but also the entire 95k pairs of color-thermal images were annotated. Next, a state-of-the-art object detection algorithm You-Only-Look-Once (Yolov3 and Yolov4) was selected for training on the custom dataset KAIST. Darknet, an open-source neural network framework written in C and CUDA, is used as the backbone for the algorithm. The processing of a huge dataset and darknet, being a neural network framework, requires high computing resources such as a GPU with sufficient memory and large disk storage.

The hardware support of a personal laptop with a 4GB NVIDIA GeForce 940MX GPU and 8GB CPU RAM was insufficient for such high and exorbitant processing. Comparatively, Google Collaboratory offered better hardware support including a 12GB RAM GPU and 68 GB disk but had its limitations too. The GPU access usage limits for Google Colab were a maximum of 12 hours, after which one cannot use their GPU for another 12 hours. In addition to that, Colab is an interactive environment and automatically disconnects after a certain period of inactivity, which makes it difficult for long-running activities, such as training the algorithm, to be kept active throughout the process.

Furthermore, as read and write operations are expensive and time-consuming on Google Drive, it was only feasible to use a subset of the colossal dataset for this research. Moreover, due to hardware limitations, 1500 training iterations were an achievable per mode of imagery for both algorithms.

VII. Conclusion

In this research, we performed an analysis of multi-modal imagery used for pedestrian detection in autonomous vehicles, particularly RGB (Visible/Color), LWIR (Thermal), and linearly blended RGB-T images using state-of-the-art object detection algorithms such as Yolov3 and Yolov4. We custom trained the algorithms on the optically aligned images from KAIST Multi-spectral Pedestrian benchmark dataset. As established in recent research, we obtained similar results proving the better performance of Yolov4 over Yolov3 in pedestrian detection.

Further, the analysis of color, thermal, and color-thermal images produced results that proved that the algorithms performed more efficiently on the linearly blended RGB-T images created using the simple linear alpha blending. Moreover, the comparison of the performance of RGB-T trained Yolov3 and Yolov4 models on images captured during daytime and night-time revealed that both the models performed much better for day and night images as compared to models trained on color and thermal images separately. Additionally, both models also performed efficiently on images captured during night-time as opposed to images captured during the daytime.

It can be concluded from the overall outcome of this research that simple linear alpha blending can be used in autonomous vehicles to combine the significant features from both color and thermal imagery for more accurate and faster pedestrian detection.

The future work of this research can include the use of various image fusion techniques[3] to improve the quality of fused RGB-T images leading to the improvement in the training and performance of state-of-the-art object detection algorithms. An approach that can be adopted as a part of future work is the development of an illumination-aware architecture capable of extracting the most significant features of a RGB and thermal imagery while feeding this combination of extracted features to a complex encoder-decoder architecture such as RTFNet[4] improving the real-time pedestrian detection performance of object detection algorithms.

Acknowledgment

The authors would like to thank practicum supervisor, Dr. Suzanne Little, for her able guidance and support in completing this research project and the department of computing at Dublin City University for this opportunity that allowed us to gain knowledgeable insights in the field of image processing, autonomous driving. and data analytics.

References

[1] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779 - 788, June 2016

[2] Choi, Y., Kim, N., Hwang, S., Park, K., Yoon, J.S., An, K. and Kweon, I.S., 2018. KAIST multi-spectral day/night data set for autonomous and assisted driving. IEEE Transactions on Intelligent Transportation Systems, 19(3), pp.934-948.

[3] D. Mishra, B. Palkar, "Image Fusion Techniques: A Review", International Journal of Computer Applications, vol. 130, pp. 0975-8887, 2015.

[4] Yuxiang Sun, Weixun Zuo and Ming Liu, "RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes", IEEE Robotics and Automation Letters, Volume 4, July 2019, pp. 2576-2583

[5] J. Wagner, V. Fischer, M. Herman, and S. Behnke. Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks. In European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 2016.

[6] Bochkovskiy, A., Wang, C.Y. and Liao, H.Y.M., 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934.

[7] Orlosky J, Kim P, Kiyokawa K, Mashita T, Ratsamee P, Uranishi Y, Takemura H. Vismerge: Light adaptive vision augmentation via spectral and temporal fusion of non-visible light. In2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) 2017 Oct 9 (pp. 22-31). IEEE.

[8] Dalal, N. and Triggs, B., 2005, June. Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 886-893). IEEE.

