Optimierungen von Objekterkennungsmodellen in IR und RGB-Bildern

Objekterkennungsmodellenoptimierungen unter verschiedenen Umweltbedingungen

Bachelorarbeit

ausgearbeitet von

Ege Çağdaş ALADAĞ

zur Erlangung des akademischen Grades
Bachelor of Science (B.Sc.)

vorgelegt an der

Türkisch-Deutschen Universität
Fakultät für Ingenieurwissenschaften

im Studiengang
Informatik

Betreuer/in: Asst. Prof. Dilek Göksel DURU
Türkisch-Deutsche Universität

Istanbul, im August 2024

Inhaltsverzeichnis

Kurzfassung

Fügen Sie hier die Kurzfassung Ihrer Arbeit, welche bestenfalls strukturiert sein sollte, z.B. Einleitung, Hintergrund, Problemstellung, Zielsetzung, Vorgehen/Methode, Ergebnis, Fazit.

Abstract

Hier folgt die Kurzfassung auf Englisch.

1 Einführung und Aufbau

Object detection is a technology related to computer vision and image processing that deals with detecting instances of objects of a certain class (such as humans, cars, animals and drones) in digital images and videos. It has roots dating back to 1990s, although there has been major leaps in techniques and algorithms over the years. In the relatively recent years, deep learning methods have been prevalent in the object detection technologies.

In the rapidly advancing field of cameras and computer vision, the development of robust object recognition models is essential for applications ranging from autonomous systems to surveillance and beyond in many different fields. As technology continues to evolve, the integration of different forms of imaging has become a key focus to enhance the adaptability and reliability of these models. Visible images are affected by environmental and illumination variations such as low lighting and sun glare; meanwhile thermal and infrared images are noisy and have low resolution. [Bustos et al., 2023, p.1] The main advantage of thermal and infrared imagery is that they are not affected by light conditions, thus they can see objects that would otherwise be very difficult or even impossible to see with visible imagery.

The increasing usage and market size of infrared cameras and imagery (see Figure 1.1) and AI-based object detection continuously require better optimized and well performing models, especially in difficult environmental conditions such as rainy, foggy weather and high or low temperatures. Improvements to these detection methods and systems can have benefits extending into fields such as autonomous vehicles, agriculture, smart cities, search and rescue operations, public safety, security and military.

Object detection in the visible spectrum is has seen a lot of interest and progress throughout the history of object detection. Deep learning methods have been developed within the past decade, that have continued to bring faster and more accurate detection performances. Some of the most prominent methods and algorithms currently used in object detection can be named as; R-CNN [Girshick et al., 2014] and it's variants such as Fast R-CNN [Girshick, 2015] and Faster R-CNN [**?**], You Only Look Once(YOLO) [Redmon et al., 2016], Single Shot Multibox Detector(SSD) [Liu et al., 2016], RetinaNet [Lin et al., 2018], EfficientDet [Tan et al., 2020].

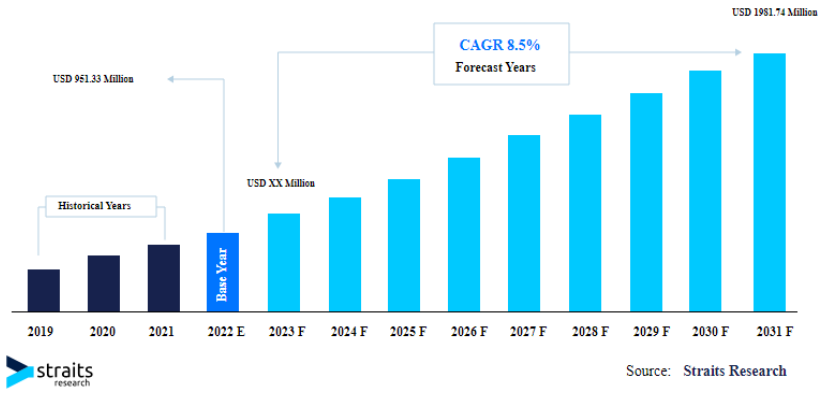

Abbildung 1.1: Infrared Camera Market Size [Straits Research, 2022]

The IR spectrum, on the other hand, is a relatively newer field in the context of object detection. Though it has been explored a lot, there really is no limit to the performance that may be achieved with further development. There has been works so far that have attempted to utilise the fusion of IR-RGB images to achieve better detection performances such as Wang et al. [2022]

2 Grundlagen

2.1 Digital Imagery

Digital imagery refers to visual content in digital form, that can be recognized and displayed by computers. For the purposes of this paper, we need to make a distinction between the following image types.

