Person Detection in Thermal Videos Using YOLO

Ivašić-Kos, Marina; Krišto, Mate; Pobar, Miran

Person Detection in Thermal Videos using YOLO

Marina Ivasic-Kos, Mate Kristo, Miran Pobar

Department of Informatics, University of Rijeka, 51000 Rijeka, Croatia
marinai@uniri.hr, matekrishto@gmail.com, mpobar@uniri.hr

Abstract. In this paper, the task of automatic person detection in thermal images using convolutional neural network-based models originally intended for detection in RGB images is investigated. The performance of the standard YOLOv3 model is compared with a custom trained model on a dataset of thermal images extracted from videos recorded at night in clear weather, rain and fog, at different ranges and with different types of movement – running, walking and sneaking. The experiments show excellent results in terms of average precision for all tested scenarios, and a significant improvement of performance for person detection in thermal imaging with a modest training set.

Keywords: Thermal imaging, Object Detector, Convolutional Neural Networks, YOLO, person detection

1 Introduction

The purpose of object detection is to classify objects in images and mark their exact position. Many successful machine learning algorithms have been developed for detecting objects such as human face [1] or human figure [2] in RGB images.

Thermal cameras are now ubiquitous in video surveillance systems that take care of the safety of people and objects in urban areas, at state borders, in guarded areas, etc. Because of global terrorist threats and illegal migration, concerns about the safety of innocent people have been intensified. To prevent unwanted events and protect people and their property, investment in security systems has reached record levels trying to utilize all available technological achievements and develop sophisticated systems.

Thermal cameras are important for surveillance and security because they can be used in unfavorable weather conditions when ordinary RGB cameras cannot be used or give poor results, such as night and full darkness (Fig. 1.) or rain and fog.

Fig. 1. Night vision vs. thermal imaging showing that tree camouflage cannot hide a person from a thermal camera[1].

Today, the best object detection results in RGB images are achieved by models based on convolutional neural networks (CNN). The popularity of convolution neural networks and deep learning began with the great success of AlexNet for the image recognition task in ImageNet Challenge in 2012 [3]. Since then, many successful CNN architectures for object detection have been developed, such as R-CNN [4], SSD [5], R-CNN mask [6], R-FCN [7] and YOLO [8].

Several methods have been proposed specifically for person detection in thermal videos. Papers [9, 10] use the HOG features that are commonly used in the task of pedestrian detection in RGB images as the fundamental feature for detecting persons in thermal images. In [9] the HOG features are extracted from previously detected regions of interest, and an AdaBoost classifier is used to make the final decision. The method presented in [10] in addition to region proposal and classification based on HOG tracks the detected persons across several frames using template matching to suppress false positive detections. Background subtraction methods are used in [11] as a first stage to identify regions with movement in images that are potentially containing moving pedestrians. These regions are then used as regions of interest from which features are extracted and used for classification with the AdaBoost classifier in the second stage.

The goal of this paper is to investigate whether the CNN method could be successful for the task of detecting people in images and videos obtained with a thermal camera. Due to the differences in visual and thermal image features, the aim is to explore how well deep learning methods successful for object detection in optical images will be successful with thermal imaging.

For the detection task, the YOLOv3 network [12] will be used, as it achieves object detection results in RGB images at the state-of-the-art level for different detection tasks [13, 14].

In the next section, the basic information about thermal imagery is provided. The detection pipeline of YOLO object detector is given in Section 3. Dataset and object detection experiments are described in Section 4. The evaluation results are presented and discussed in Section 5. The work ends with the conclusion and direction for future research.

[1] https://www.youtube.com/watch?v=rAvnMYqj2c0

2 Thermal imagery

Thermal cameras record the heat emitted by the subject being monitored and form an image by infrared (IR) radiation. IR radiation is the electromagnetic radiation emitted by the relative heat generated or reflected by an object, and therefore IR imaging is termed thermal imaging and an IR image thermogram. The wavelengths of IR range from 400 nm to 1400 nm, Fig. 2. IR wavelengths are longer than those of visible light, so IR is invisible to humans [15].

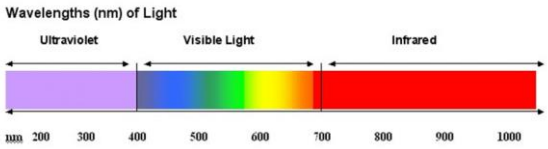


Fig. 2. Wavelengths of light in nm[2]

As thermal sensors form an image of the environment or object solely according to the detected amount of heat energy of the recorded object, unlike visible sensors, they are insensitive to light conditions and changes in light and robust to different weather conditions and a wide range of light variations [16, 17].

However, thermal sensors provide much less detail than visible light cameras, because instead of the information that gives color in the visible spectrum, it only has the temperature ranges in the thermograms, Fig. 3.

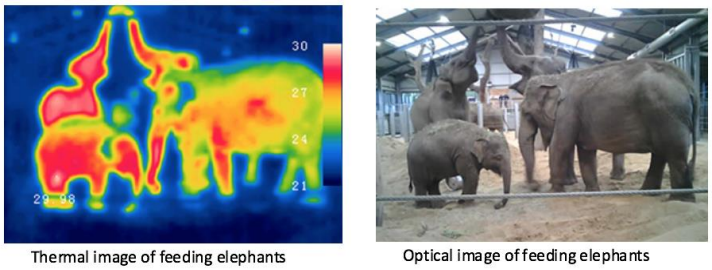


Fig. 3. IR sensors provide much fewer details than the optical sensor of visible light[3]

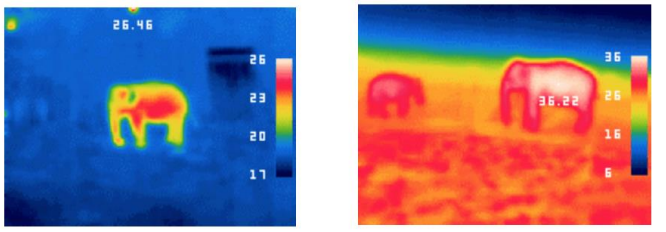
Unlike optical sensors, IR sensors are very sensitive to changes in ambient temperature. The heat that some object emits is not constant but depends on the internal state of the object and the ambient temperature, Fig. 4. For example, the human body temperature makes the core temperature (the temperature of the abdomen, the thoracic and the cranial cavities) and the skin temperature (skin and subcutaneous tissue). The thermoregulatory system of the human body maintains a constant core temperature

[2] https://www.scienceoflight.org/ir-light/

[3] https://www.wildlabs.net/resources/case-studies/hwc-tech-challenge-update-comparing-thermopile-and-microbolometer-thermal

(about 37° C during rest) by establishing a balance between the production of the heat of metabolism and the release of heat into the environment. But during running or intense exercise, metabolic heat production can be increased by 10 to 20 times in relation to heat production at rest. Also, skin temperature is more affected by blood flow to the skin and by environmental conditions [18].

The color on the temperature scale does not always match the same temperature for all images, but the lowest color (dark blue) corresponds to the coolest part of the image and the brightest (white) hottest part of the image. E.g., in Fig. 4. (left) the white color corresponds to 26 degrees C and to the Fig. 4. (right) to 36 degrees C.



Elephant indoors 15 meters away from sensor Elephant outdoors 12 meters away from sensor

Fig. 4. IR sensors are very sensitive to changes in ambient temperature[4]

3 YOLO Detector

The original YOLO object detector (YOLOv1) [8] uses a convolutional neural network to simultaneously predict the bounding boxes of multiple objects in an image and associate them with the confidence of the class they belong to.

The YOLOv1 architecture consists of 24 convolutional layers and two fully connected layers. The convolutional layers perform feature extraction while the fully connected layers predict locations of bounding boxes and their probabilities. The input image is first divided into an S x S grid. Two bounding boxes and corresponding class confidences are associated with each grid cell, so at most two objects can be detected within a cell. If an object occupies more than one cell, the center of all cells is selected as the holder of prediction for that object. The bounding box that doesn't hold any object has a confidence value assigned to zero when training the model. The confidence value of a bounding box that contains an object or has an intersection with an object corresponds to the intersection-over-union (IoU) score of the bounding box and the ground truth box.

[4] https://www.wildlabs.net/resources/case-studies/hwc-tech-challenge-update-comparing-thermopile-and-microbolometer-thermal

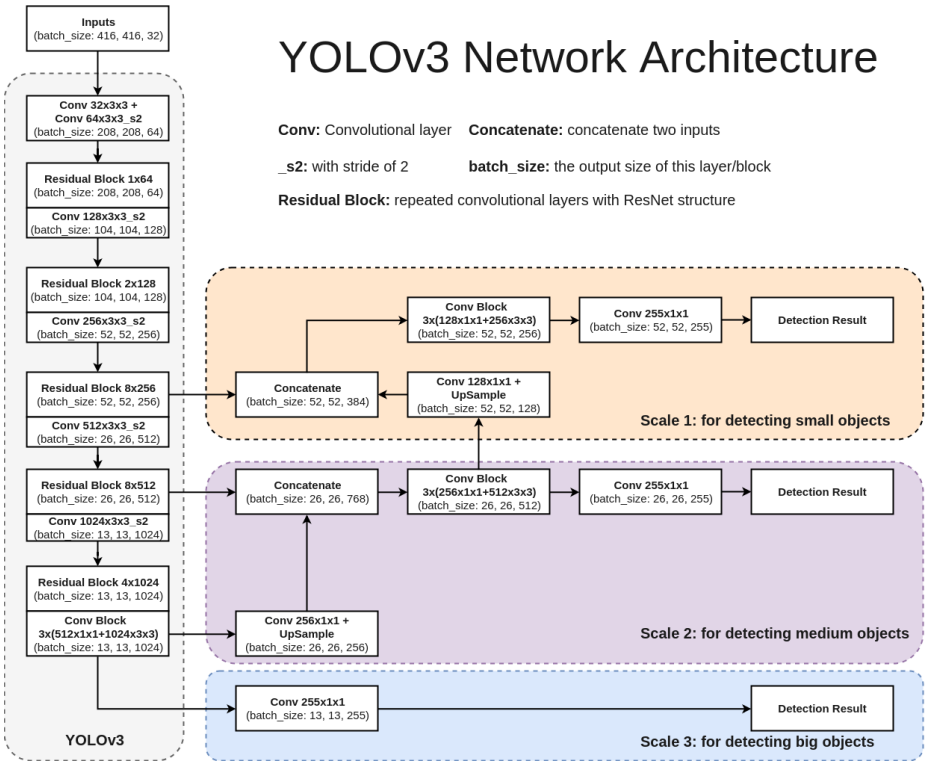


Fig. 5. YoloV3 network architecture[5]

In the version 2 of the Yolo detector (YoloV2) [19], five of the convolutional layers of the original model are replaced with max-pooling layers, and the way in which the bounding box suggestions are generated was changed. Instead of predicting coordinates of the bounding box for each cell, predefined anchor boxes are used. Anchor boxes are defined in a training set, using k-means clustering of ground truth bounding boxes where boxes translations are relative to a grid cell.

In the third version of the Yolo detector (YoloV3) [12, 20], instead of the 19-layer feature extraction network, a much deeper network is used, consisting of 53 layers of 3x3 and 1x1 filters with skip connections and without fully-connected layers, Fig. 5. Instead of pooling layers, a convolutional layer with stride 2 is used to down-sample the feature map and pass size-invariant feature forward. Also, the bounding box prediction was refined, using features at 3 different scales to make 3 sets of box predictions for each location. The classification method has also been changed, so now multilabel classification is used. An object may, in that case, belong to more than one class simultaneously, which is achieved by replacing the soft-max with logistic regression.

[5] https://www.cyberailab.com/home/a-closer-look-at-yolov3

4 Experiment Setup

In the experiment, the effectiveness of the YoloV3 detector in surveillance applications when using IR cameras is examined. The task is to detect a person on thermal images collected at different weather conditions.

As the baseline model, the original YOLOv3 network (referred to as bY) trained on the COCO RGB [21] image datasets is used, which proved successful in detecting different object classes [22] in RGB images. Although the thermal image differs significantly in color and detail from RGB images, reasonably good results are expected with the bY model for IR imaging for two reasons. First, it can be assumed that some convolutional layers trained on RGB images will extract shape features that will be similar in thermal imaging, so those features trained on RGB images will also be useful for thermal imaging. Then, the experiment [8] showed that the effectiveness of the YOLO detector, when applied to the task of person detection in art images, was less degraded than with other detectors, even though art images were not used for training the model.

The original model was further trained for person class on thermal images from a custom data set (called tY), and the results of both models are compared on thermal images taken in different weather conditions.

A. **Dataset**

The data used in the experiment were collected by recording people during the night in different weather conditions and at different distances from the camera. FLIR ThermalCAM P10 thermal focal surface (FPA) camera with uncooled bolometer that covers the spectral range between 7.5 and 13 μm (LWIR) was used. Telephoto lens series FLIR P / B with 7°x5.3V FOV and 3.5x magnification was used along with cameras and lenses with 24°x18°/0.3 m field of view as basic equipment.

Five men and two women were recorded in the winter period (in February 2017) moving in a normal and hunched position, either with a normal walking speed or running, in several lenses and range configurations [23].

The recording was done in different weather conditions, with appropriate settings. In the clear weather, the distance of people from the camera was 110 m (baseline) or 165 m. In the heavy fog, with a minimum visibility of up to 5 m, the shooting was done with people moving at less than 30 m from the camera and at 50 m from the camera. In the fog conditions, it was not possible to use standard lenses or to record at larger distances, so only telephoto lenses were used. In the heavy rain, people were moving at 30 m, 70 m, 110 m, 140 m, 170 m, 180 m and 215 m from the camera.

After the videos had been captured, individual frames were extracted from the video to create a data set. About 15,000 images in the set were shot with a telephoto lens in clear, fogy or rainy weather conditions and about 6,000 images were shot with a standard lens in clear weather [24].

The images were manually annotated using the VGG Image Annotator (VIA) [25]. In this experiment, about 1,000 images for each weather condition were used for training.

B. Evaluation measure

The mean average precision (mAP) criteria like the one used in the PASCAL VOC 2012 competition [26] is used to evaluate the performance of the models. To get the mAP value, the mean of Average Precision (AP) values of all classes is calculated, but in this experiment, only the *Person* class is considered.

The detection results are compared with the ground truth. A detection is counted as a true positive if the intersection-over-union (IoU) score of the detected bounding box and the corresponding ground truth bounding box is greater than or equal to 50%. An example of positive and negative object detection with respect to intersection-over-union score in the case of person detection is shown in Fig. 6.

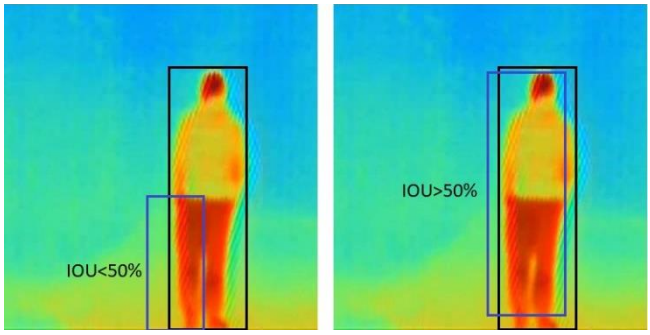


Fig. 6. Negative (left) and positive (right) representation of IoU criteria

When the same object is detected multiple times, only one detection is counted as a true positive.

5 Evaluation Results and Discussion

A precision versus recall curve for the desired class (here *Person*) is produced by varying the confidence threshold of the detector. The AP score is then the area underneath the precision versus recall curve. Fig. 7 presents the AP score for the original YOLO model, bY, that was not trained on thermal images, while Fig. 8 corresponds to the AP score for model tY that was additionally trained for the class *Person* on the custom dataset. Additional training improved the baseline results, so the AP score of 97.93% achieved with the model tY significantly exceeds the AP score of 19.63% achieved by bY.

For example, the bY model achieves 100% precision with the recall of 15.5%, while the model tY achieves the same precision with a much higher recall of approx. 50%, meaning that the tY model detects a lot more people present in the images with the same precision.

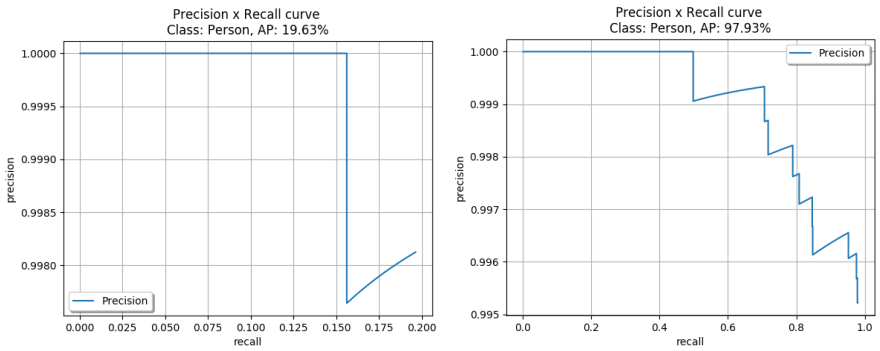
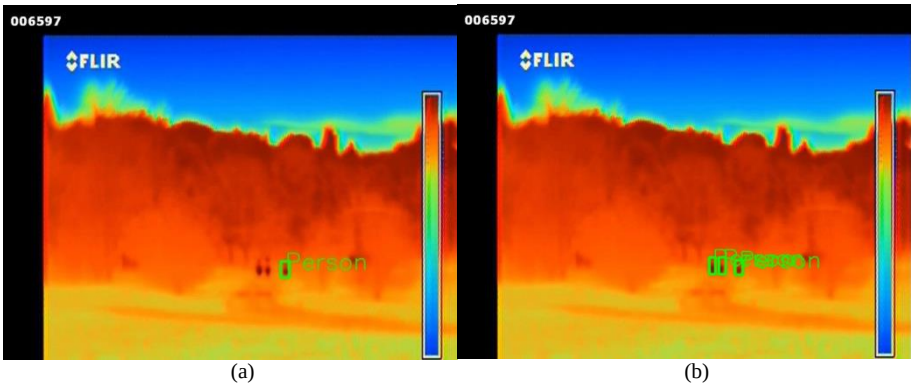


Fig. 7. AP score and precision versus recall curve for baseline YOLO model, bY (left), and for custom trained YOLO model, tY (right)

Fig. 8 shows an example (Image no. 6597) of the detection results of both bY and tY models on clear weather, recorded with a standard lens.



(a) (b)

Fig. 8. Results of person detection on images recorded with a normal lens in clear weather condition, distance 110-160m, using the bY (a) opposite to tY model (b).

The tY model has detected all three persons in the image (true positive detection, TP), even though they are about 150 m away from the camera and take up only a few pixels. The bY Model managed to detect only one out of three persons present in the image, (two false negative detections, FN). This is also an unexpectedly good result because the silhouettes of persons are tiny and the temperature difference between persons and vegetation is not so large, so it is not easy to notice people at that distance even for a security guard.

Figs. 9 to 11 show the AP score for the different weather conditions. The tY model achieves an AP score of 97.85% for clear and foggy weather, while for the rain the AP score is even better, at 98.8%. Looking at the results fixed at 100% precision, the tY model achieves a recall score of 35% in the clear weather, 50% in the foggy weather and 75% in the rain.

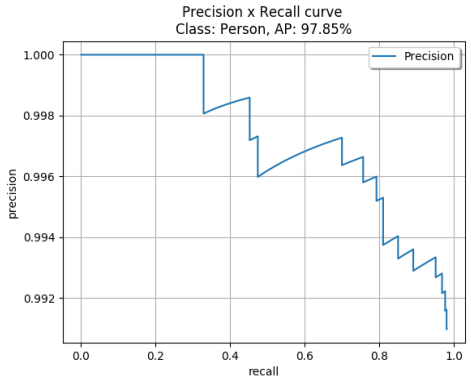


Fig. 9. Precision versus recall curve for clear weather images

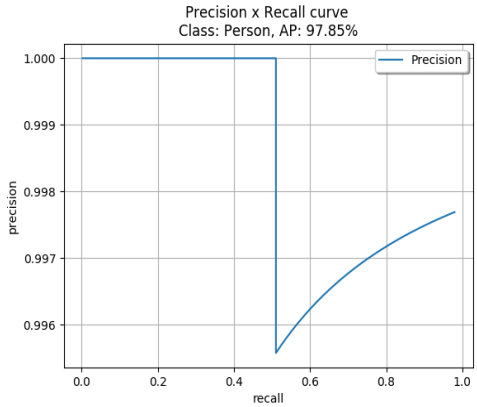


Fig. 10. Precision versus recall curve for images taken in foggy conditions.

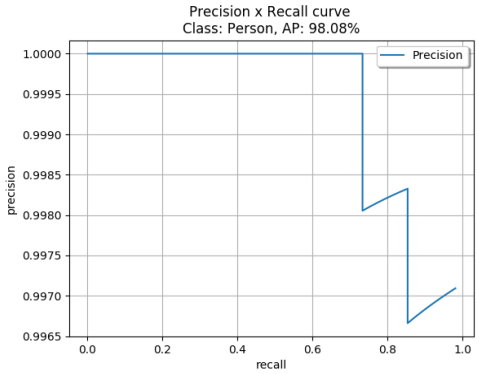
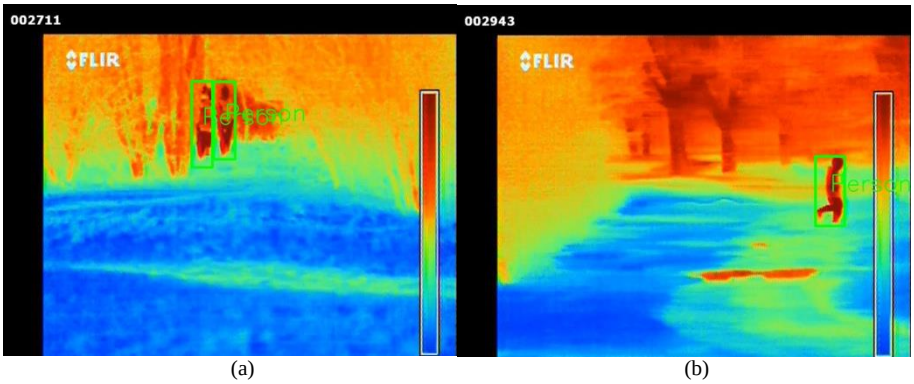


Fig. 11. Precision versus recall curve for images taken in the rain.

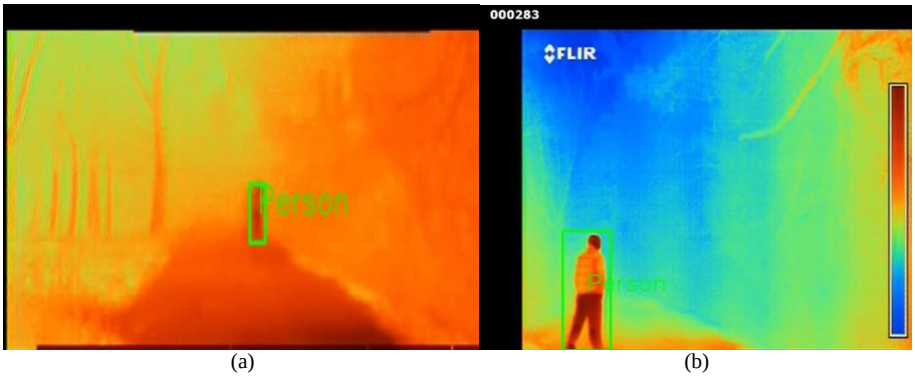
Figs. 12 to 15 show a few examples of the detection results in thermal images for the tY model. The examples show that the model provides good detection results for people at different shooting distances, with different types of motion including normal walking, running and sneaking, and in all tested weather conditions. It is interesting that the model managed to detect people at large distances regardless of the mode of movement, even when the thermal difference between a person and a tree trunk was low. The model also managed to detect people when they stood still (Fig. 12 (a)). Although there are similarities between persons and tree trunks in their contours and temperature curves, there are no false positive detections in the image.



(a) (b)

Fig. 12. Results of person detection on images recorded with telephoto lens in clear weather condition, 110 m distance, (a) normal walk (b) running.

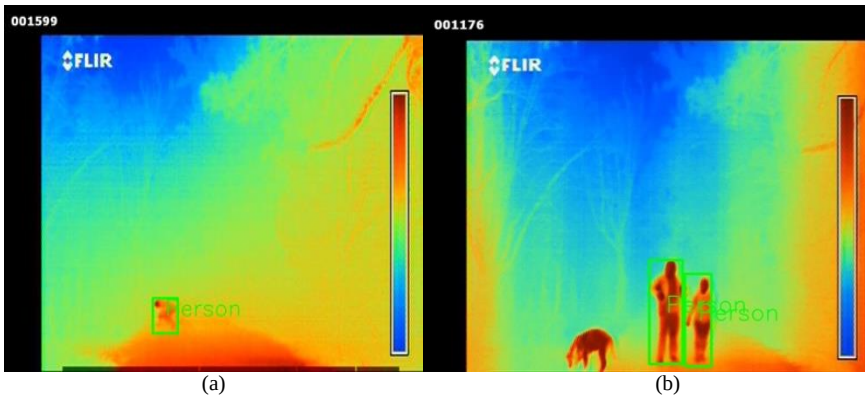
In the case of fog, the dispersion of the temperature is the greatest then in the other observed weather conditions. However, despite this, the model has achieved positive detection of persons both near (Fig. 13. (b)) and at large distances from the camera, Fig. 13. (a).



(a) (b)

Fig. 13. Results of person detection (normal walk) on images recorded in foggy weather with a normal lens, (a) 50 m, (b) <30 m.

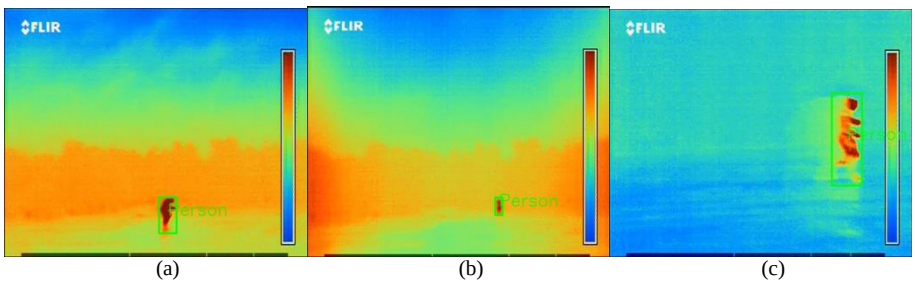
Fig. 14. (a) is an example of a positive detection in the case of fog and when people are moving hunched or are crouching. The increase in the number of people in the image did not harm the detection result, nor did the presence of animals, in this case, a dog, Fig. 14. (b). Note: The dog is not detected because detection is only observed for the class Person.



(a) (b)

Fig. 14. Results of person detection on images recorded with a normal lens in foggy weather, (a) crouched walk, 50 m, (b) standing still, <30 m.

In the case of rain, the detector has successfully detected people while walking, running or creeping, at different distances from the camera (Fig. 15). In rainy conditions, a significant temperature difference exists between the person and the environment, especially when the telephoto lens is used (Fig. 15. (c)). Also, it is evident that the temperature difference between a person and the environment is not constant, even for the same weather conditions (Fig. 15. (b), and 15. (c)). Both Figs. 15. (b) and 15. (c) are recorded in the rain but in Fig. 15. (c) the environment is cold and shown with blue and green while in Fig. (b) it seems to be warmer and is presented with most orange and green although images are recorded on the same day. In both cases, a person is warmer than the environment and has a dark orange and red color.



(a) (b) (c)

Fig. 15. Results of person detection on images recorded in the rainy weather, a) 30 m normal lens, b) 70 m normal lens, c) 30 m telephoto lens.

The given examples have shown that additional training of the original YOLO model (bY) for the class Person in thermal imaging results in new tY model that achieves excellent detection results in different weather conditions. Person detection has proved successful even in the case of large distances from the camera when people appear as a tiny object in the image, as well as with different attempts of avoiding detection such as walking, running or sneaking in a hunched position.

6 Conclusion

The performance of a convolutional neural network-based method for the task of person detection in thermal images was investigated in this paper. Thermal imaging is increasingly employed in surveillance and security because it can be used in the night and in weather conditions such as rain and fog when ordinary RGB cameras cannot be used or give poor results,

The performance of the original YOLOv3 network trained on the COCO RGB as the baseline model was compared with a model that was further trained for person class on a subset of thermal images from a custom dataset. The set consisted of images from videos captured during the winter time in the fog, it the rain and in the clear weather, with people moving at different distances from the camera, ranging from 30 m to 215 m. The movement varied from a normal walk to running and sneaking.

The YOLOv3 model that was not trained on thermal images achieved an AP score of 19.63% and a recall score of 15.5% at 100% precision, which is significantly lower than the reported AP score of about 90% for the Person class in the RGB images [13]. This is expected, as thermal images differ significantly in appearance from the RGB images. However, the model has served as a good starting point for training a specific model for person detection and recognition in thermal imaging.

The performance of the model that was additionally trained on a set of about 3,000 thermal images improved significantly, reaching the AP score of 97.93% for all weather conditions. A modestly sized training set proved to be sufficient to achieve excellent detection results in all tested weather conditions, pose and camera distance variations, encouraging further research for this task. Since the dataset used in this experiment was limited to winter weather conditions, the performance of person detection in other diverse weather conditions such as exceptionally hot weather will be investigated in the future work, as well as the influence of other potentially confusing objects such as wild animals present in the scene.

Acknowledgment

This research was fully supported by the Croatian Science Foundation under the project IP-2016-06-8345 “Automatic recognition of actions and activities in multimedia content from the sports domain” (RAASS).

References

1. P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2001, vol. 1, pp. I–I.
2. N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2005, vol. 1, pp. 886–893.
3. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.
4. R. Girshick, “Fast r-CNN,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
5. W. Liu et al., “SSD: Single shot multi-box detector,” at a European conference on computer vision, 2016, pp. 21–37.
6. K. He, G. Gkioxari, and Dollar, “Mask r-CNN,” in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.
7. J. Dai, Y. Li, K. He, and J. Sun, “R-fcn: Object detection via region-based fully convolutional networks,” in Advances in neural information processing systems, 2016, pp. 379–387.
8. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
9. S. Chang, F. Yang, W. Wu, Y. Cho and S. Chen, "Nighttime pedestrian detection using thermal imaging based on HOG feature," Proceedings 2011 International Conference on System Science and Engineering, Macao, 2011, pp. 694-698.
10. J. Ge, Y. Luo and G. Tei, "Real-Time Pedestrian Detection and Tracking at Nighttime for Driver-Assistance Systems," in IEEE Transactions on Intelligent Transportation Systems, vol. 10, no. 2, pp. 283-298, June 2009.
11. J. W. Davis and M. A. Keck, "A Two-Stage Template Approach to Person Detection in Thermal Imagery," 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1, Breckenridge, CO, 2005, pp. 364-369.
12. J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” arXiv preprint arXiv:1804.02767, 2018.
13. M. Buric, M. Pobar, and M. Ivasic-Kos, “Object Detection in Sports Videos,” in 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018
14. M. Burić, M. Pobar, M. Ivašić-Kos, “Ball detection using YOLO and Mask R-CNN,” in 5th Annual Conf. on Computational Science & Computational Intelligence (CSCI’18), Las Vegas, USA, 2018.
15. M. Kristo and M. Ivasic-Kos, “An Overview of Thermal Face Recognition Methods,” in 2018 41st International Convention on Information and Communication Technology, Electronics, and Microelectronics (MIPRO), 2018.
16. Z. Wu, N. Fuller, D. Theriault, and M. Betke, “A thermal infrared video benchmark for visual analysis,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 201–208.
17. M. K. Bhowmik et al. “Thermal infrared face recognition–a biometric identification technique for robust security system,” in Reviews, refinements and new ideas in face recognition, InTech, 2011.

18. G. Tanda "The use of infrared thermography to detect the skin temperature response to physical activity." Journal of Physics: Conference Series. Vol. 655. No. 1. IOP Publishing, 201
19. J. Redmon and A. Farhadi, “YOLO9000: better, faster, stronger,” arXiv preprint, 2017.
20. T.-Y. Lin, et al. “Feature Pyramid Networks for Object Detection,” arXiv preprint arXiv:1804.02767, 2018.
21. T.-Y. Lin et al. “Microsoft coco: Common objects in context,” in European conference on computer vision, 2014, pp. 740–755.
22. M. Buric, M. Pobar, and M. Ivasic-Kos, “Adapting YOLO network for Ball and Player Detection,” In ICPRAM 2019.
23. M. Krišto and M. Ivašić-Kos, Thermal Imaging Dataset for Person Detection, in 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2019 (In press)
24. M. Ivasic-Kos, M. Kristo, and M. Pobar, “Human detection in thermal imaging using YOLO,” In ICCTA 2019.
25. A. Dutta, A. Gupta, and A. Zissermann, “VGG image annotator (VIA),” URL: http://www.robots.ox.ac.uk/~vgg/software/via, 2016.
26. M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (VOC) challenge,” International Journal of computer vision, vol. 88, no. 2, pp. 303–338, 2010.