ACCEPTED AUTHOR-SUBMITTED VERSION of the article:
A. Bañuls, A. Mandow, R. Vázquez-Martín, J. Morales and A. García-Cerezo, "Object Detection from Thermal Infrared and Visible Light Cameras in Search and Rescue Scenes," *2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, Abu Dhabi, United Arab Emirates, 2020, pp. 380-386, doi: 10.1109/SSRR50563.2020.9292593.

Object Detection from Thermal Infrared and Visible Light Cameras in Search and Rescue Scenes

Adrián Bañuls, Anthony Mandow, Ricardo Vázquez-Martín, Jesús Morales and Alfonso García-Cerezo[1]

***Abstract*— Visual object recognition is a fundamental challenge for reliable search and rescue (SAR) robots, where vision can be limited by lighting and other harsh environmental conditions in disaster sites. The goal of this paper is to explore the use of thermal and visible light images for automatic object detection in SAR scenes. With this purpose, we have used a new dataset consisting of pairs of thermal infrared (TIR) and visible (RGB) video sequences captured from an all-terrain vehicle moving through several realistic SAR exercises participated by actual first response organisations. Two instances of the open source YOLOv3 convolutional neural network (CNN) architecture are trained from annotated sets of RGB and TIR images, respectively. In particular, frames are labelled with four representative classes in SAR scenes comprising both persons (*civilian* and *first-responder*) and vehicles (*Civilian-car* and *response-vehicle*). Furthermore, we perform a comparative evaluation of these networks that can provide insight for future RGB/TIR fusion.**

I. INTRODUCTION

A thermal infrared (TIR) camera was employed in the first reported life save by a robot in 2013 [1]. At this point, infrared imagery is a decisive imaging modality not only for search and rescue (SAR) [2][3][4], but also for other robotic applications such as surveillance [5], military [6] and autonomous driving [7][8]. In comparison with visible light cameras (RGB), TIR cameras can be more robust against smoke, fog and lighting conditions [9]. Nevertheless, thermal radiation produces images lacking contrast and texture information [10], so the combination with other modalities can be advantageous for effective object identification [2].

Specially, synergy between thermal and visible images can be helpful to distinguish between rescuers and civilians, to identify survivors, or to recognise different kinds of vehicles. Besides, the RGB/TIR combination can produce an intuitive modality output for human rescuers [11] and can also benefit from recent deep learning tools for automatic object detection and scene understanding. Recently, state-of-the-art convolutional neural networks (CNN) models for object detection, such as single shot multi-box detector (SSD) [12] and YOLO [13] have achieved impressive real time performance with visible light images [7][14] in growing application domains [15][16].

A few works have extended the use of YOLO to thermal imaging, mainly with a focus on nighttime person detection. Thus, [17] addressed the problem of detecting distant persons and vehicles with small pixel sizes for surveillance and border control. In [5], a model trained on a TIR dataset clearly outperformed the original RGB-trained model for person detection under different weather conditions. Furthermore, the real-time qualities of YOLO were exploited in [18] for nighttime pedestrian detection from a moving TIR camera by applying a prior saliency stage. Other works have used CNNs to boast performance and accuracy when visible images are combined with other sensing modalities [19][20]. Thus, YOLO networks were used in [21] for semantic mapping from RGB images with depth information by incorporating a three-dimensional (3D) segmentation algorithm, and in [8] for combining frame- and event-driven images for pedestrian detection.

Fig. 1. All-terrain vehicle used to obtain the datasets (left), and the Oculus TI dual camera [22] (right).

Another indication of the growing interest on TIR image processing is the recent publication of different datasets, such as a far infrared (FIR) dataset for on-road pedestrian detection [23], a combination of visual and thermal data for person tracking in urban environments [24], and a multispectral dataset for day and nighttime driving [25]. Moreover, a specific dataset for SAR robotics has been constructed with multimodal (RGB, small field-of-view thermal, and depth) measurements of several indoor search scenarios as well as semi-synthetic images of victims [26].

In this work, our goal is to contribute to filling the gap in combined use of TIR and visible light images in the disaster robotics field. In particular, we explore automatic object detection in SAR scenes with TIR images and their complementarity with visible images. The major novel contributions of the paper are the following:

- We use a specific SAR dataset consisting of pairs of thermal and visible video sequences captured from an all-terrain vehicle (see Fig. 1) moving through several realistic SAR exercises performed by actual first re-

This work has been done in the framework of the TRUST-ROB project, funded by the Spanish Government (RTI2018-093421-B-I00).