- Visible Spectrum(RGB) Imagery
- Thermal(Infrared) Imagery

2.1.1 Visible Spectrum(RGB) Imagery

NASA Science Mission Directorate [2010] defines the visible light spectrum as the part of the electromagnetic spectrum visible to the human eye, ranging from approximately 380 to 700 nanometers in wavelength. This range encompasses all the colors perceivable by the human eye, from violet to red.

2.1.2 Thermal(Infrared, IR) Imagery

SPI Corp [2014] defines thermal imaging, or thermography as the detection and measuring of radiation in the infrared spectrum being emitted from an object with the use of thermographic cameras. This type of imagery can collect temperature data from its field of view and display it using a variety of color palettes. Each pixel in a thermal image represents a temperature data point, and these data points are assigned a unique color or shade based on their value, meaning that as the thermal sensor detects changes in heat energy, it will express this change by adjusting the color or shade of a pixel. [Teledyne FLIR, 2021]

2.2 Machine Learning (ML)

> *"The studies reported here have been concerned with the programming of a digital computer to behave in a way which, if done by human beings or animals, would be described as involving the process of learning."* [Samuel, 1959]

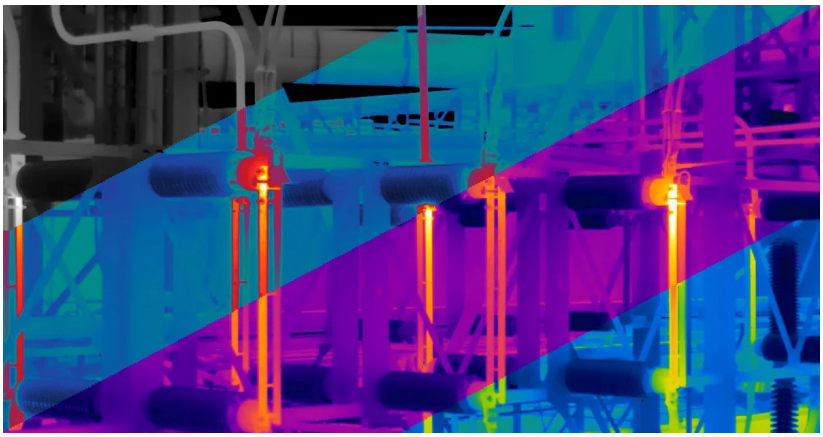

Abbildung 2.1: Thermal Color Palettes [Teledyne FLIR, 2021]

Machine Learning is recognized to be a term originally coined by Arthur L. Samuel in Samuel [1959]. It refers to the behavior of a computer to learn by processing data, improving it's performance on certain tasks by working through a dataset and adjusting the algorithms to achieve better results. The idea is to require minimal human input while still performing well over a previously unseen set of data.

For a task to be fit for a machine learning application, a definite goal must exist, and at least one criterion or intermediate goal must exist which has a bearing on the achievement of the final goal and for which the sign should be known. [Samuel, 1959] Today, machine learning systems are used to identify objects in images, transcribe speech into text, match news items, posts or products with users' interests, and select relevant results of search. [LeCun et al., 2015]

Machine learning algorithms essentially work by processing training data and refining certain parameters to increase performance. This happens through iterations on the data and adjusting parameters by increments. The types of machine learning can be defined as so:

Supervised Learning is a very common method of machine learning, where the training data contains the expected output of an ML model. The training algorithm can then compute an objective function that measures the error (or distance) between the output scores and the desired pattern of scores [LeCun et al., 2015] and adjust the feature weights to improve the result by minimizing the error. The goal is to make correct predictions for new, unseen data.

Unsupervised Learning in contrast, does not contain any specific expected output within the training data. Thus the aim of the training is to recognize patterns within the dataset and categorize the data, rather than predicting an accurate label.

Reinforcement Learning algorithms that mimics how humans learn by trial and error. These algorithms use a reward and punishment paradigm, where the actions that work towards the goal are reinforcend and the actions that detract from the goal are punished or ignored. This feedback loop allows the algorithm to reach the goal by learning what is good and what is bad.

