Human Detection from Thermal Images

Yiğit KOTAMAN, Alperen TEKİN
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi, 34220 Istanbul, Türkiye
{yigit.kotaman, alperen.tekin}@yildiz.edu.tr

***Özetçe* —Bu çalışma; düşük aydınlatma, sis, yağmur ve benzeri zorlu çevre koşullarında insan tespitini iyileştirmek amacıyla termal görüntüler üzerinde CNN tabanlı YOLOv11 modeli ile transformer tabanlı DETR modelini karşılaştırmaktadır. Çeşitli açık kaynak veri setleriden özenle seçilip ön işlemeden geçirilen 24.389 termal görüntü ortak bir havuza aktarılarak hem YOLO hem de COCO biçimlerine dönüştürülmüştür ve böylece her iki model aynı veriler üzerinde eğitilmiştir.**

Modellerin performansı doğruluk, duyarlılık, F1 skoru ve mAP@0.50 skoru metrikleriyle sunulmuştur. Sonuçlar, YOLOv11'in daha kısa eğitim süresi ve daha düşük hesaplama maliyetiyle DETR'ye kıyasla daha yüksek performans değerlerine ulaştığını ortaya koymuştur. DETR ise uzun eğitim süresi ve yüksek donanım gereksinimine rağmen dikkat mekanizmaları sayesinde bağlam bilgisini etkili kullanarak dikkate değer sonuçlar vermiştir.

Anahtar Kelimeler—termal görüntü, insan tespiti, YOLO, DETR, kıyaslama, derin öğrenme.

***Abstract*—This study compares the CNN-based YOLOv11 model and the transformer-based DETR model on thermal images to improve human detection in harsh environmental conditions such as low illumination, fog, rain and so on. A total of 24,389 carefully selected and pre-processed thermal images from various open source datasets were transferred to a common repository and converted into both YOLO and COCO formats, so that both models were trained on the same data.**

The performance of the models is presented in terms of accuracy, sensitivity, F1 score and mAP@0.50 score metrics. The results show that YOLOv11 achieves higher performance values compared to DETR with shorter training time and lower computational cost. DETR, on the other hand, despite its long training time and high hardware requirements, achieved remarkable results by utilising context information effectively through attention mechanisms.

Keywords—thermal imagery, human detection, YOLO, DETR, comparison, deep learning.

I. Introduction

Dim light, fog, rain, snow, hail and similar environmental factors pose some difficulties in human and object detection for today's image processing technology.[1] Since thermal image technology uses the temperature of objects, it reduces such difficulties and offers an alternative method to image processing. In current studies, various methods are used in human detection on thermal images. Two of the prominent methods are YOLO, a CNN-based model, and DETR, a transformer-based model.[2], [3] Moreover, in addition to method differences, the variety of data sets used, the angles, distances and ambient conditions at which images are captured directly affect model performance.[4] It has been shown in academic studies that deep learning based models perform better than general purpose trained models when they are trained in accordance with the situations and conditions (fire images, closed circuit city camera images, drone images, etc.) in which they will be used.[5], [6], [7]

In this study, YOLOv11 and DETR models were trained on a specially compiled dataset and the performance of the models were compared. The performance comparison is based on metrics such as accuracy, sensitivity, F1 score and mAP score. Thus, it is aimed to bring a new perspective to the studies in this field and to pave the way for future studies.

II. The Proposed Method

In this study, two different object detection methods, CNN-based YOLOv11 and transformer-based DETR, were used. Both models aim to detect people in thermal images and have been developed with certain optimisations.

A. YOLO

YOLO is a model that processes the image at once with a "you only look once" approach. It divides the image into grid cells and detects objects in each cell. Significant performance improvements have been achieved in recent versions. It offers an optimum balance between speed and accuracy, especially suitable for real-time applications. It is also the model generally used in image processing studies.

B. DETR

It offers an end-to-end approach that combines CNN and Transformer architectures. It eliminates the manual components of traditional methods, such as bounding boxes and non-maximal suppression. It has the ability to perform parallel inference and better utilise comprehensive context information. However, training time is long and computational cost is high. It is not an option generally used in image processing studies. This study also stands out with the use of the DETR model.

III. Experiment and Result

A. Dataset

The dataset used in this study is composed of quality samples selected from different thermal image datasets. Due to the incompatibility of the labelling systems of the different datasets, the entire dataset was preprocessed and converted

to the YOLOv11 labelling format as a common labelling format.

The dataset contains a total of 24,389 thermal images. The images were taken in different weather conditions and from various distances, providing the necessary diversity for a more general and robust performance of the model. The majority of the images were taken at close and medium distances. This allows the models to better detect people at these distances. With this diversity of the dataset, it is aimed that the model can provide effective results in various environmental conditions and distances.

1) Converting the labels of the datasets: The datasets are usually labelled using popular annotation formats (such as YOLO, COCO and XML). In this study, a conversion between different formats was performed.