[9] Dollár, P., Wojek, C., Schiele, B. and Perona, P., 2009, June. Pedestrian detection: A benchmark. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 304-311). IEEE.

[10] Enzweiler, M. and Gavrila, D.M., 2008. Monocular pedestrian detection: Survey and experiments. IEEE transactions on pattern analysis and machine intelligence, 31(12), pp.2179-2195.

[11] Ess, A., Leibe, B., Schindler, K. and Van Gool, L., 2008, June. A mobile vision system for robust multi-person tracking. In 2008 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). IEEE.

[12] Geiger, A., Lenz, P. and Urtasun, R., 2012, June. Are we ready for autonomous driving? the kitti vision benchmark suite. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 3354-3361). IEEE.

[13] Hwang, S., Park, J., Kim, N., Choi, Y. and So Kweon, I., 2015. Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1037-1045).

[14] Davis, J.W. and Keck, M.A., 2005, January. A two-stage template approach to person detection in thermal imagery. In 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)-Volume 1 (Vol. 1, pp. 364-369). IEEE.

[15] Khellal, A., Ma, H. and Fei, Q., 2015, August. Pedestrian classification and detection in far infrared images. In International Conference on Intelligent Robotics and Applications (pp. 511-522). Springer, Cham.

[16] Portmann, J., Lynen, S., Chli, M. and Siegwart, R., 2014, May. People detection and tracking from aerial thermal views. In 2014 IEEE international conference on robotics and automation (ICRA) (pp. 1794-1800). IEEE.

[17] Wu, Z., Fuller, N., Theriault, D. and Betke, M., 2014. A thermal infrared video benchmark for visual analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 201-208).

[18] Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M.C., Qi, H., Lim, J., Yang, M.H. and Lyu, S., 2015. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. arXiv preprint arXiv:1511.04136.

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

[20] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[21] Redmon, J., 2020. YOLO: Real-Time Object Detection. [online] Pjreddie.com. Available at: ¡https://pjreddie.com/darknet/yolo/¿.

[22] Girshick, R., 2015. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[23] Felzenszwalb, P.F., Girshick, R.B., McAllester, D. and Ramanan, D., 2009. Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence, 32(9), pp.1627-1645.

[24] Se.mathworks.com. 2020. Darknet-53 Convolutional Neural Network - MATLAB Darknet53- Mathworks Nordic. [online] Available at: ¡https://se.mathworks.com/help/deeplearning/ref/darknet53.html¿ [Accessed 16 August 2020].

[25] Padilla, R., Netto, S.L. and da Silva, E.A., 2020, July. A Survey on Performance Metrics for Object-Detection Algorithms. In 2020 International Conference on Systems, Signals and Image Processing (IWSSIP) (pp. 237-242). IEEE.

[26] Hui Jonathan, Medium. 2020. Yolov4. [online] Available at: ¡https://medium.com/@jonathan_hui/yolov4-c9901eaa8e61¿.

[27] Pugliese, M., 2020. A Very Shallow Overview Of YOLO And Darknet. [online] ClearlyErroneous. Available at: ¡https://martinapugliese.github.io/recognise-objects-yolo/¿.

[28] GitHub. 2020. Alexeyab/Darknet. [online] Available at: ¡https://github.com/AlexeyAB/darknet¿.

[29] Venkatesh, G.M., Hu, F., O'Connor, N.E., Smeaton, A.F., Yang, Z. and Little, S., 2019, September. Saliency Guided 2D-Object Annotation for Instrumented Vehicles. In the 2019 International Conference on Content-Based Multimedia Indexing (CBMI) (pp. 1-7). IEEE.

[30] Team, P., 2020. Thermal Workflow - Plantcv. [online] Plantcv.readthedocs.io. Available at: ¡https://plantcv.readthedocs.io/en/stable/thermal_tutorial/¿ [Accessed 16 August 2020].

[31] Davis, J.W. and Sharma, V., 2007. Background-subtraction using contour-based fusion of thermal and visible imagery. Computer vision and image understanding, 106(2-3), pp.162-182.

[32] Torabi, A., Massé, G. and Bilodeau, G.A., 2012. An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications. Computer Vision and Image Understanding, 116(2), pp.210-221.

[33] Gerber, R., Intel Corp, 1998. Alpha blending palettized image data. U.S. Patent 5,831,604.

[34] Flir.com. 2020. Thermal Datasets For ADAS Algorithm Training — FLIR Systems. [online] Available at: ¡https://www.flir.com/oem/adas/dataset/¿