2.2.1 Deep Learning

A deep-learning architecture is a multilayer stack of simple modules, all (or most) of which are subject to learning, and many of which compute non-linear input–output mappings. [LeCun et al., 2015] The multilayered structure allows automatic feature extraction from the data, which normally would need to be done manually and that often is a difficult task. Deep learning has the same fields of application as machine learning.

2.2.2 Object Detection

Object detection in digital images involves identifying and localizing objects within an image. This process not only recognizes the presence of objects but also depicts their precise boundaries, typically using bounding boxes. Unlike humans; computers can not easily identify objects inside an image and classify them as to what kind of object they are, whether that is an animal, a vehicle, a human or something else that such images could contain.

Object detection in digital images has been its own field of study for a number of decades. The underlying algorithms for object detection have evolved significantly over the years, from traditional methods like sliding windows and handcrafted features to modern deep learning approaches. These deep learning models are trained on large datasets annotated with object classes and bounding boxes. During training, the network learns to minimize a loss function that penalizes incorrect classifications and inaccurate bounding boxes.

2.2.3 Traditional Object Detection Methods

Feature Selection by Hand

Early object detection algorithms relied on features such as Haar-like features [Viola and Jones, 2001], Scale-Invariant Feature Transform (SIFT) [Lowe, 2004], and Histogram of Oriented Gradients (HOG) [Dalal and Triggs, 2005]. These features are designed to capture relevant visual information about objects, which are then fed into classifiers like Support Vector Machines (SVMs) [Cristianini and Ricci, 2008] or Deci-

sion Trees [Breiman et al., 1984] to detect objects. However, these methods struggle with variations in object appearance and background clutter.

Sliding Window Approach

One of the earliest techniques, this method involves moving a fixed-size window across the image and applying a classifier at each position to detect objects. Despite its simplicity, this approach is computationally intensive and often inefficient because it requires the classifier to be evaluated at every possible location and scale within the image. NT et al. [2024] utilizes the sliding window approach to object detection and classification for driver assistance systems.

2.2.4 Modern Deep Learning Approaches to Object Detection

Convolutional Neural Networks (CNN)

Convolutional Neural Networks' introduction to the object detection field has been revolutionary. CNNs are deep learning models that automatically learn feature representations from the data. CNN layers can progressively extract more complex features, from edges and boundaries in early layers to parts of objects and complete objects in the later layers.

Region-Based Convolutional Neural Networks (R-CNN)

Region-Based Convolutional Neural Networks (R-CNN) have built upon the groundwork of CNNs by using selective search to propose regions likely to contain objects. Thus it reduces the number of locations to be checked. Each of these regions are then fed into a CNN.

There are R-CNN variants that have further optimized this process to achieve better speed and accuracy such as Fast R-CNN [Girshick, 2015] and Faster R-CNN [Ren et al., 2016]

You Only Look Once (YOLO)

YOLO [Redmon et al., 2016] models use a different approach by dividing the image into a grid and predicting bounding boxes and class probabilities directly for each grid cell in a single network pass. This architecture allows YOLO to achieve real-time object detection with greater speed. YOLO models may struggle with small or closely packed objects because of the way they are designed.

Single Shot MultiBox Detector (SSD)

Similar to YOLO, SSD predicts bounding boxes and class scores directly from feature maps. However, SSD uses multiple feature maps at different scales, allowing it to handle objects of various sizes more effectively.

3 Stand Der Technik

4 Modell-Optimerung

4.1 Purpose

The purpose of optimizations in the context of object detection is to achieve better overall performance in terms of prediction accuracy and speed, with the ultimate goal of defining a good set of standards that can provide future model creators a helpful guideline to base their models off of.

4.2 Goals

We will work on optimizing models in three different metrics:

- Accuracy
- Speed
- Accuracy & Speed (with Thresholds)

4.2.1 Accuracy

In object detection, a model's accuracy refers to how well the model can correctly identify and locate objects within an image. Accuracy in object detection can be broken down into several key metrics:

Precision is the fraction of true positive detections (correctly identified objects) among all detections made by the model. Higher precision means the model better distinguishes true objects from false positives.

Recall, or Sensitivity , is the proportion of true positive detections to all actual objects in the image. High recall means that the model is good at finding most of the objects present in the image.

Intersection over Union (IoU) measures how well the predicted bounding box overlaps with the ground truth bounding box. It is calculated as the area of overlap divided by the area of union of the predicted and ground truth boxes.

Average Precision(AP) is a combined measure that takes both precision and recall into account, often used for evaluating the performance across different confidence thresholds.

