UNIVERSITÀ DI PARMA
ARCHIVIO DELLA RICERCA

University of Parma Research Repository

Thermal Imaging on Smart Vehicles for Person and Road Detection: Can a Lazy Approach Work?

This is the peer reviewd version of the followng article:

Original
Thermal Imaging on Smart Vehicles for Person and Road Detection: Can a Lazy Approach Work? / Humblot-Renaux, G; Li, V; Pinto, D; Marchegiani, M. - (2020). ((Intervento presentato al convegno IEEE CONFERENCE ON INTELLIGENT TRANSPORTATION SYSTEMS [10.1109/ITSC45102.2020.9294671].

Availability:
This version is available at: 11381/2931522 since: 2022-11-04T13:08:43Z

Publisher:
Institute of Electrical and Electronics Engineers Inc.

Published
DOI:10.1109/ITSC45102.2020.9294671

Terms of use:
openAccess

Anyone can freely access the full text of works made available as "Open Access". Works made available

Publisher copyright

(Article begins on next page)

19 November 2022

Thermal Imaging on Smart Vehicles for Person and Road Detection: Can a Lazy Approach Work?

Galadrielle Humblot-Renaux*, Vivian Li*, Daniela Pinto*, and Letizia Marchegiani

***Abstract*— This paper proposes the addition of a thermal camera to an RGB system with the goal of improving person and road detection reliability in unfavorable weather and illumination conditions. Custom data is gathered on an experimental vehicle and used for development and testing. For person detection, we propose a novel multi-modal approach, where bounding boxes are initially obtained from RGB and thermal images using YOLOv3-tiny. We then identify high-intensity connected components in thermal images to compensate for missed detections. Detections from the two cameras and the two algorithms are finally weighed and combined into a confidence map. Using the proposed fusion method, recall and precision are improved compared to using RGB only, without the need to retrain the network. For thermal-based road segmentation, we achieve an average precision of 94.2% after re-training MultiNet's KittiSeg decoder on a small thermal dataset, while using pre-trained weights for MultiNet's VGG-based encoder. These results show that the addition of thermal cameras to perception systems of autonomous vehicles can bring substantial benefits with minimal labelling, implementation effort and training requirements.**

I. Introduction

Autonomous vehicles rely on exteroceptive sensors to find a navigable path while avoiding obstacles. Traditionally, cameras are the sensor of choice for detecting obstacles in the scene. However, these often fall short when facing non-ideal illumination and weather conditions, as they are inherently sensitive to any visual change in the scene, such as darkness, fog, rain or glare from the sun [1]. Other sensor modalities have been used for similar purposes, such as LIDAR [2] and microphones [3], [4]; yet, while LIDAR also suffers in harsh weather conditions (*e.g.* heavy rain, fog), acoustic sensing cannot, alone, provide full understanding of the environment. Radar is currently considered a valid solution, as quite resilient to a wide range of weather conditions, and able to detect objects at long range [5]. However, despite the recent progress in this direction (*e.g.* [6]), the interpretation of radar data remains challenging, due also to the presence of noise and unwanted artifacts. This imposes significant limitations when trying to leverage existing tools in computer vision to parse the data, and when creating a labelled dataset for object detection tasks in radar scans.

In this work, we evaluate the benefits and potential of adding a thermal camera to an autonomous vehicle for urban environment understanding. The vehicle we employ in this study is an experimental golf-cart which operates on a university campus, driving primarily on unmarked roads and bicycle lanes with heavy pedestrian traffic. Given this application context, we focus our investigation on two crucial tasks: person and road detection. However, our findings could be extended to other detection tasks (*e.g.* vehicle detection).

Much like traditional cameras, thermal cameras provide the visual cues necessary to not only detect obstacles, but also to distinguish among different types of objects. They also share many of the useful properties of the radar: indeed, they are not sensitive to visible light, they do not rely on any illumination source, and do not "see" on-coming headlights, smoke, haze, etc. For this reason, they can be used to detect heat sources, such as people, through rain, snow or fog, even though these conditions may lead to a decrease in range or contrast [7]. Compared to radars, thermal cameras provide a much more *humanly intuitive* representation of the environment, simplifying the labelling process. Furthermore, given the nature of the data, computer vision methods and techniques normally adopted in the RGB domain could be adapted and employed with minimal effort.