[1]Universidad de Málaga, Robotics and Mechatronics Group, Andalucía Tech, 29071 Málaga, Spain. amandow@uma.es, ajgarcia@uma.es

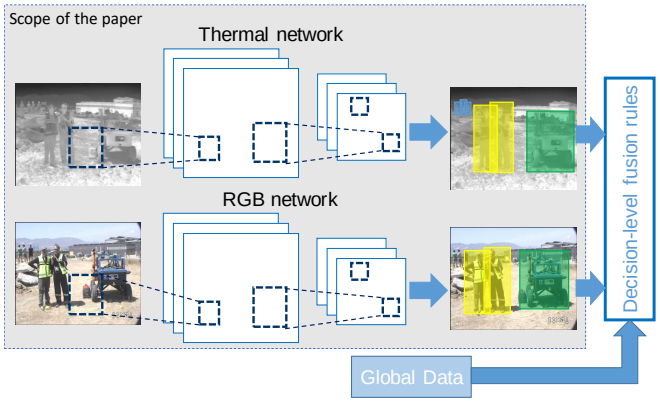

Fig. 2. RGB and TIR networks for representative SAR classes detection.

sponse teams. We plan to make this dataset publicly available [27].

- We evaluate the performance of the the open source YOLOv3 convolutional neural network (CNN) architecture for training both thermal infrared and visible light networks to detect four representative classes in SAR scenes comprising both persons (*civilian* and *first-responder*) and vehicles (*Civilian-car* and *response-vehicle*). With this purpose we have labeled selected frame pairs from the dataset with the corresponding classes.
- We analyze results from both networks to gain insight for RGB/TIR fusion for robust object detection.

This paper is organized as follows: section II presents the proposed system overview; section III describes the dataset and training; section IV discusses and analyses the results; and section V offers conclusions and ongoing work.

II. Overview

An overview for the proposed framework of object recognition in SAR scenes from RGB and TIR images is illustrated in Fig. 2. Two independent networks are trained from thermal and RGB images obtained from video sequences during realistic SAR exercises. The images have been annotated with representative classes in SAR scenes comprising both persons (*civilian* and *first-responder*) and vehicles (*Civilian-car* and *response-vehicle*). Quantitative and qualitative analysis and comparison of the results from both networks by considering context and global image parameters is used to gain insight for defining fusion criteria in a future work (e.g., decision-level fusion rules, as suggested in Fig. 2).

The thermal and RGB networks have been implemented with YOLO [13]. This is a state-of-the-art network architecture for multiple object recognition in full images consisting of a single network with convolutional layers that perform feature extraction as well as two fully connected layers for simultaneous prediction of bounding box locations and class probabilities. In particular, we have used YOLOv3 [28], which is an improvement over previous versions that performs feature extraction with 53 convolutional layers for feature extraction (Darknet-53). All in all, the network has 106 layers, 75 of which are convolutional.

III. Dataset, Data Model and Training

A. Dataset Acquisition

This work uses thermal and visible light images selected from a new multi-modal dataset [27] that we obtained during realistic SAR exercises carried out in Málaga (Spain) in June 2018 and in June 2019. These exercises were participated by first responders from emergency response organizations [29] in an experimental SAR site that can be seen in Fig. 3.

The video stream images in the dataset were recorded from an all-terrain vehicle equipped with a sensor suite that included an Oculus TI dual camera as well as a Velodyne HDL-32 lidar, an inertial unit and a differential global positioning system (see Fig. 1). The image datasets were captured as video sequences by an onboard Intel NUC NUC715BNKP i5-7260U/8GB/256GB SSD computer running Ubuntu 16.04 with ROS Kinetic.

The Oculus TI is a compact pan-and-tilt system that houses a TIR and RGB cameras. A microbolometer provides thermal images in the Longwave infrared (LWIR) wavelength band (8 to 14μm) and a horizontal field of view (FOV) of 44°. The thermal images have been recorded with white hot polarity. Besides, the system includes a visible light camera that has been used to obtain RGB images with a horizontal FOV of 57.8°. Both cameras capture images with a resolution of 704×576 and at a 25 Hz rate.

