DOI 10.5162/sensoren2024/A1.3

Differentiation of Humans and Robots with Thermal Images and Convolutional Neural Networks for Human-Robot Collaboration

Sinan Süme[1], Katrin-Misel Ponomarjova[1], Thomas M. Wendt[1], Stefan J. Rupitsch[2]

[1]Offenburg University of Applied Sciences, Work-Life Robotics Institute, 77652 Offenburg, Germany
[2]University of Freiburg, IMTEK - Department of Microsystems Engineering, 79110 Freiburg, Germany

Abstract

This paper introduces the use of convolutional neural networks to detect and classify humans and robots in Human-Robot Collaboration workspaces based on their thermal radiation power. The measurement setup includes an infrared camera, two cobots and up to two persons walking or interacting with the cobots in industrial settings. The chosen architectures are the YOLOv5 and YOLOv8 in different model sizes. The results are promising, showing real-time object detection in industrial settings with up to 303 fps with the YOLOv8n model. YOLOv5m achieves the best mAP50 result at 99.2% and the YOLOv5m achieves the best mAP50-95 at 85.8%.

Keywords: Real-time object detection, Human-Robot Collaboration, Human-Robot Differentiation

Introduction

As collaborative robots (cobots) become more prevalent in production, it is expected that humans and robots will be able to work together without compromising efficiency or safety [1]. The ISO/TS 15066 describes four types of safe collaborations between robots and humans [2]. The focus of this research is on speed and separation monitoring and the safety-rated monitored stop. Another emerging trend is the use of autonomous mobile robots (AMR) with potentially mounted cobots for dynamic and collaborative workspaces. Some of the challenges in the use of AMRs are dynamic obstacle avoidance and autonomous navigation and path planning [3]. Differentiating between humans and robots can lead to increased safety and efficiency in collaborative, dynamic and smart workplaces. Robots must slow down or stop in a dangerous situation when humans are approaching. In the case of approaching AMR, Automated Guided Vehicle or another robot in multi-robot applications, there is, however, no need to slow down or even stop. Consequently, it is important to know the features of a human so a distinguishment between a robot, a human or other objects is possible. This contribution deals with the possibilities of the detecting humans and cobots in collaborative workspaces with Convolutional Neural Networks (CNN) based on their thermal radiation power.

Related Work and Motivation

Fraden [5] gives an overview of methods to detect human presence. One promising approach is to measure the thermal radiation of humans. Earlier work [4] proved that infrared sensors with low resolution (32 x 32 px) are capable of detecting a human in infrared images with the use of the CNN GoogLeNet and MobileNetV2 with up to 99.48% accuracy. Interfering heat sources were not considered which can lead to distortions in human detection results.

Materials and Methods

In order to capture environmental information regarding thermal radiation, an infrared sensor or camera is required. The camera used to collect the data and detect the presence of cobots and humans is the FLIR-Camera T440. The features of the infrared camera are presented in Tab. 1. Training and testing of the utilized CNN was performed on Windows 10. The computer has a 10-core CPU, 64 GB of main memory and an Nvidia RTX3080 with an integrated GPU memory of 12 GB. The focus of the detection algorithm will be on the YOLO architectures. The YOLO architectures deliver the best overall results in accuracy and inference speed [6].
The proposed detectors are the YOLOv5 and YOLOv8 architectures, which are single-stage object detection algorithms, intended for real-time applications.

The model sizes of the networks chosen for comparison are the between nano (n), small (s), and medium (m) size of both YOLO versions. Real-time applications have to consider the tradeoff between higher accuracy rate of larger architectures and faster processing with smaller architectures [7]. Instead of pretrained models on a dataset, we took default YOLO parameters. The training itself lasted 50 epochs. The dropout as well as the erase function were set to 80%, to prevent overfitting. A fixed seed allows to compare the test results across the models.

Table 1: List of features of the T440-Camera.

Features	Value
Frame rate	60 Hz
Resolution	320 x 240 px
Field of view	25° x 19°
Thermal sensitivity at 30° C	0.045 °C
Spectral range	7.5 - 13 µm

Measurement Setup

As shown in Fig. 1, the infrared camera was placed at a distance of 10 m. This enables to capture a wide range of different interactions between the humans and cobots.