In this study, we propose a novel method for multi-modal person detection, where the predictions obtained on RGB and thermal images are weighed and combined into a single confidence map. Firstly, we generate bounding box estimates by employing a YOLOv3-tiny (You Only Look Once) architecture [8] on both kinds of data (*i.e.* RGB and thermal images). The network is used with pre-trained weights, without the need for additional retraining, or the need to generate a labelled training set of thermal images. Secondly, connected components in thermal images are identified and employed to compensate for missed detections. Predictions are lastly scored and integrated into a confidence map. Additionally, we present a thermal image-based road detection framework, implemented through a MultiNet architecture [9], using pre-trained VGG16 weights for the encoder [10], and only re-training the KittiSeg segmentation decoder, such that the network is trained with very little thermal data. Note that we use the terms "road segmentation" and "road detection" interchangeably throughout this paper. Our evaluation, based on real data collected with our experimental vehicle, demonstrates that thermal cameras could be a compelling alternative for vision-based systems operating on autonomous vehicles, both if used as a single modality, and in combination with RGB cameras. By taking a "lazy" approach which leverages existing deep learning networks pre-trained on RGB data, we also show that enabling thermal vision on smart vehicles does not necessarily require developing dedicated architectures or annotating large datasets.

Authors are members of the Department of Electronic Systems, Aalborg University, Denmark, {ghumbl19, vli16, dpinto16}@student.aau.dk; lm@es.aau.dk

*Authors contributed equally to this work.

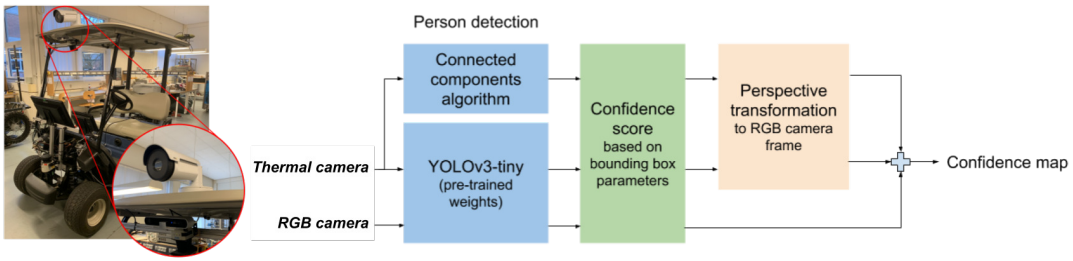


Fig. 1. Person Detection: High-level diagram showing how detections from thermal and RGB images are combined into a confidence map. The figure also reports the experimental platform employed in this study, where the position of the two visual sensors is highlighted.

II. Related Work

A. Person Detection

Detecting people accurately is a crucial task to guarantee safety in autonomous vehicles. While a vast number of approaches have been proposed over the years for this purpose, Convolution Neural Networks (CNNs) are now considered the ones yielding the greatest performance [11] [12]. Thermal imaging is an attractive approach for person detection since humans will always appear "heated" in the scene, regardless of the illumination. One of the first attempts to rely only on the use of thermal images for pedestrian detection has been presented in [13]. The approach makes use of Support Vector Machines (SVMs), but manages to achieve only a 35% detection rate. More recently, in [14] a YOLOv3 network is retrained on thermal images, obtaining remarkable results in the context of surveillance. In [15], the authors introduced a modified version of YOLO, YOLO Darknet for object recognition on long range thermal images. A system exploiting the stereo information from two thermal cameras has been described in [16]. The possibility of combining RGB and thermal data have been also approached in a number of works, starting from [17], which demonstrate that even adding a low-cost, low-resolution infra-red sensor to an RGB person tracker can significantly improve the performance of the system. Lately, [18] presented a CNN-based object recognition framework, where the training dataset is augmented by model images, created using object 3D models textured by real color and thermal images. Our study, which employs a mono high-resolution thermal camera is close in spirit to the analysis proposed in [19]. Yet, compared to that work, where a set of CNN architectures is retrained by mixing different combinations of RGB, thermal and VOC data [20], our approach does not require additional training and labelling (*cf.* Section I), crucial aspects which can greatly facilitate its use.