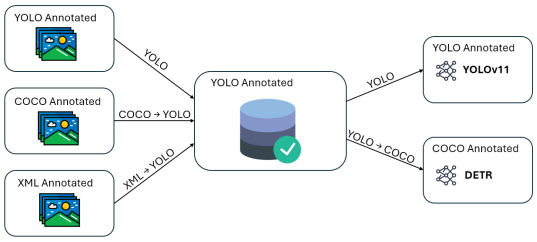

Figure 1 Creating the database

The following steps are followed to convert the data in COCO or XML formats into YOLO format:

1) The coordinates of the boundary boxes in the COCO or XML file are extracted.
2) These coordinates are converted to normalised coordinates in YOLO format.
3) The class label for each object is determined and written appropriately in YOLO format.
4) The converted data is saved in a new file in YOLOv11 format.

The following steps are followed to convert the data set in YOLO format to COCO format:

1) The coordinates of the boundary boxes in YOLO format are converted to 'xmin', 'ymin', 'width', 'height' according to COCO format.
2) The class label for each object is mapped to the class identification number in COCO format.
3) Additional parameters such as 'track id', 'keyframe' in COCO format are assigned for each object in the data set.
4) Finally, all necessary metadata in COCO format is generated and written to the JSON file.

This transformation process makes it possible to train both models on the same dataset, enabling accurate comparisons.

YOLOv11, one of the models used in this study, can be trained with YOLO annotated data set, while another model, DETR, requires COCO annotated data set for training. Since these models need to be trained with a common data set in order to compare the two models, the data were collected in a common pool and then converted to COCO format so that both models could be trained on the same data set.

B. Training

Although there are some differences in the training stages of the object detection models used, they are mostly similar. In this section, the training steps of YOLO and DETR models are explained.

The following steps were followed for the training of the YOLO model:

1) **Preparation of Data Set**: Initially, the datasets from different sources were organised in YOLOv11 format. This format contains boundary boxes and class labels for each image. After checking the accuracy of the dataset, the dataset was divided into three parts: training data, validation data and test data for training the model. These parts are 71% training data, 9% validation data, 20% test data.
2) **Model Training**: In the training phase of the model, the model was trained in the most optimal way by looking at the performance per cycle. The hyperparameters were trained and tested separately to reach the optimal version of the model. The number of revolutions was adjusted to fully train the model.
3) **Performance Evaluation**: The performance of the YOLOv11 model was evaluated with metrics such as accuracy rate, sensitivity rate, mAP score and F1 score. In order to improve these values, the model was returned to the training phase when necessary.

The following steps were followed for the training of the DETR model:

1) **Data Format Conversion**: The labels in YOLOv11 format were converted to COCO format. This process involved accurately reconstructing each boundary box and class label. During the conversion of the dataset, the extra context information contained in the COCO format and used by the DETR model was added to the dataset. After checking the accuracy of the dataset, the dataset was divided into three parts: training data, validation data and test data for training the model. These parts are 71% training data, 9% validation data, 20% test data.
2) **Model Training**: The data set converted to COCO format was made suitable for the training of the DETR model. The training process of the model was optimised taking into account the diversity of the data set. The hyperparameters were trained and tested separately to reach the optimal version of the model. The number of revolutions was adjusted to fully train the model.
3) **Performance Evaluation**: The performance of the DETR model was evaluated with metrics such as accuracy rate, sensitivity rate, mAP score and F1 score. In order to improve these values, the training phase was returned to the training phase when necessary.

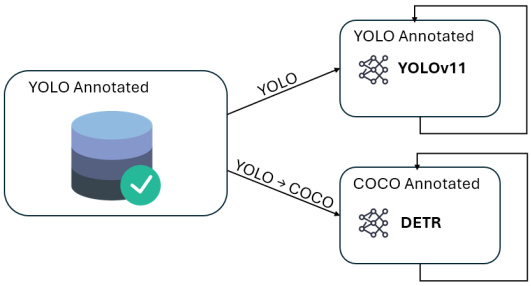

Figure 2 Training of the models

C. Results

The performance results of the models are presented in terms of accuracy score, sensitivity score, F1 score, mAP@0.50 score and mAP@0.50-0.95 score. The calculation of the performance metrics is shown in the formula below. The concepts can be thought of as follows: **TP(True Positive):** True positive means true, **TN(True Negative):** To call a false event false, **FP(False Positive):** To call true what is false, **FN(False Negative):** To call a true event false.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

$$mAP = \frac{1}{n} \sum_{k=1}^{k=n} AP_k \tag{4}$$

D. YOLOv11

Table 1 YOLOv11 Results and Hiperparameters

Precision	0,94
Recall	0,85
F1	0,89
mAP@0.50	0,92
mAP@0.50-0.95	0,65

Epoch	100
Batch	16
Learning Rate	0,02
Confidence	0,28

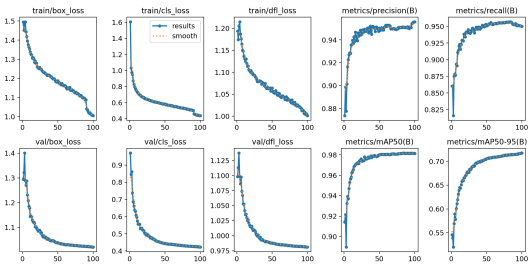

Figure 3 Graphics of the YOLOv11

Figure 4 Images of the YOLOv11

The main reasons for the satisfactory performance of the YOLOv11 model are the CNN structure and the non-maximum suppression (NMS) feature used by the model. Thanks to the CNN structure, it performs feature extraction by applying special filters to the images, and the NMS feature makes it easier to detect objects close to each other. For these reasons, the model detected people more easily in thermal images and showed high performance.