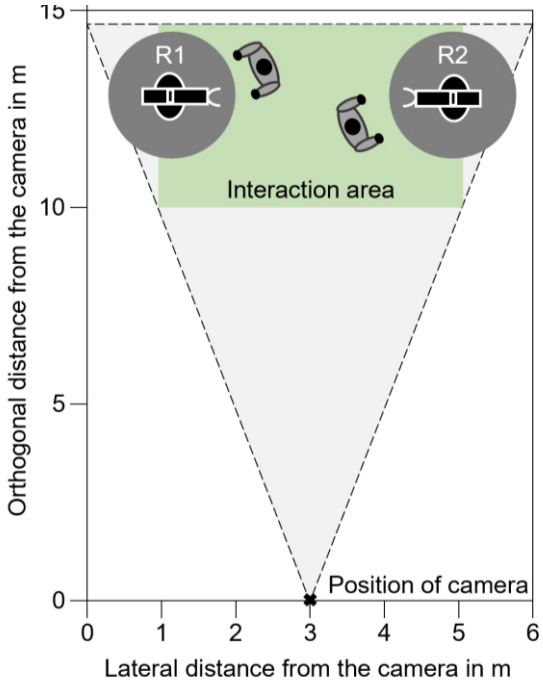

Figure 1: Illustration of the measurement setup with two humans in the interaction area.

The interaction area has a length of 4.5 m and a width of 4 m. The experimental setup includes two cobots and up to two humans in one image. The cobots used in the setup are the KUKA LBR iiwa 7 (R2) and the KUKA LBR iiwa 14 (R1). Both KUKA models differ in size, payload and the R2 has the handguiding functionality. Handguiding enables manipulating the cobot by applying force to the flanch of the cobot. In this research, we exploited the handguiding function to simulate the interaction between the human and the cobot, as shown in Fig. 2. The data contains different scenarios, which includes a person walking in the interaction area, two persons walking and interacting differently with the cobots and handguiding the R2. While moving, the persons can leave the scene in lateral directions, to create partially visible body parts or occluded images of persons or cobots. The cobots were moving while the persons were walking. The environment temperature during the measurement was between 21.5 °C and 22.9 °C. The humidity was 37% and the ambient lighting in the interaction area was between 908 lx and 1381 lx.

Figure 2: Infrared image in industrial settings with two cobots, moving human (left) and human cobot interacting (right).

Data Preprocess

The first step of data preprocessing was to convert the video-stream based on the temperature value into grayscale. The procedure of generating grayscale images from colored images decreases the information to process for the model, which can lead to faster training and prediction times. The transformation into grayscale should be considered only, if the colorization is irrelevant for the classification [8]. In this case, the use of the radiation power for the detection does not differ between the classes or contribute to the differentiation between cobots and persons in a specific way. The frames of the infrared video stream were extracted.

One frame was extracted every second of the video stream with a Python script, which led to a total sum of 3083 images. The images were labeled with the classes “person” and “robot” and split randomly into 70% training, 20% validation and 10% test data as shown in Tab. 2. Data augmentation methods like cropping up to 30%, blurring up to 4.5 px and adding noise with up to 1.9 px were employed. These methods quadrupled the number of the training data to 4536 images and helps preventing overfitting. The datasets were resized before employing the model from the original resolution from the FLIR camera to 640 x 640 px to save processing time.

Table 2. Distribution of the dataset into training validation and test sets.