B. Road Detection

While the literature provides many different approaches for road detection in RGB images (*e.g.* [21], [22]), little research is available on road detection using thermal imaging. One of the reasons might be that many traditional road detection algorithms rely on white markings which cannot be recovered by a thermal camera. Furthermore, most large publicly-available datasets used for neural network training only contain RGB images. The authors in [23] present a robust road detection method based on a thermal system, which, however, requires two cameras to compute disparity information. In [24], a scene-adaptive sampling method for road detection in the thermal domain is presented. The approach yields good performance; however, the extensive evaluation carried out shows that Fully Convolutional Networks (FCNs) out-perform the approach, as well as other deep learning architectures. Building on those results, we opt for the employment of an FCN-based KittiSeg network, which, even when trained on a small dataset of mono thermal images (*i.e.* no stereo information or large amount of data needed) achieves remarkable performance.

III. Methods

A. Person Detection

The chosen person detection method is the real-time object detection system, YOLOv3-tiny [8]. A major advantage of the YOLO architecture is that it is able to perform detections in a single network pass, which ranks it amongst the fastest state-of-the-art object detectors without sacrificing performance. YOLOv3-tiny is a smaller model of the original YOLOv3 which satisfies the requirements of working real-time with non-dedicated hardware with an accuracy trade-off. The network consists of 23 layers and is pre-trained on the COCO dataset [25] to detect over 80 different object classes. Yet, using pre-trained YOLOv3-tiny on our own data may result in missed or erroneous detections. For this reason, we take advantage of the unique features of thermal images to develop an alternative algorithm based on connected components labelling to use in combination with YOLOv3-tiny. Since people stand out in their intensity level from the rest of the scene due to their body heat, they can be extracted with simple adaptive thresholding. This gives a binarized image. Connected components labelling is then used to identify groups of connected pixels likely to belong to the same object. Further analysis of each resulting bounding box is required to discard boxes which are unlikely to be people.

We use the following intuitions to give each bounding box a confidence score based on its shape and position:

- aspect ratio of the box: a highly imbalanced aspect ratio (eg. a very long & thin box) would not correspond to a person;

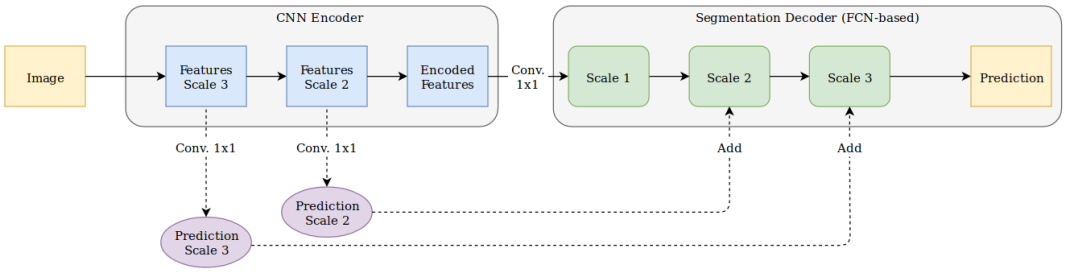


Fig. 2. Road Segmentation Architecture. Modified image from [9].

- relation between the box position and its height: the further away a person is, the smaller they appear and the smaller their y-position.

We also discard any detection which is positioned higher than the horizon, as well as bounding boxes below a specific height (since we are mostly concerned with detecting people in the vicinity of the vehicle). The score calculation and constraints are described in Section IV-B. The workflow of the proposed method is illustrated in Fig. 1.