The images show first responders, civilian and rescue vehicles, rescue robots, civil observers, actors performing as victims and different SAR-related objects in an outdoor envirnment. The data was captured in different disaster simulations, including an earthquake and a man-made attack. In the images, rescuers from different organizations wear their corresponding working uniforms, which make them distinguishable from survivors and other civilians.

B. Data Model

The proposed SAR object detection data model considers $N_C = 4$ target classes, which are $\mathcal{C} = \{$*first-responder, civilian, response-vehicle, civilian-car*$\}$. Different colors have been assigned to represent the bounding boxes for these classes, as illustrated by labeled training frame shown in Fig. 4. These four classes are defined as follows:

- *First-responder*. A person with any kind of uniform and/or high visibility jackets is considered in this class. Recognizing the differences between all types of uniforms in the SAR exercise is out of the scope of this work.
- *Civilian*. This class corresponds to persons with ordinary clothing. Therefore, the *civilian* class comprises victims (represented by actors), unharmed survivors, civil observers, and journalists without distinctive uniform. In an actual disaster site, all of these could be potential survivors.
- *Response-vehicle*. The class includes all specialized vehicles such as ambulances, trucks, vans, robotic vehicles

Fig. 3. SAR exercise site.

Fig. 4. An example of the four classes in a labeled training image: the *first-responder* class is in yellow, *civilian* in blue, *response-vehicle* in green, and *civilian-car* in red.

and any other vehicle with distinctive signs of security forces and civil protection.

- *Civilian-car*. This class is defined for ordinary cars that lack of any recognizable sign of search and rescue organizations.

Our motivation for considering these four classes was the relevance of differentiating between survivors and rescuers in a disaster site. Furthermore, recognizing civilian cars can be useful to detect unseen victims.

C. Training and evaluation data

The RGB and TIR networks were trained using the parameters shown in Table I [30], where the number of epochs, the batch size and the warmup epochs have been adjusted empirically. We used a computer equipped with an AMD Ryzen 7 2700X 4.3GHz CPU and a NVidia GeForce RTX 2060 Ventus OC 6GB GDDR6 GPU running Ubuntu 18.04 operating system. Both networks were trained using transfer learning, but no pre-trained networks for similar classes are available. In order to take advantage of a pre-trained network, training for the new classes has been done by changing only the weights of the fully-connected layers, using fine-tuning to perform the classification task of the new SAR classes [31].

TABLE I
TRAINING PARAMETERS CONFIGURATION

Training parameters	Value
Number of epochs	65
Batch size	2
Moving average decay	0.9995
Initial learning rate	1e-4
Final learning rate	1e-6
Warm-up epochs	2

In this work, a total of 2288 frames containing objects of interest have been selected from the full SAR dataset and manually labeled. The selection criterion was to have images with different number and types of persons and vehicles, points of view, scales and backgrounds.

This complete set of labeled images was split into three different groups, as follows:

- Training (70%): Images used for network training.
- Validation (15%): Images used to fit the model parameters during training, avoiding overfitting.
- Test (15%): Labeled images for evaluating the network after training is complete.

Furthermore, data augmentation is especially useful for application fields, such as SAR, where obtaining measurements from actual and even simulated disaster sites can be very difficult. Thus, we have used the data augmentation features provided by the YOLOV3 framework used in this work [32]. These include random translations, rotations and scale changes up to 20% of the original image. Besides, an early stopping criterion has been used during training, in order to prevent overfitting.