Data split	No. images	No. images after data augmentation
Train	1134	4536
Validation	1300	1300
Test	649	649
Total	**3083**	**6485**

Metrics

The metrics in this paper are precision, recall, average precision (AP) and mean average precision (mAP) to evaluate the accuracy of the models on the test data. The precision P can be calculated by (1) and the recall R by (2). P is the capability of a model to identify the relevant object and determines the percentage on correct positive predictions. R defines the percentage of correct predictions based on all ground truths. The confidence-threshold value defines the number of predictions the model makes. A higher threshold leads to less predictions and so the model tends to have a higher P value, while the R value tends to decrease and vice versa [9].

$$P = \frac{True_{\ pos}}{True_{\ pos} + False_{\ pos}} \tag{1}$$

$$R = \frac{True_{\ pos}}{True_{\ pos} + False_{\ neg}} \tag{2}$$

The metric of the intersection over union (IOU) allows to determine when a prediction of a bounding box is correct. The IOU overlays the predicted bounding box B_p on the ground truth bounding box B_{gt} and divides the intersection area by the area of both boxes, as shown in

$$IOU = \frac{area\ (B_p \cap B_{gt})}{area\ (B_p \cup B_{gt})}. \tag{3}$$

If the value is above a certain threshold, then the prediction will be defined as true positive. So the given threshold to define whether a prediction is correct or not has to be stated [9]. The mean average precision (mAP) is used to describe the average precision over all classes as in (4), while N describes the number of classes and AP_i is the average precision of the i-th class[9]. For example, the mAP50 is the AP of all classes at an IOU of 50, while the mAP50-95 varies the threshold of IOU between 50 and 95 in incremental 5% steps.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{4}$$

The evaluation of the real-time capability combines the pre-processing ($proc_{pre}$), inference and post-processing ($proc_{post}$) time and will be calculated for each model with the test dataset in FPS (frames per second), as shown in

$$FPS = \frac{1000}{proc_{pre} + inference + proc_{post}}. \tag{5}$$

Results

The trained models have been deployed on the unseen test dataset to make predictions and the results are available in Tab. 3. Fig. 3 illustrates the prediction of the YOLOv5n. YOLOv5n has the best P value with 98.8%. The best R value is achieved by the YOLOv5m with 97.4%. The highest mAP50 value is 99.2% accomplished by the YOLOv5m and YOLOv8s. As for mAP50-95, the medium size models achieve the highest value, where the YOLOv5m (85.8%) is slightly higher as the YOLOv8m (85.6%). As for FPS, the lowest result was reached by YOLOv8m with 117.6 fps. The highest result was achieved by YOLO8n with 303 fps.

Discussion

The YOLOv5 and the YOLOv8 architectures show similar results, with high values for all metrics. We applied various augmentation methods, dropout, and erasing techniques during training to avoid overfitting. The high performance can mainly be attributed to the model’s ability to differentiate between the distinct features of humans and robots based on their infrared radiation intensity. The difference between the results of these architectures is mainly in the frames predicted per second. YOLOv8n surpassed all other architectures, while the larger models showed reduced FPS.

Table 3: Results of the different architectures and models on the test dataset.

Model	P	R	mAP50: All	AP50: Robot	AP50: Person	mAP 50-95: All	AP 50-95: Robot	AP 50-95: Person	FPS
YOLO v5n	0.988	0.956	0.991	0.993	0.99	0.839	0.836	0.841	217.4
YOLO v5s	0.982	0.945	0.987	0.993	0.981	0.849	0.858	0.84	232.6
YOLO v5m	0.984	0.974	0.992	0.993	0.991	0.858	0.857	0.859	144.9
YOLO v8n	0.975	0.965	0.989	0.992	0.986	0.848	0.84	0.855	303.0
YOLO v8s	0.971	0.969	0.992	0.993	0.99	0.848	0.854	0.843	222.2
YOLO v8m	0.982	0.972	0.991	0.99	0.993	0.856	0.847	0.866	117.6

Figure 3: Results of the prediction with the YOLOv5n model, detecting occluded person and interacting person and both cobots in different poses.

Conclusion and Future Work

The paper's approach to differentiate between human and robots in a collaborative workspace based on the infrared radiation intensity delivers promising results. These initial explorative results provide a foundation for a more general differentiation in industrial settings. The dataset could be further enhanced by collecting more frames from various camera angles and heights. Some scenes were more complicated for the model to predict, such as where one person is positioned behind another, as shown in Fig. 4. This can be prevented by using two or more cameras at different angles. More data can be collected including individuals wearing hard hats, safety shoes, and safety glasses could enhance the model's ability to specialize in industrial environments. Recordings can be made at different temperature ranges and at different ambient temperatures. After some additional refinement, it could be conceivable to implement such a model in real-time detection tasks in industrial environments.

