Human detection in thermal imaging using YOLO

Marina Ivašić-Kos
Department of Informatics
University of Rijeka
Rijeka, Croatia
marinai@uniri.hr

Mate Krišto
Department of Informatics
University of Rijeka
Rijeka, Croatia
matekrishto@gmail.com

Miran Pobar
Department of Informatics
University of Rijeka
Rijeka, Croatia
mpobar@uniri.hr

ABSTRACT
In this paper, we consider the problem of automatic detection of humans in thermal videos and images. The thermal videos are recorded on a meadow with a small forest with up to three persons present on the scene at different positions and ranges from the camera. To simulate realistic conditions that can happen during surveillance and monitoring of protected areas, all videos are recorded at night but different weather conditions– clear weather, rain, and fog. We present the results of human detection on a custom dataset of thermal videos using the out-of-the-box YOLO convolutional neural network and the YOLO network trained on a subset of our dataset. YOLO is an object detector pretrained on the COCO image dataset of RGB images of various object classes. Test experimental results have shown significantly improved performance of human detection in thermal imaging in terms of average precision for trained YOLO model over the original model.

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

1. INTRODUCTION
Security is nowadays a rising concern, and thus security technologies are becoming more relevant and researched. This concerns the domain of personal security, national security, border protection due to global terrorism threat and illegal migrations, as well as the security of important government and private infrastructure. Investment in security systems reach record highs, and video surveillance systems and protected area control systems are becoming more sophisticated and capable. A significant factor in this development is the improvements in computer vision technologies and successful application of neural networks for object detection.

The goal of object detection is to classify certain objects in images and to provide their exact position. Many successful machine learning algorithms have been developed in the past for the detection of objects such as human faces [1] or full human figures [2] in RGB images.

Currently, the most successful paradigm for object detection in RGB images is based on convolutional neural networks (CNNs). Development started with the great success of AlexNet in the ImageNet Large Scale Visual Recognition Challenge in 2012 [3] for the image recognition task. Since then, several successful CNN architectures have been developed for the object detection task as well, such as R-CNN [4], SSD [5], Mask R-CNN [6], R-FCN [7] and YOLO [8].

In this paper, we consider the application of the CNNs for the task of detecting persons in images and videos obtained with a thermal camera. The thermal camera captures the heat emitted by the subject of the surveillance and forms an image using infrared (IR) radiation, so-called thermogram.

The IR radiation is electromagnetic radiation emitted in proportion to the heat generated/ reflected by an object and, therefore IR imaging is referred to as thermal imaging. The wavelengths of IR are longer than those of visible light, ranging from 400 nm to 1400 nm, Fig. 1, so IR is not visible to humans [9].

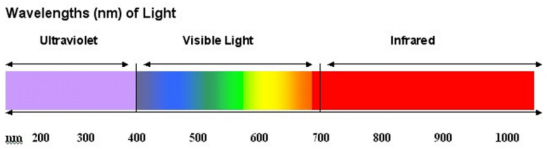


Figure 1. Wavelengths of light in nm (https://www.scienceoflight.org/ir-light/)

Since thermal sensors form imagery of the environment or object solely by the detected amount of thermal energy emission of recorded object, they are, unlike the visible sensors, invariant to illuminating conditions, robust to a wide range of light variations and weather conditions [10, 11].

Thermal cameras can be used in security applications in weather conditions in which regular RGB cameras produce poor results, such as rain and fog or are not useful at all, such as in the total darkness, Fig. 2.

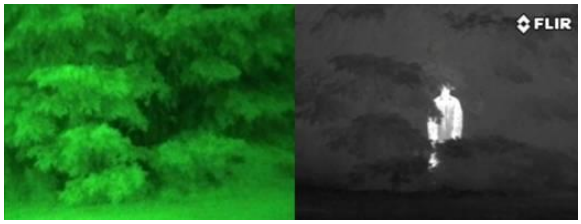


Figure 2. Night vision vs. thermal imagine showing that tree cover cannot hide a person from a thermal camera (https://www.opticsplanet.com/howto/how-to-thermal-imaging-vs-night-vision-devices.html)

On the other hand, IR cameras are susceptible to the variations of surrounding's temperature and provide fewer

details than visible light cameras since color captured in visible spectrum provides much more information and is easier to interpret.