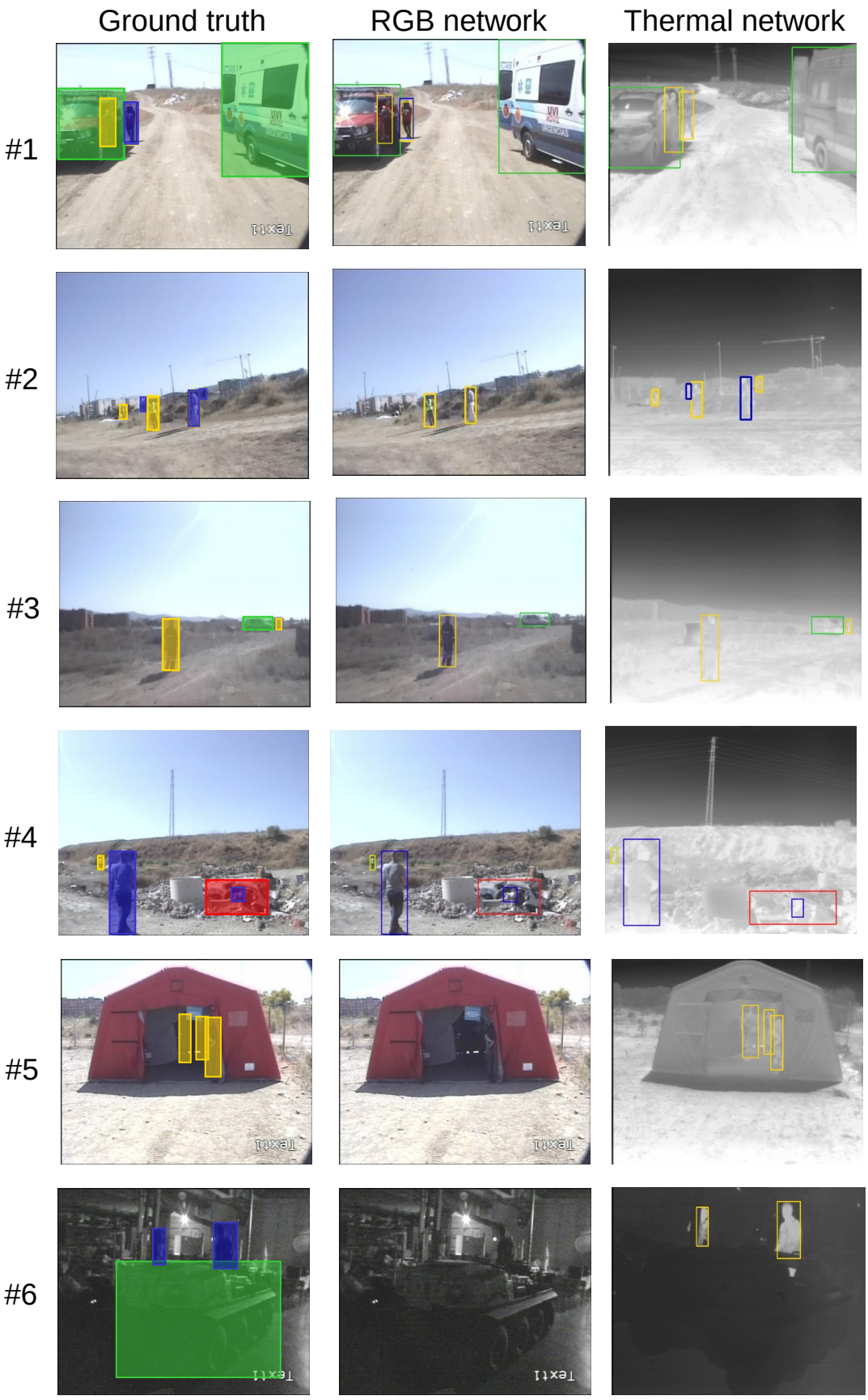

Fig. 5. Six representative examples of object detection with the RGB and the thermal networks. Ground truth is illustrated on an RGB image.

TABLE II

PERFORMANCE RESULTS OF RGB NETWORK FOR IoU THRESHOLD = 0.5

Class	AP (%)	Actual objects	TP	FP	FN	P (%)	R (%)
First-responder	**87.91**	769	700	192	69	78.47	91.02
Civilian	**77.83**	220	179	39	41	82.11	81.36
Response-vehicle	**94.28**	342	323	27	19	92.28	94.44
Civilian-car	**79.62**	32	29	7	3	80.55	90.62

TABLE III

PERFORMANCE RESULTS OF TIR NETWORK FOR IoU THRESHOLD = 0.5

Class	AP (%)	Actual objects	TP	FP	FN	P (%)	R (%)
First-responder	**89.33**	703	643	187	60	77.47	91.46
Civilian	**85.54**	247	216	48	31	81.82	87.45
Response-vehicle	**91.38**	398	373	62	25	85.75	93.72
Civilian-car	**65.02**	78	55	11	23	83.33	70.51

IV. RESULTS

This section evaluates and compares results from the RGB and the TIR image networks for the four representative classes considered in this work.

A. Qualitative Results

Resulting bounding boxes for six representative examples of RGB/TIR image pairs are presented in Fig. 5. The figure also shows the corresponding ground truth, which has been defined by a human expert after observing both the RGB and thermal frames.