After a detection algorithm is run on both camera frames, the resulting two sets of bounding boxes may not overlap. In order to compare and combine both sets of detections, all bounding boxes should be mapped onto a single reference frame. Since two cameras cannot have the exact same position, and may also have different orientations or Fields of View (FOV), it is necessary to find the homography relating their respective image planes. Due to the different nature of RGB and thermal images, features cannot be extracted and matched algorithmically. Therefore, in order to find a transformation from one camera frame to the other, a set of feature points is manually selected. The two sets of feature points are matched using model fitting (with all points considered inliers) to estimate a perspective transformation. The perspective transformation is then applied to the bounding box coordinates from one camera to map them to the other camera's frame. This results in a single image in the reference frame featuring bounding boxes from both cameras. The goal is then to combine these 3 sets of bounding boxes into a single image. As introduced above, for each bounding box, a score is computed indicating how likely it is that it really corresponds to a person. A confidence map is then generated for each algorithm as follows:

- start with an empty map (an array of zeros with the same dimensions as the captured image);
- sort bounding boxes by their weight, starting by the lowest;
- for each bounding box, set the value of the corresponding map area to the bounding box score.

The confidence maps for each algorithm are then simply added into a single map and normalized.

B. Road Segmentation

The goal is to identify navigable road area without relying on markings. We do so by relying on the semantic segmentation pipeline of the MultiNet model presented in [9]. This network is chosen as it achieves high performance on the Kitti Benchmark [26] (for both marked and unmarked roads), has the capability to run in real-time, supports grayscale images and only requires a small dataset for training.

MultiNet follows an encoder-decoder structure, as illustrated in Fig. 2. The encoder is based on the VGG16 architecture, and trained on the ImageNet dataset [27]. The segmentation decoder KittiSeg is based on a FCN architecture and was originally trained on the Kitti Vision Benchmark dataset [26]. In order to obtain good performance on the benchmark, the authors disabled data augmentation during training. This causes over-fitting, which makes the pre-trained weights perform poorly on data outside of the Kitti dataset, especially thermal data which differs from the original RGB dataset by its very nature. Therefore, instead of using the original weights, we re-train the KittiSeg decoder on our custom dataset of thermal images.

IV. Experiments

We perform three different experiments: firstly, we investigate whether using the YOLOv3-tiny network with pre-trained weights is a viable option for person detection in thermal images. Secondly, we analyse the behaviour of the multi-modal scheme we propose in this paper. Lastly, we evaluate the performance of the road segmentation method. For all the experiments, we rely on data collected with our golf cart, equipped with a ZED RGB camera (with resolution of 1280×720), positioned at the front of the vehicle, and an AXIS Q1942-E thermal camera, mounted on the roof (with resolution of 800×600, operating in the long-wave infrared (LWIR) range). Our custom RGB and thermal datasets are collected during sunny and cloudy weather at different times of the day in winter (average temperature of 7°C).

A. Person Detection with YOLOv3-tiny (pre-trained weights)

This experiment investigates how YOLOv3-tiny with pre-trained weights performs on thermal images, using RGB images as a base-line. A custom dataset consisting of 200 RGB images and 200 thermal images featuring people is annotated and used for testing. Default parameters [8] are used during inference. In order to compare the performance of YOLOv3-tiny on RGB and thermal images, the output of approximately the same time-frames are compared and evaluated. Bounding box precision and recall are computed for different overlap thresholds λ. Figure 3 shows the precision-recall

curve for both thermal and RGB images, using YOLOv3-tiny with pre-trained weights: for almost every λ, both the precision and recall are higher for thermal images. This suggests that even though it was pre-trained on an RGB dataset, the YOLOv3-tiny network can be directly used on our thermal images without compromising performance.