Due to the difference between visual and thermal images, we are interested in exploring how the common deep learning methods successful for object detection and recognition in RGB images [12, 13] will perform with thermal images.

For the detection task, we decided to use the YOLOv3 network [14], which performs at or near state-of-the-art levels in the object detection task in RGB images [15].

In the next section, we briefly describe the YOLO object detector. The dataset and the experimental setup are discussed in Section 3. The results are presented and discussed in Section 4, followed by the conclusion.

2. THE YOLO OBJECT DETECTOR

The original YOLO paper [8] describes the object detection model that uses a single convolutional network to simultaneously predict multiple object bounding boxes in full images as well as class probabilities for those boxes, Fig. 3.

The network architecture of this model has 24 convolutional layers and two fully connected layers. The convolutional layers perform feature extraction while the fully connected layers predict the bounding box locations and their probabilities. The system first divides the input image 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, and if an object occupies more than one cell, the center cell is selected to be the holder of prediction for that object. When training the network, a bounding box that holds no objects has a confidence value of zero, a bounding box around an object has a confidence value that corresponds to the intersection-over-union (IoU) score of the bounding box and the ground truth box.

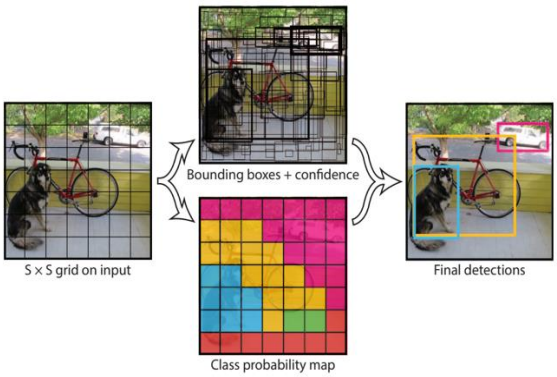


Figure 3. The detection pipeline of YOLO: the input image is divided into a S x S grid where the bounding boxes are simultaneously predicted with corresponding confidence and class probability values [8]

Version 2 of the YOLO detector (YoloV2) [16] replaces five convolution layers of the original model with max-pooling layers and changes the way bounding box proposals are generated. Instead of fully connected layers that predict the bounding box coordinates for each cell, predefined anchor boxes are used. To define the anchor boxes, YoloV2 uses k-means clustering in a training set of GT bounding boxes where boxes translations are relative to a grid cell.

In Yolov3 [14], the 19-layer feature extraction network has been replaced with a much deeper network consisting of 53 layers of 3x3 and 1x1 filters with skip connections. Also, the bounding box prediction was refined, using features at three different scales to make three sets of box predictions for each location. The classification method has also been changed, so now multi-label 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.

3. EXPERIMENT SETUP

In the experiment, we focus on the application of YOLO in the area of surveillance using thermal imaging for human detection in different weather conditions.

We compare the performance of the YOLOv3 network pretrained on the COCO image dataset of RGB images of various object classes [17], used as baseline model and referred to as bYOLO, with the performance of YOLO with additional training on thermal images from our dataset (referred to as tYOLO). Even though thermal images differ significantly in appearance from the RGB images, it is expected that individual layers of RGB images still sufficiently resemble the thermal image, so that the features learned on training data of RGB images should still provide a reasonable baseline for thermal images. Additionally, an experiment in [8] has shown that the YOLO detector performance degraded less than of other detectors when applied to person detection in artwork, a domain that was not used in training the network.

We have compared the baseline performance with the YOLO network that was trained on thermal image data for the class Person.