The major detection errors appreciated in the thermal network consist on the confusion between classes corresponding to persons. Thus, the thermal network missclassifies civilians as *first-responder* in cases #1 and #6. These errors can be due to the limited contrast and texture information offered by TIR images, where there is only one single channel, whose content correspond to the distribution of temperature. Conversely, the colour channels in RGB images provide complementary information about the patterns of the person clothes, which results in a correct recognition of the *first-responder* in case #1.

Another relevant issue is the detection of distant objects. In principle, persons can be detected with the Oculus TI camera up to a distance of 40 m. Even a human expert can find difficulty in spotting distant objects with low RGB resolution, not to mention distinguishing between the two classes of persons. Thus, the RGB network fails to detect far persons in cases #2 and #3, which are successfully detected by the thermal network due to the characteristic temperature pattern of the human body. Interestingly, all objects in case #4, where the farthest *first-responder* person wears high-visibility clothing, are correctly detected by both networks.

Finally, cases #5 and #6 correspond to challenging situations with extreme visibility. In case #5 there are persons inside a tent that are not detected in the RGB image. Case #6 shows a dark scene recorded in our lab specially for this work. Again, none of the classes is detected in the RGB image, but the TIR network succeeds in identifying two persons, even if they are classified as *first-responder* instead of *civilian*. On the other hand, the vehicle is not detected in the thermal image, which indicates that it has been parked for some time and the motor temperature cannot be distinguished from the rest of the scene.

B. Quantitative Results

The two different networks for RGB and TIR images are evaluated using the standard performance metrics that describe the accuracy and quality of object detection:

- Mean Average Precision (mAP) is the average precision at different recall values; i.e., the area under the precision-recall curve, where precision is $P = TP/(TP + FP)$ and recall is $R = TP/(TP + FN)$. TP, FP and FN stand for true positive, false positive and false negative, respectively. The Average Precision for a single class is denoted AP.
- Intersection Over Union (IoU) measures how predicted bounding boxes fit the location of an object. Thus, IoU is the relation between the area of intersection and the union of predicted and real bounding boxes.

Tables II and III present the results obtained with the RGB and the thermal networks, respectively. The average precision for both person classes with the TIR network is greater than that for the same classes with RGB network. This can be explained by the temperature difference between the human body and the surrounding environment, which favors TIR detection. This difference in AP is greater for persons of the *civilian* class, who do not wear high visibility uniforms. For vehicle classes, where temperature difference is not indicative, the use of visible spectrum cameras offers better results, especially for the *Civilian-car* class.

Performance results for the RGB and thermal networks are summarized in Table IV. The global mAP and IoU are very similar for both networks, with a better IoU value for RGB. One interesting aspect to consider is that the TIR network was pre-trained with RGB images from the COCO dataset,

TABLE IV

RGB AND TIR NETWORKS RESULTS FOR FOUR CLASSES.

Network	mAP (%)	IoU	First-responder AP (%)	Civilian AP (%)	Response-vehicle AP (%)	Civilian-car AP (%)
RGB	84.91	65.15	87.91	77.83	**94.28**	**79.62**
TIR	82.82	60.57	**89.33**	**85.54**	91.38	65.02

TABLE V

RGB AND TIR NETWORKS RESULTS FOR TWO CLASSES: PERSON AND VEHICLE.

Network	mAP (%)	IoU	Person AP (%)	Vehicle AP (%)
RGB	91.44	60.76	87.40	**95.49**
TIR	90.59	57.93	**88.37**	92.81

TABLE VI

PRECISION (mAP) FOR DIFFERENT IoU THRESHOLD VALUES.

IoU	RGB mAP (%)	TIR mAP (%)
0.5	84.91	82.82
0.55	82.69	80.60
0.6	78.03	75.29
0.65	71.84	63.50
0.7	62.38	49.70
0.75	50.37	34.35
0.8	32.46	20.07
0.85	16.15	9.91
0.9	3.33	1.94
0.95	0.04	0.04
Average	**48.22**	**41.82**

and the fine-tuning training of the new SAR representative classes were performed using TIR images. In spite of this, the resulting similar metrics for both networks could be explained by the similarity of the image features filtered in the first layers of the network architecture.