Mean Average Precision (mAP) is the mean of Average Precision across different classes, providing a single metric to evaluate the overall accuracy of the model.

Precision: The proportion of true positive detections (correctly identified objects) out of all detections made by the model. High precision means that when the model predicts an object, it is likely to be correct.

Recall: The proportion of true positive detections out of all actual objects in the image. High recall means that the model is good at finding most of the objects present in the image.

Intersection over Union (IoU): A measure of how well the predicted bounding box overlaps with the ground truth bounding box. It is calculated as the area of overlap divided by the area of union of the predicted and ground truth boxes.

Average Precision (AP): A combined measure that takes both precision and recall into account, often used for evaluating the performance across different confidence thresholds. Mean Average Precision (mAP): The mean of Average Precision across different classes, providing a single metric to evaluate the overall accuracy of the model.

4.3 Methodology

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and Regression Trees*. Taylor & Francis, 1984. ISBN 9780412048418. URL https://books.google.com.tr/books?id=JwQx-WOmSyQC.

N. Bustos, M. Mashhadi, S. K. Lai-Yuen, S. Sarkar, and T. K. Das. A systematic literature review on object detection using near infrared and thermal images. *Neurocomputing*, 560:126804, 2023. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.126804. URL https://www.sciencedirect.com/science/article/pii/S092523122300927X.

N. Cristianini and E. Ricci. *Support Vector Machines*, pages 928–932. Springer US, Boston, MA, 2008. ISBN 978-0-387-30162-4. doi: 10.1007/978-0-387-30162-4_415. URL https://doi.org/10.1007/978-0-387-30162-4_415.

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'05)*, volume 1, pages 886–893 vol. 1, 2005. doi: 10.1109/CVPR.2005.177.

R. Girshick. Fast r-cnn. 2015.

R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. 2018.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. *SSD: Single Shot MultiBox Detector*, page 21–37. Springer International Publishing, 2016. ISBN 9783319464480. doi: 10.1007/978-3-319-46448-0_2. URL http://dx.doi.org/10.1007/978-3-319-46448-0_2.

D. G. Lowe. Distinctive image features from scale-invariant keypoints. volume 60, pages 91–110, Nov 2004. doi: 10.1023/B:VISI.0000029664.99615.94. URL https://doi.org/10.1023/B:VISI.0000029664.99615.94.

NASA Science Mission Directorate. Visible light, June 2010. URL https://science.nasa.gov/ems/09_visiblelight. Last accessed 14 May 2024.

S. K. NT, S. Yadav, and P. Rajalakshmi. A sliding window technique based radar and camera fusion model for object detection in adverse weather condition. *IEEE Sensors Letters*, pages 1–4, 2024. doi: 10.1109/LSENS.2024.3401233.

J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. 2016.

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. 2016.

A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959. doi: 10.1147/rd.33.0210.

SPI Corp. What is thermal imaging?, 2014. URL https://www.x20.org/knowledgebase/what-is-thermal-imaging/. Last accessed 15 May 2024.

Straits Research. Infrared camera market size, share & trends analysis report by technology (cooled ir camera, uncooled ir camera), by end-user (defense and military, industrial, commercial surveillance, automotive, bfsi, healthcare, residential, others) and by region(north america, europe, apac, middle east and africa, latam) forecasts, 2023-2031, 2022. URL https://straitsresearch.com/report/infrared-camera-market. Last accessed 30 April 2024.

M. Tan, R. Pang, and Q. V. Le. Efficientdet: Scalable and efficient object detection. 2020.

Teledyne FLIR. Picking a thermal color palette, Apr. 2021. URL https://www.flir.com/discover/industrial/picking-a-thermal-color-palette/. Last accessed 15 May 2024.

P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. 1:I–I, 2001. doi: 10.1109/CVPR.2001.990517.

Q. Wang, Y. Chi, T. Shen, J. Song, Z. Zhang, and Y. Zhu. Improving rgb-infrared object detection by reducing cross-modality redundancy. *Remote Sensing*, 14(9), 2022. ISSN 2072-4292. doi: 10.3390/rs14092020. URL https://www.mdpi.com/2072-4292/14/9/2020.

Eidesstattliche Erklärung

Ich versichere, die von mir vorgelegte Arbeit selbständig verfasst zu haben.

Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer entnommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben.

Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.

Istanbul, 17. Juni 2024

Max Mustermann