The evaluation is performed using the mean average precision (mAP) criteria, like the one used in the PASCAL VOC 2012 competition [18].

3.1. Dataset

The data for our experiment was collected by recording humans during the night in different weather conditions and different ranges from the camera. The videos are taken using the FLIR ThermaCAM P10 thermal camera. It is a focal plane array (FPA) camera with uncooled bolometer which covers the spectral range between 7.5 and 13 m (LWIR). The camera has a sensor resolution of 320 x 240 pixels, but we used a digital recorder which converted and unsampled the video to the resolution of 1280 x 960 pixels in AVI format. We recorded five men and two women during the winter time (in February 2017.) in several lens and range configurations. The people moved in normal walking

position and hunched position, and with normal walking speed and running.

The base range for recording was 110 m, conducted using basic camera equipment and lenses with 24° x 18° / 0.3 m field of view. Additionally, we used the FLIR P/B series telephoto lenses with 7°x5.3V FOV and 3.5x magnification.

Recordings were made in different weather conditions, with appropriate setups. In clear weather, the distance of people from the camera was either 110 m (base) or 165 m. Recordings in heavy fog, with minimum visibility to about 5 m were made with people less than 30m from the camera and with people moving at a distance of 50 m from the camera. In the fog, using standard lenses or recording at larger distances was not possible, so we only used the telephoto lens here. In the heavy rain condition, the people were moving at 30m, 70m, 110m, 140m, 170m, 180m, and 215m from the camera.

After recording the videos, we extracted individual video frames to create the dataset. We got 15.000 images taken with a telephoto lens on clear, fog and rain condition and about 6.000 images taken with a standard lens on clear weather conditions.
For the training, we used approximately 1.000 images for each weather condition. Images were manually annotated using the VGG Image Annotator (VIA) [19].

4. RESULTS AND DISCUSSION

The Average Precision (AP) measure is used to evaluate the performance of the models. The detection results are compared with the ground truth so that for a detection to be counted as a true positive, intersection over union (IoU) score of the detection bounding box and the corresponding ground truth bounding box should be at least 50%. An example of positive and negative object detection concerning intersection over union (IoU) score in case of ball detected is shown in Fig. 4.

Figure 4. Visual representation of IoU criteria [20]

When the same object is detected multiple times, only one detection is counted as a true positive. To get the mAP value, mean of AP value of all classes is calculated, but in this experiment, we consider only one class, Person.

By varying the confidence threshold of the detector, a precision-recall curve can be produced for the desired class. The AP score is then the area underneath the precision-recall curve, Fig. 5.

Fig. 5 presents AP score for original YOLO model, bYOLO, that was not trained on our datasets, while Fig. 6 corresponds to AP score obtained with the tYOLO model that is additionally trained on our custom datasets. The AP score achieved with tYOLO of 29% significantly exceeds the AP score achieved by bYOLO of 7%.

For example, the bYOLO model achieves a precision of 97% with a recall of 6%, while the model tYOLO achieves the same precision with a much higher recall of approx. 20%, meaning the tYOLO model detects a lot more people in the images with the same precision.