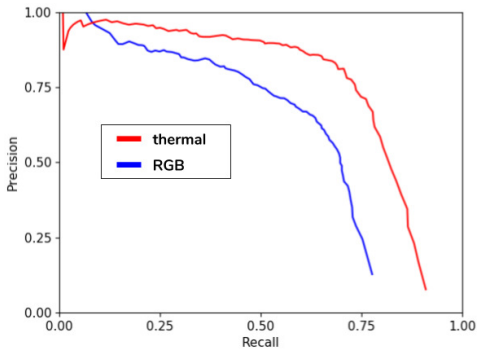


Fig. 3. Bounding box Precision-Recall curve for YOLOv3-tiny's performance on thermal (red) and RGB images (blue).

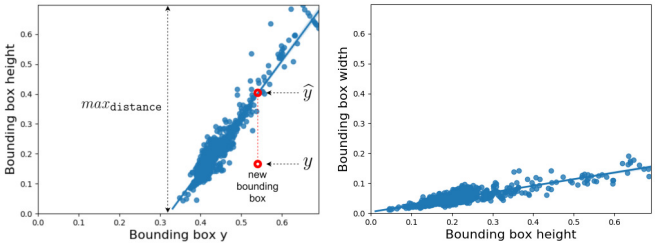
The multi-modal scheme described below requires choosing a λ for YOLOv3-tiny predictions. In the context of autonomous vehicles, failing to detect a pedestrian is a critical safety issue: it is more important to minimize false negatives than false positives. Therefore a higher recall value is preferred. Based on the curve in Fig. 3, we thus set λ to 0.2 for inference on new images, corresponding to a precision/recall pair of 0.781/0.715 for thermal and 0.594/0.665 for RGB images.

B. Person Detection with the Multi-Modal Scheme

The initial step consists in identifying connected components in thermal images. This requires that we first obtain a binary image which successfully separates people from the background. For this, a fixed constant threshold is a poor choice since the intensity distribution may vary across captures. Therefore, an adaptive threshold is preferable. To pick a suitable threshold, we look at the intensity distribution of different captures. For each image, the mean intensity of the whole scene is compared with the mean intensity of a rectangular area containing a person, and the average ratio between them is computed. Out of seven images, the lowest ratio is 1.149. We therefore set the threshold to a "conservative" value of 1.14 multiplied by the mean intensity of the image. An example of a binarized image is shown in Fig. 6b. Connected components labelling is then applied to the binarized image using the block-based decision tree (BBDT) implementation described in [28].

A transformation matrix from the thermal camera's to the RGB camera's reference frame is found and applied to bounding box coordinates from thermal images, such that all bounding boxes are expressed in the RGB camera's higher-resolution coordinate system.

Next, a confidence score is assigned to each of the bounding boxes. Based on statistical analysis of 850 bounding boxes from ground truth annotations, two linear models



(a) Height vs y-position. (b) Height vs width.

Fig. 4. Linear regression fit based on 850 bounding box annotations, used for calculating the position score and aspect ratio score for new bounding box measurements.

relating the height & y-position, and the width & height respectively are obtained using a regression fitting algorithm. The data points and regression line are shown in Fig. 4. A *position* score and an *aspect ratio* score ranging from 0 to 1 are calculated for each box based on its distance from each of the two regression lines:

$$score = \left(1 - \frac{(y - \hat{y})}{max_{\texttt{distance}}}\right)^2 \quad (1)$$

where $\hat{y}$ is the predicted value, y the corresponding measured value, and $max_{\texttt{distance}}$ acts as a normalization factor (as illustrated in Fig. 4). The total score is obtained by multiplying these two scores. Thus, the total score of a bounding box ranges between 0 and 1 and cannot be larger than either the position or the aspect ratio score. This bounding box score represents the likelihood that the bounding box corresponds to a person. To illustrate this, four arbitrary boxes are scored and shown in Fig. 5. The two red boxes are unlikely to correspond to people, while the green boxes have a high score due to their aspect ratio and position, and size.

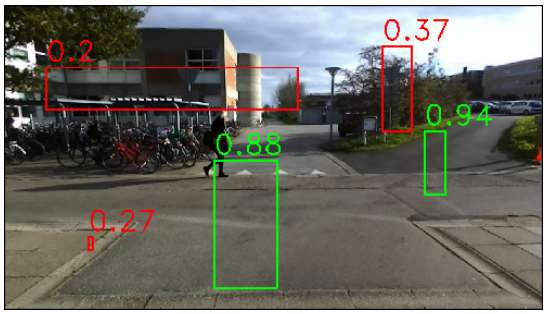


Fig. 5. Confidence score computed for arbitrary boxes, overlayed on an RGB image.

Since the cameras are mounted at a fixed angle and we assume navigation on a flat terrain (*i.e.* changes in elevation can be considered insignificant and ignored in our scenario), the horizon level can be estimated by visual inspection. We consider the top 30% of our images to be above the horizon level. We also set the minimum allowable height of a bounding box to be 10% of the image height, as this would correspond to people far in the distance. Two examples of the confidence scoring applied on the connected components detection are shown in Fig. 6.

(a) Original thermal image. (b) Detections overlayed on binarized image.

Fig. 6. Detected objects after applying adaptive thresholding, connected components detection and bounding box scoring on an thermal image. Only boxes with a score > 0.5 are shown.

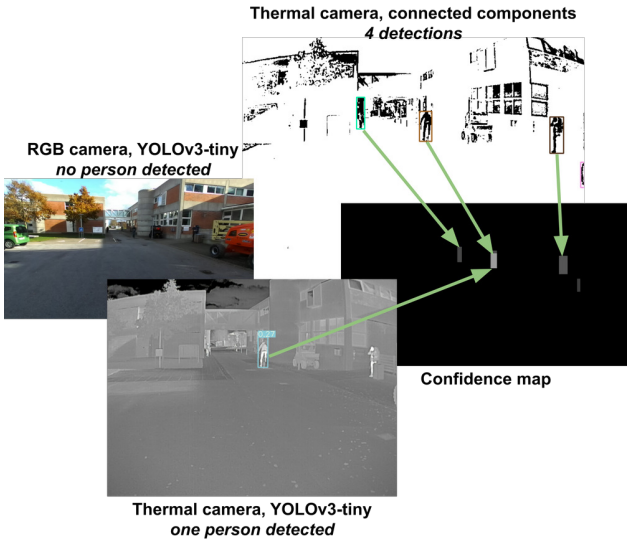


Fig. 7. Example showing a generated confidence map.

An example of a generated confidence map is shown in Fig. 7. A higher intensity indicates that the detection occurred in several modalities and/or has a high confidence score.

We use the same ground truth dataset as in Section IV-A. Binary images are generated from annotations. Each generated confidence map is thresholded into a binary image and compared to the corresponding binary ground truth image. The average recall and precision is then calculated for different threshold values. Unlike the previous experiment (Section IV-A), recall and precision are calculated pixel-wise. For comparison, a confidence map is generated for the 3 following cases:

- using the RGB camera only;
- combining RGB and thermal YOLOv3-tiny detections;
- running YOLOv3-tiny and the connected components algorithm on thermal, combined with YOLOv3-tiny on RGB.

Evaluation results are plotted in Fig. 8 and show that combining YOLOv3-tiny person detections from RGB and thermal images yields higher precision and recall than using RGB images alone. Combining YOLOv3-tiny with the connected components algorithm further improves recall with a small precision trade-off.

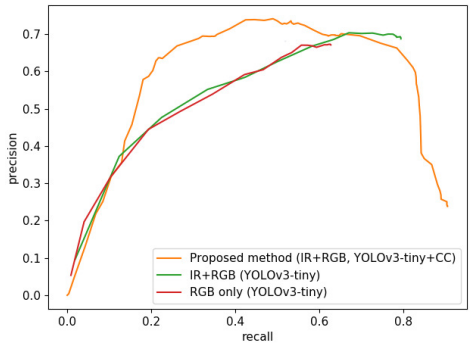
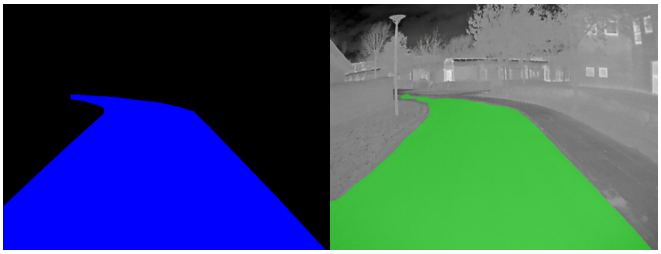


Fig. 8. Pixel-wise Precision-Recall curve of the proposed method compared to using RGB alone or YOLOv3-tiny on RGB and thermal images. Unlike Fig. 3, bounding boxes are scored with our proposed scoring method.

C. Road Segmentation with KittiSeg

A small dataset of around 350 images is used for training and testing. The images are captured with the experimental golf cart along different bicycle lanes, main marked roads and parking lots, providing a quite diverse dataset. Image augmentation was enabled during training to increase the size of the dataset. This includes random changes in brightness, contrast and hue, as well as random cropping. The rest of the training parameters were left as default. [9]

Table I shows the recorded performance. The values for both training and testing dataset are very similar, meaning that the network is not over-fitting to the training dataset. Fig. 9 shows an example result.



(a) Ground truth image (black: background class, blue: road). (b) KittiSeg output (estimated road area in green).

Fig. 9. Road segmentation output with KittiSeg trained on thermal dataset.

	Training dataset	*Testing dataset*
Average precision	94.2771	94.2186
Maximum F1	98.7994	97.4229

TABLE I. Performance evaluation of KittiSeg trained on thermal images.

V. Discussion

Person detection - Pedestrians sometimes fail to be detected in RGB captures while being detected in thermal captures or vice-versa. For instance, on a sunny day, significant glare in RGB images may completely obstruct people in the scene, and thus result in missing detections. However, since the thermal camera is insensitive to visible light,

glare does not appear in the thermal images and thus these pedestrians are successfully recovered in the confidence map. Furthermore, running the connected components algorithm in parallel with YOLOv3-tiny on thermal images allows many missing detections to be recovered. Even though the connected components algorithm also generates false positive detections, this is an acceptable trade-off considering the context of autonomous driving: it is much safer to mistake certain objects for people than to miss people altogether. Combining captures nevertheless presents several implementation challenges. Our multi-modal implementation assumes that thermal and RGB images are captured synchronously. In practice, the two cameras operate at different frame-rates, therefore detections of the same object may not perfectly overlap. Differences in the camera's FOVs also introduces "blind spots" in which the RGB camera is a single point of failure. Investigation of how the YOLOv3-tiny evaluation and proposed multi-modal scheme generalize to different datasets and experimental set-ups is left for future work.

Road Segmentation - Traversable road areas are precisely detected in a variety of different road configurations, without relying on road marks. Some inaccurate detections sometimes occur when there are foreign objects on the road (*e.g.* leafs, pothole). This could be limited by adding more of these examples to the training dataset.

VI. Conclusion

This paper confirms thermal imaging to be a very promising modality for both road and person detection, and shows that existing RGB image-based methods can be transferred to the thermal domain with little effort or complexity. Our proposed multi-modal scheme, indeed, achieves remarkable performance without the need to re-train the YOLOv3-tiny network which we use to generate an initial estimate of the bounding boxes, nor generate a fully labelled training dataset. We also show that the MultiNet deep learning architecture is able to achieve state-of-the-art road segmentation performance on high-resolution thermal images, using only a small labelled dataset for training the FCN-based decoder, and without needing to re-train the VGG-based encoder. The main limitation of the evaluation method is that our custom datasets did not include images captured in challenging weather conditions. However, given the nature of thermal cameras, we expect comparable performance in rain or fog, similarly to what has been already proved in [14].

References

[1] D. S. P. T. T. R. Shizhe Zang, Ming Ding and M. A. Kaafar, "The impact of adverse weather conditions on autonomous vehicles," *IEEE Vehicular Technology Magazine*, vol. 14, pp. 103–111, 2019.