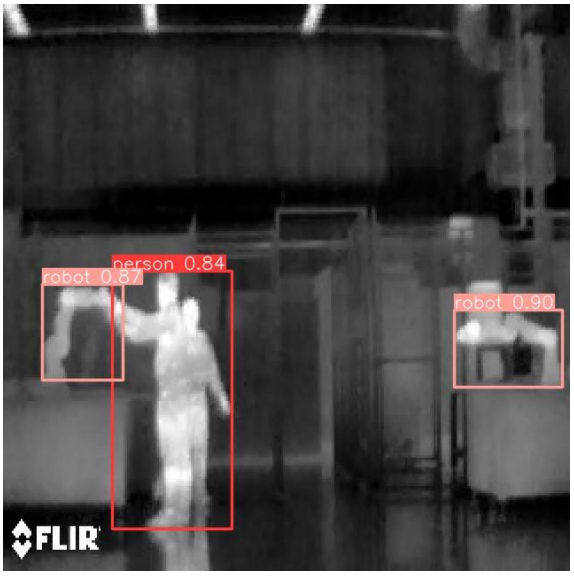

Figure 4: Results of the prediction with the YOLOv5n model, wrongly detecting one person and not two due to occlusion of the second person.

Literature

[1] S. Proia, R. Carli, G. Cavone, and M. Dotoli, "Control Techniques for Safe, Ergonomic, and Efficient Human-Robot Collaboration in the Digital Industry: A Survey," *IEEE Trans. Automat. Sci. Eng.*, vol. 19, no. 3, pp. 1798–1819, Jul. 2022, doi:10.1109/TASE.2021.3131011.

[2] International Organization for Standardization, "ISO/TS 15066: Robots and robotic devices - collaborative robots." Geneva, Switzerland, 2016.

[3] M. B. Alatise and G. P. Hancke, “A Review on Challenges of Autonomous Mobile Robot and Sensor Fusion Methods,” *IEEE Access*, vol. 8, pp. 39830–39846, 2020, doi: 10.1109/ACCESS.2020.2975643.
[4] U. B. Himmelsbach, S. Süme, and T. M. Wendt, “Classification of Thermal Images for Human-Machine Differentiation in Human-Robot Collaboration Using Convolutional Neural Networks,” in *2023 20th International Conference on Ubiquitous Robots (UR)*, Honolulu, HI, USA: IEEE, Jun. 2023, pp. 730–734. doi: 10.1109/UR57808.2023.10202384.
[5] J. Fraden, *Handbook of Modern Sensors: Physics, Designs, and Applications*. Cham: Springer International Publishing, 2016. doi: 10.1007/978-3-319-19303-8.
[6] S. Srivastava, A. V. Divekar, C. Anilkumar, I. Naik, V. Kulkarni, and V. Pattabiraman, “Comparative analysis of deep learning image detection algorithms,” *J Big Data*, vol. 8, no. 1, p. 66, Dec. 2021, doi: 10.1186/s40537-021-00434-w.
[7] C. Li *et al.*, “YOLOv6 v3.0: A Full-Scale Reloading.” arXiv, Jan. 13, 2023. Accessed: Apr. 15, 2024. [Online]. Available: http://arxiv.org/abs/2301.05586
[8] Y. Xie and D. Richmond, “Pre-training on Grayscale ImageNet Improves Medical Image Classification,” in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds., Cham: Springer International Publishing, 2019, pp. 476–484. doi: 10.1007/978-3-030-11024-6_37.
[9] R. Padilla, S. L. Netto, and E. A. B. Da Silva, “A Survey on Performance Metrics for Object-Detection Algorithms,” in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, Niterói, Brazil: IEEE, Jul. 2020, pp. 237–242. doi:10.1109/IWSSIP48289.2020.9145130.