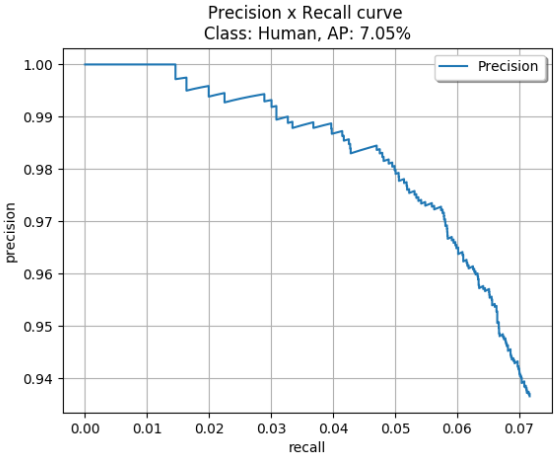


Figure 5. AP score and precision/recall curve for baseline YOLO model, bYOLO

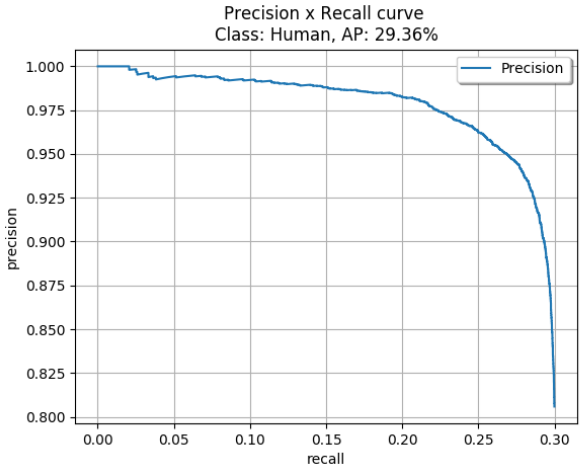
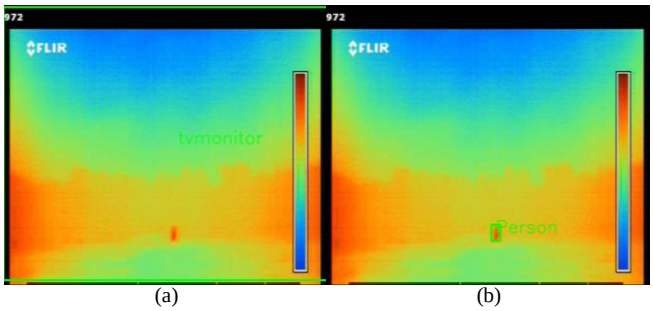


Figure 6. AP score and precision/recall curve for custom trained YOLO model, tYOLO

Below, Figs. 7 to 11 show examples of the detection results of both models bYOLO and tYOLO with respect to different weather conditions and different distances from the camera.

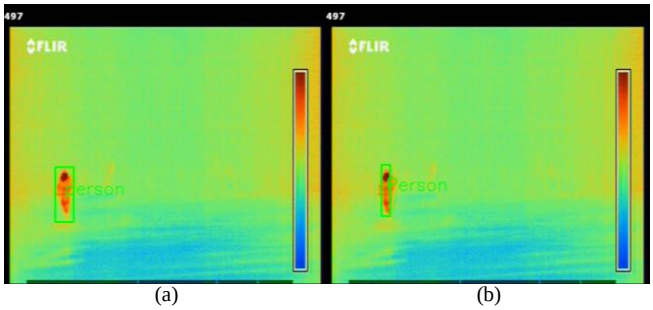
In all examples, the model tYOLO has a true positive detection (TP) of people, Figs. 7 to 11 (b), while model bYOLO had positive detection only in the case of a normal walk on images recorded with telephoto cameras on rain condition, with a distance of 70m from the camera, Fig. 8 (a). Interestingly, in the case of the same weather conditions but at a distance of 100 meters, (for 30 meters longer), when the person was hunched, the model bYOLO failed to detect the person. On the other hand, change of people behavior and activity from a person walking to hiding and hunched walking and running did not affect the detection result for model tYOLO, Figs. 9 (b) and 10 (b).



(a) (b)

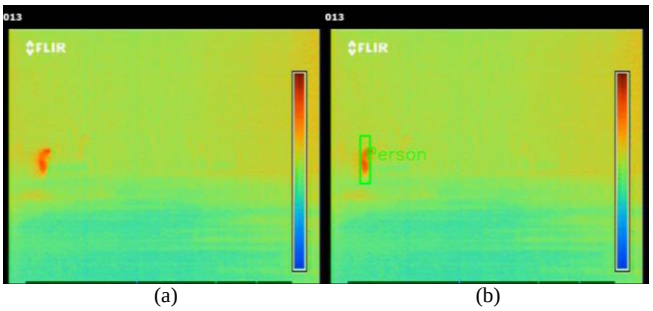
Figure 7. Results of person detection (hunched walk) on images recorded with a standard lens on rain condition, 70m distance, using bYOLO (a) opposite to tYOLO model (b).