[2] M. Engelcke, D. Rao, D. Zeng Wang, C. Hay Tong, and I. Posner, "Vote3Deep: Fast Object Detection in 3D Point Clouds Using Efficient Convolutional Neural Networks," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[3] L. Marchegiani and I. Posner, "Leveraging the urban soundscape: Auditory perception for smart vehicles," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[4] L. Marchegiani and P. Newman, "Listening for sirens: Locating and classifying acoustic alarms in city scenes," *arXiv preprint arXiv:1810.04989*, 2018.

[5] S. Chadwick, W. Maddern, and P. Newman, "Distant vehicle detection using radar and vision," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[6] R. Weston, S. Cen, P. Newman, and I. Posner, "Probably unknown: Deep inverse sensor modelling radar," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[7] F. C. V. Systems, "Metrological effects of fog & rain upon ir camera performance."

[8] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv*, 2018.

[9] M. Teichmann, M. Weber, J. M. Zöllner, R. Cipolla, and R. Urtasun, "Multinet: Real-time joint semantic reasoning for autonomous driving," *CoRR*, vol. abs/1612.07695, 2016.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv 1409.1556*, 2014.

[11] Z. Zhao, P. Zheng, S. Xu, and X. Wu, "Object detection with deep learning: A review," *CoRR*, vol. abs/1807.05511, 2018.

[12] H. M. N. R. S. S. C. E. M. . K. S. Ahmed, S., "Pedestrian and cyclist detection and intent estimation for autonomous vehicles: A survey." *Applied Sciences*, 2019.

[13] Fengliang Xu, Xia Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 1, pp. 63–71, 2005.

[14] M. Ivasic-Kos, M. Kristo, and M. Pobar, "Person detection in thermal videos using yolo," in *Proceedings of SAI Intelligent Systems Conference*. Springer, 2019, pp. 254–267.

[15] V. Ghenescu, E. Barnoviciu, S.-V. Carata, M. Ghenescu, R. Mihaescu, and M. Chindea, "Object recognition on long range thermal image using state of the art dnn," in *2018 Conference Grid, Cloud & High Performance Computing in Science (ROLCG)*. IEEE, 2018, pp. 1–4.

[16] A. L. M. Bertozzi, A. Broggi and M. D. Rose, "Infrared stereo vison-based pedestrian detection," *IEEE Proceedings. Intelligent Vehicles Symposium, 2005*, pp. 24–29, 2005.

[17] S. Kumar, T. K. Marks, and M. Jones, "Improving person tracking using an inexpensive thermal infrared sensor," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 217–224.

[18] V. Knyaz, "Multimodal data fusion for object recognition," in *Multimodal Sensing: Technologies and Applications*, vol. 11059, 2019, p. 110590P.

[19] M. Zilkha and A. B. Spanier, "Real-time cnn-based object detection and classification for outdoor surveillance images: daytime and thermal," in *Artificial Intelligence and Machine Learning in Defense Applications*, vol. 11169. International Society for Optics and Photonics, 2019, p. 1116902.

[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results."

[21] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[22] Q. Wang, J. Gao, and Y. Yuan, "Embedding structured contour and location prior in siamesed fully convolutional networks for road detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 230–241, 2017.

[23] G. A. Pelez, D. Bacara, A. de la Escalera, F. Garca, and C. Olaverri-Monreal, "Road detection with thermal cameras through 3d information," in *IEEE Intelligent Vehicles Symposium*, 2015, pp. 255–260.

[24] J. S. Yoon, K. Park, S. Hwang, N. Kim, Y. Choi, F. Rameau, and I. so Kweon, "Thermal-infrared based drivable region detection," in *IEEE Intelligent Vehicles Symposium*. IEEE, 2016, pp. 978–985.

[25] T. L. et al., "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014.

[26] J. Fritsch, T. Kuehnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *International Conference on Intelligent Transportation Systems (ITSC)*, 2013.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[28] C. Grana, D. Borghesani, and R. Cucchiara, "Optimized block-based connected components labeling with decision trees," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1596–1609, 2010.