E. DETR

Table 2 DETR Results and Hyperparameters

AP@0.50	0,264
AR@0.50:0.95	0,430
Epoch	50/100/100/50/34
Batch	8/16/24/24/24
Learning Rate	1e-4/1e-4/2e-4/2e-4/9e-4

Figure 5 Images of the DETR

The main reason for the average performance of the DETR model is that the model is based on the transformer structure. DETR achieves the results that other object detection models achieve with fewer cycles, but with higher cycles. Therefore, training the model is costly compared to other models.

Although it uses a different technique compared to classical approaches, it has been observed that the DETR model can perform human detection in thermal images with a certain success. The use of transformer architecture models in this field is not common, but the performance of DETR shows that transformer architecture models can be an alternative in this field.

The advantages of the DETR model are that it can infer between objects at every point of an image with its general context detection feature, solve complex object relations with attention mechanisms, predict both the location and class of the object at the same time, and provide better performance with big data. However, the disadvantages are the high computational cost of the model, the need for high quality data and the long training time.

F. Comparative Analysis

When the outputs in the results section are analysed, it is seen that the YOLOv11 model gives better results than the DETR model in human detection from thermal images. The technology used by the models, the number of cycles, training parameters, and the size of the data set are effective in the emergence of these results.

The YOLOv11 model is ahead of DETR with its ability to perform better with a more limited data set, to have lower computational and training costs, and to extract and analyse the features of images with CNN architecture. DETR has brought a new perspective to the field of image processing with its capabilities such as attention mechanisms and general context analysis. As seen in this study, YOLOv11 model is more advantageous than DETR in human detection in thermal images under today's conditions. However, the advanced techniques offered by the DETR model can be used more easily and perform more effectively with advanced systems in the future. Therefore, in this study, it is shown that the DETR model is also an alternative in this field and studies on this model are encouraged.

IV. Conclusion and Further Work

In this study, the human detection performance of two object detection models with different architectures on thermal images is compared. The first of these models, YOLOv11, outperformed the other model, transformer-based DETR, thanks to its CNN-based architecture. The feature extraction capability of the CNN architecture, which is the basis of the YOLOv11 model, and the long training time of the DETR model were effective in the difference in the results of the performance of the models.

While both models had no problem in detecting people at medium and close distances, they had difficulty in detecting small human figures at far distances. One reason for this is the small number of small human figures in the dataset, but a more important reason is that it is difficult to recognise small human figures and train the models in this way. In order to overcome this problem, such data can be added to the dataset or the parameters of the models can be changed to detect objects at a distance.

In order to increase the performance of the models, various and high quality data can be added to the training data set. In addition, high parameterised model versions can be used depending on the expanded data set. These changes will be effective especially in increasing the performance of the YOLOv11 model. In order to increase the performance of the DETR model, the number of cycles should be increased. It is expected that the DETR model will show high performance like the YOLOv11 model if the necessary time and energy is given.

In future studies, the appropriate ones of these models can be applied to the specific data set by transfer learning method and can be used for more specific areas. In addition, thermal images obtained from image sources such as UAVs, security cameras and satellites can be processed instantaneously and human detection can be provided.

References

[1] M. Ivašić-Kos, M. Krišto, and M. Pobar, "Human detection in thermal imaging using yolo," in *Proceedings of the 2019 5th International Conference on Computer and Technology Applications*, 2019, pp. 20–24.

[2] N. Shahid, G.-H. Yu, T. D. Trinh, D.-S. Sin, and J.-Y. Kim, "Real-time implementation of human detection in thermal imagery based on cnn," , vol. 17, no. 1, pp. 107–121, 2019.

[3] T. Guo, C. P. Huynh, and M. Solh, "Domain-adaptive pedestrian detection in thermal images," in *2019 IEEE international conference on image processing (ICIP)*. IEEE, 2019, pp. 1660–1664.

[4] K. Akshatha, A. K. Karunakar, S. B. Shenoy, A. K. Pai, N. H. Nagaraj, and S. S. Rohatgi, "Human detection in aerial thermal images using faster r-cnn and ssd algorithms," *Electronics*, vol. 11, no. 7, p. 1151, 2022.

[5] T.-D. Do, N.-N. Truong, and M.-H. Le, "Real-time human detection in fire scenarios using infrared and thermal imaging fusion," *arXiv preprint arXiv:2307.04223*, 2023.

[6] P.-F. Tsai, C.-H. Liao, and S.-M. Yuan, "Using deep learning with thermal imaging for human detection in heavy smoke scenarios," *Sensors*, vol. 22, no. 14, p. 5351, 2022.

[7] J. Portmann, S. Lynen, M. Chli, and R. Siegwart, "People detection and tracking from aerial thermal views," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 1794–1800.