Fig. 7(a) shows a false positive detection of the bYOLO model for class tv-monitor and false negative detection for a person since no person was detected although the person exists on the image. In all other cases, Figs. 9 to 11 (a), the model bYOLO did not detect the person in the image even though it was present (false negative detection).



(a) (b)

Figure 8. Results of person detection (normal walk) on images recorded with a telephoto lens on rain condition, 70m distance, using bYOLO (a) opposite to tYOLO model (b).

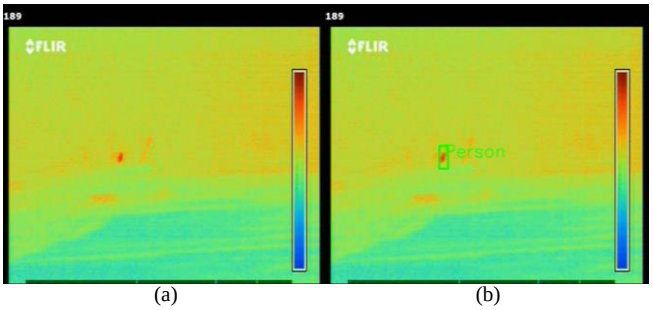


(a) (b)

Figure 9. Results of person detection (hunched walk) on images recorded with a telephoto lens on rain condition, 100m distance, using bYOLO (a) opposite to tYOLO model (b).

On all images in rain conditions (Figs. 7 to 10), there is a large temperature difference between the person on image who is hot and marked with red and the environment that is cold and marked ranging from blue to green. The heat difference makes it easier to detect a person, but that temperature difference is not present at all weather conditions. E.g., in the case of fog, (Fig. 11) the temperature difference between the person and the environment is much smaller and the detection according to the heat map is much harder.

In the case of fog, bYOLO could not detect any person, Fig. 11(a), and tYOLO has detected one of the two people present on the scene, Fig. 11(b).



(a) (b)

Figure 10. Results of person detection (running) on images recorded with a telephoto lens on rain condition, 215m distance, using bYOLO (up) opposite to tYOLO model (b).

(a)

(b)

Figure 11. Results of detection of a group of people (2 persons) on images recorded with a telephoto lens on fog, 50m distance, using bYOLO (a) opposite to tYOLO model (b).

The selected detection examples, as well as the results presented in Figs. 5 and 6. have shown that the original YOLO model (bYOLO) trained for class Person on COCO dataset of RGB images does not achieve good detection results in any of the weather conditions and thus cannot be used directly on thermal images. This is especially so in the case of large distances from the camera when people are hunched or are running and appear as a tiny object in the image. Additional learning on thermal imaginary results in significant improvement of people detection in different weather conditions, especially when the temperature difference between persons and the environment is large.

5. CONCLUSION

In this paper, we wanted to examine how the common deep learning methods that are successful for object detection and recognition in RGB images, such as the YOLO detector, perform with thermal images.

The task was to detect persons in videos captured during the winter time in different weather conditions during the night and with different distance from the camera, ranging from 30m to 215m. The persons were walking and running or walking hunched and trying to stay out of sight.

Even though thermal images differ greatly in appearance from the RGB images, we have assumed that the features that YOLO has learned on large COCO dataset of RGB images for the class Person will still provide a reasonable baseline for thermal images. Unfortunately, due to the difference between visual and thermal images, the original YOLO model (bYOLO) has achieved average precision (AP) of only 7% for person detection in the thermal images. That result is significantly worse than the results YOLO achieves on the images of the visible spectrum where the results depending on the scenario range around 90% [21]. Therefore, we have additionally trained the bYOLO model on thermal images from our custom dataset, and after training the model, tYOLO has achieved significantly better results of AP approx. 30% for person detection in different weather conditions and with different distances from the camera.