Besides, for the sake of performance comparison, we have trained simpler versions of the RGB and TIR networks with just two conventional classes $\mathcal{C}_s = \{person,\ vehicle\}$, with the results shown in Table V. In this case, a mAP accuracy of 91.44% for the RGB network and 90.59% for the thermal network have been obtained. These results offer only a slight improvement over the four classes case in Table IV, in spite of the greater difficulty in differentiating between two classes of vehicles and two classes of persons in the latter. Furthermore, the IoU values are better for both four-class networks. This comparison indicates a good performance for the networks trained with four SAR classes in comparison to two standard classes. As for the evaluation of location precision for bounding boxes, the two networks have been tested using different IoU thresholds and computing the average mAP, as in the COCO detection challenge [33]. The results for an IoU that range from 0.5 up to 0.95 are shown in table VI with the total average mAP at the bottom. Although YOLOv3 reaches 33% in the average mAP for 80 classes [28], the results reveal a precise location of the bounding boxes for the four SAR classes in the RGB (48.22%) and TIR (41.82%) networks.

V. CONCLUSIONS

In this work, we have offered a preliminary analysis of the use of thermal and visible range images for automatic object detection in SAR scenes. With this purpose, we have obtained a custom dataset consisting of pairs of thermal and visible video sequences captured from an all-terrain vehicle moving through several realistic SAR exercises participated by actual first response organizations. Two instances of the open source YOLOv3 convolutional neural network (CNN) architecture have been trained from annotated sets of RGB and TIR images, respectively. In particular, frames have been labeled with four representative classes in SAR scenes corresponding to both persons (*civilian* and *first-responder*) and vehicles (*Civilian-car* and *response-vehicle*). To the best of our knowledge, this is the first work that addresses automatic CNN detection of these specific SAR classes.

Qualitative results have shown a good performance of both the RGB and the thermal networks in the detection and identification of the four SAR classes. These results indicate that the YOLOv3 architecture could be trained for a larger number of classes in the SAR domain, such as identifying victims. Moreover, results have indicated that the network for TIR images can benefit from transfer learning from RGB networks. Besides, the strengths and limitations of both modalities have been identified by the quantitative and the qualitative analysis, which has confirmed the potential synergies of both modalities.

The insight gained from this work can be considered for future development of an RGB/TIR fusion mechanism for robust object detection. Future work will be needed to evaluate different fusion strategies, which could include a data fusion approach, e.g. combining infra-red and color images at the input, and decision-level fusion, where sensor fields of view do not need to overlap completely [34]. Finally, we plan to make the datasets used in this work publicly available in the near future.

ACKNOWLEDGMENTS

We are grateful to the Chair for Safety, Emergencies and Disasters at UMA (Cátedra de Seguridad, Emergencias y Catástrofes, Universidad de Málaga) and in particular Prof.

Jesús Miranda-Páez for organizing the exercises and allowing us to capture the dataset.

REFERENCES

[1] R. R. Murphy, S. Tadokoro, and A. Kleiner, *Springer Handbook of Robotics*. Springer, Cham, 2016, ch. Search and Rescue Robotics.

[2] P. Rudol and P. Doherty, "Human body detection and geolocalization for UAV search and rescue missions using color and thermal imagery," in *IEEE Aerospace Conference*, 2008, pp. 1–8.

[3] S. Kim, S. Jun, and J. Park, "Thermal stereo system for visible range extension of disaster robot," in *IEEE International Symposium on Safety, Security, and Rescue Robotics*, 2018, pp. 1–2.

[4] S. P. Kleinschmidt and B. Wagner, "Visual multimodal odometry: Robust visual odometry in harsh environments," in *IEEE International Symposium on Safety, Security, and Rescue Robotics*, 2018, pp. 1–8.

[5] M. Ivăsić-Kos, M. Krišto, and M. Pobar, "Human detection in thermal imaging using YOLO," *ACM International Conference Proceeding Series: International Conference on Computer and Technology Applications*, vol. Part F148262, pp. 20–24, 2019.

[6] A. D'Acremont, R. Fablet, A. Baussard, and G. Quin, "CNN-based target recognition and identification for infrared imaging in defense systems," *Sensors*, vol. 19, no. 9, 2019.

[7] D. Chaves, S. Saikia, L. Fernández-Robles, E. Alegre, and M. Trujillo, "A systematic review on object localisation methods in images [Una revisión sistemática de métodos para localizar automáticamente objetos en imágenes]," *Revista Iberoamericana de Automática e Informática Industrial*, vol. 15, no. 3, pp. 231–242, 2018.

[8] Z. Jiang, P. Xia, K. Huang, W. Stechele, G. Chen, Z. Bing, and A. Knoll, "Mixed frame-/event-driven fast pedestrian detection," in *International Conference on Robotics and Automation*, 2019, pp. 8332–8338.

[9] Y. S. Shin and A. Kim, "Sparse depth enhanced direct thermal-infrared slam beyond the visible spectrum," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2918–2925, 2019.

[10] M. Leingartner, J. Maurer, G. Steinbauer, and A. Ferrein, "Evaluation of sensors and mapping approaches for disasters in tunnels," in *IEEE International Symposium on Safety, Security, and Rescue Robotics*, 2013, pp. 1–7.

[11] L. Zalud and P. Kocmanova, "Fusion of thermal imaging and CCD camera-based data for stereovision visual telepresence," in *IEEE International Symposium on Safety, Security, and Rescue Robotics*, 2013, pp. 1–6.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European Conference on Computer Vision*. Springer International Publishing, 2016, pp. 21–37.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[14] Z. Zhao, P. Zheng, S. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. Early access, pp. 1–21, 2019.

[15] M. Wang, X. Long, P. Chang, and T. Padlr, "Autonomous robot navigation with rich information mapping in nuclear storage environments," in *IEEE International Symposium on Safety, Security, and Rescue Robotics*, 2018, pp. 1–6.

[16] M. Fulton, J. Hong, M. J. Islam, and J. Sattar, "Robotic detection of marine litter using deep visual detection models," in *International Conference on Robotics and Automation*, 2019, pp. 5752–5758.

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

[18] D. Heo, E. Lee, and B. Ko, "Pedestrian detection at night using deep neural networks and saliency maps," *Journal of Imaging Science and Technology*, vol. 61, no. 6, 2017.

[19] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *British Machine Vision Conference*, 2016, pp. 73.1–73.13. [Online]. Available: https://dx.doi.org/10.5244/C.30.73

[20] Y. Sun, W. Zuo, and M. Liu, "RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.

[21] Y. Nakajima and H. Saito, "Efficient object-oriented semantic mapping with object detector," *IEEE Access*, vol. 7, pp. 3206–3213, 2019.

[22] Silent Sentinel, "Oculus Scout datasheet," 2018, accessed on 2019-7-01. [Online]. Available: https://silentsentinel.com/product/oculus-scout/

[23] Z. Xu, J. Zhuang, Q. Liu, J. Zhou, and S. Peng, "Benchmarking a large-scale FIR dataset for on-road pedestrian detection," *Infrared Physics and Technology*, vol. 96, pp. 199–208, 2019.

[24] E. Gebhardt and M. Wolf, "CAMEL dataset for visual and thermal infrared multiple object detection and tracking," in *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2018, pp. 1–6.

[25] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, "KAIST multi-spectral day/night data set for autonomous and assisted driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 934–948, 2018.

[26] T. Petříček, V. Šalanský, K. Zimmermann, and T. Svoboda, "Simultaneous exploration and segmentation for search and rescue," *Journal of Field Robotics*, vol. 36, no. 4, pp. 696–709, 2019.

[27] J. Morales, R. Vázquez-Martín, A. Mandow, D. Morilla-Cabello, and A. García-Cerezo, "The UMA-SAR dataset: Multimodal data collection from a ground vehicle in disaster response training exercises," *submitted for publication*.

[28] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv e-prints*, vol. abs/1804.02767, 2018. [Online]. Available: http://arxiv.org/abs/1804.02767

[29] J. J. Fernández-Lozano, A. Mandow, M. Martín-Guzman, J. Martín-Avila, J. Serón, J. L. Martínez, J. A. Gomez-Ruiz, C. Socarrás-Bertiz, J. Miranda-Paez, and A. García-Cerezo, "Integration of a canine agent in a wireless sensor network for information gathering in search and rescue missions," in *IEEE International Conference on Intelligent Robots and Systems*, 2018, pp. 5685–5690.

[30] Alexey, "Yolo: How to train (to detect your custom objects)," 2019. [Online]. Available: https://github.com/AlexeyAB/darknet

[31] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, ch. Optimization for Training Deep Models, http://www.deeplearningbook.org.

[32] YunYang1994, "Tensorflow-yolov3," 2019. [Online]. Available: https://github.com/YunYang1994/tensorflow-yolov3

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*. Springer International Publishing, 2014, pp. 740–755.

[34] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Information Fusion*, vol. 45, pp. 153 – 178, 2019.