The experiment has shown that the performance of YOLO model on thermal imagery can improve significantly with additional training on the thermal dataset. We plan to investigate further how the different conditions affect the detection performance, such as distances of the object from the camera, different time conditions, and for each of the situations, we will examine the success of the model.

In the future, we intend to expand a dataset with other objects that may occur in the observed scenario of forests and meadows such as wolves, foxes, wild boars, and other forest animals. The idea is to learn a model that can detect them to avoid misleading detection and alarms of illegal access to the controlled area in cases when it comes to the passing of animals.

6. ACKNOWLEDGMENT

This research was supported by Croatian Science Foundation under the project IP-2016-06-8345 “Automatic recognition of actions and activities in multimedia content from the sports domain” (RAASS) and by the University of Rijeka under the project number 18-222-1385.

7. REFERENCES

[1] Viola, P., Jones M. 2001. “Rapid object detection using a boosted cascade of simple features,” in 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. I–I.

[2] Dalal, N., Triggs B. 2005. “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] Krizhevsky, A., Sutskever, I., and Hinton. G. E. 2012. “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, pp. 1097–1105.

[4] Girshick, R. 2015. “Fast r-CNN,” in Proceedings of the IEEE international conference on computer vision, pp. 1440–1448.

[5] Liu, W. et al. 2016. “SSD: Single shot multi-box detector,” in a European conference on computer vision, 2016, pp. 21–37.

[6] He, K., Gkioxari, G., and Dollar, P. 2017. “Mask r-cnn,” in 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988.

[7] Dai, J., Li, Y., He, K., and Sun, J. 2016. “R-fcn: Object detection via region-based fully convolutional networks,” in Advances in neural information processing systems, pp. 379–387.

[8] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. 2016. “You only look once: Unified, real-time object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788.

[9] Kristo, M., and Ivasic-Kos, M. 2018. “An Overview of Thermal Face Recognition Methods,” in 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO).

[10] Wu, Z., Fuller, N., Theriault, D., and Betke, M. 2014. “A thermal infrared video benchmark for visual analysis,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 201–208.

[11] Bhowmik, M. K., et al. 2011. “Thermal infrared face recognition–a biometric identification technique for robust

security system," in Reviews, refinements and new ideas in face recognition, InTech.

[12] Pobar, M., Ivašić-Kos, M. 2018. "Detection of the leading player in handball scenes using Mask R-CNN and STIPS" in 11th International Conference on Machine Vision (ICMV 2018), Muenchen, Germany: SPIE.

[13] Burić, M., Pobar, M., Ivašić-Kos, M. 2018. "Ball detection using YOLO and Mask R-CNN," in 5th Annual Conf. on Computational Science & Computational Intelligence (CSCI'18), Las Vegas, USA.

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

[15] Ivašić-Kos, M., Pobar, M. 2018. "Building a labeled dataset for recognition of handball actions using mask R-CNN and STIPS", in 7th IEEE European Workshop on Visual Information Processing (EUVIP), Tampere, Finland, pp. 1-6

[16] Redmon, J. and Farhadi, A. 2017. "YOLO9000: better, faster, stronger," arXiv preprint.

[17] Lin, T.-Y. et al. 2014. "Microsoft coco: Common objects in context," in a European conference on computer vision, pp. 740–755.

[18] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. 2010. "The pascal visual object classes (VOC) challenge," International Journal of computer vision, vol. 88, no. 2, pp. 303–338.

[19] Dutta A., Gupta, A., and Zissermann A. 2016. "VGG image annotator (VIA)," URL: http://www.robots.ox.ac.uk/~vgg/software/via.

[20] Buric M., Pobar M., and Ivasic-Kos, M. 2019. "Adapting YOLO network for Ball and Player Detection", In ICPRAM 2019.

[21] Buric, M., Pobar, M., and Ivasic-Kos, M. 2018. "Object Detection in Sports Videos," in 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO).