NTNU
Kunnskap for en bedre verden

Department of Mechanical and Industrial Engineering

Master's Thesis - robotics & automation

Efficient, accurate, and privacy-preservant object detection in edge devices

Student:
Hallvard Enger Bjørgen

Supervisor at NTNU:
Amund Skavhaug

Trondheim/Esbjerg Spring 2024

Table of Contents

1 Introduction

On-device processing is emerging as a vital component of modern human detection and tracking systems as an approach to ensure privacy. The ability to detect and track humans in real-time is crucial for a wide range of applications, from security surveillance to visitor analytics in cultural institutions. However, the deployment of such systems raises concerns about privacy and data security, particularly in sensitive environments like museums and aquariums. This project aims to develop a privacy-preserving human analytics system that can be deployed in these environments while ensuring visitor privacy and data security.

Faller følgende to avsnitt under bakgrunn/motivasjon, generell introduksjon, eller problembeskrivelse?

The method of human detection and tracking in public spaces has evolved significantly over the past decade, driven by advancements in computer vision and machine learning. Traditional surveillance systems relied on centralized processing, where video feeds were transmitted to a remote server for manual human analysis. However, this approach raised privacy concerns as it involved transmitting raw video data over the network, potentially exposing sensitive information. Additionally, it also required a human to manually analyze the video feed, which was time-consuming, prone to errors, lacking of scalability, and not privacy-preserving. On-device processing addresses this issue by performing analytics locally on the edge device, reducing the need to transmit raw video data and enhancing privacy.

A device was deployed in the acquarium of "Fiskeri og Søfartsmuseet" in Esbjerg, Denmark, to demonstrate the feasibility and effectiveness of on-device human detection and tracking in a practical and realistic setting. On top of the inherent challenges of running the analysis on-device in real-time, the system also faced the challenge of a suboptimal lighting environment, common in acquarium settings. This thesis serves as a demonstration of how to overcome said challenges by the use of a Raspberry Pi 4 with a camera, running a pre-trained yolov9 object detection model. The thesis further explores the effects of adding labeled images from the museum environment to the training dataset to improve the model's performance in the specific setting. The implementation is described in great detail, and the dataset is available at the following link: *TODO add link*.

Inkluderer også noe (mer) om background and motivation, problem description, scope (research questions og research objectives) og struktur av oppgaven...

2 Literature Review

The advent of "modern" object detection has enabled more sophisticated and automated methods for understanding visitor engagement and flow in cultural institutions. This literature review aims to explore the current state of research on object detection and visitor behaviour analysis in cultural institutions, focusing on privacy-preserving techniques, dataset specialization for enhanced object detection accuracy, and case studies of technology implementation.

2.1 Visitor Behavior Analysis in Cultural Institutions

Studies on traditional methods for analyzing visitor behavior (surveys, manual counting, direct observation) and their use cases and limitations.

la2017museumbehaviouranalysis conducted a study on the behaviour of museum visitors, and the perceived value of their findings to museum curators. They found that the use of technology for visitor behaviour analysis was generally well-received by museum curators, and that the data collected could be used to improve the visitor experience. They also avoided the use of manual counting and surveys, by giving the visitors wearable RFID trackers, communicating their position to the system when close to one of several beacons deployed at positions deemed important by the museum curators. The study was able to provide insights into the visitors' behaviour, such as

the most popular exhibits, the average time spent in the museum, and the most common paths taken by visitors. The biggest draw back of such a system is having to give the visitors wearable trackers, which can be unfavourable. TDOO: finne ut om de spurte deltakere om hva de synes om å bha med seg devices.

On the usability of a visitor behavior analysis systems, **la2017museumbehaviouranalysis** found split opinions. Administrators and department heads were generally more enthustistic, while the museum curators were generally more sceptical: "A temporary exhibition won't change after you deploy it, and understanding how it is used by the public would not help me in my next exhibition, since they are each very different. My main reason to analyze behavior would be to satisfy my curiosity." On the contrary, one museum administrator stated: "Some exhibitions cost a fortune. We really need to know if this expenditure was worthwhile" displaying how opinions can vary greatly depending on the role of the individual in the museum.

2.2 Privacy-Preserving Technologies in Surveillance

Exploration of techniques for ensuring privacy in surveillance, such as blurring, pixelating faces etc..

TODO insert chapter from forstudien om hva folk tolker som privacy preservant og ikke. Poengtere at det mest personvernvennlige er å ikke lagre data i det hele tatt.

2.2.1 Privacy and Ethics

Research addressing the ethical implications of surveillance in public spaces, including visitor perceptions and legal considerations.

2.3 Introduction to Object Detection and Visitor Tracking

Review of the evolution of object detection and tracking technologies, including the transition from traditional methods to modern computer vision techniques. What are some of their challenges, and how have they been addressed? Specifically focusing on suboptimal lighting conditions.

Har fått yolov9 oppe og kjøre på pcen så tenker å kortfattet oppsummere hvordan yolo-algoritmen er ulik noen andre algoritmer, hva yolo er brukt til, og hvordan v9 skiller seg fra tidligere versjoner. Ta for seg hva vekter er i denne konteksten og at vektene bestemmes av treningsdata, som gir en glidende overgang til 2.3.3

2.3.1 Yolov9 Object Detection

The YOLO (You Only Look Once) object detection algorithm is a popular choice for real-time object detection due to its speed and accuracy. The YOLO algorithm processes images in a single pass, making it faster than traditional object detection algorithms that require multiple passes. YOLO divides the image into a grid and predicts bounding boxes and class probabilities for each grid cell. The YOLOv9 model is an improved version of the original YOLO algorithm, incorporating various enhancements to improve detection accuracy and speed. YOLOv9 uses a pre-trained model that has been trained on a large dataset to detect a wide range of objects. The model's weights are determined by the training data, allowing it to learn to detect objects with high accuracy.

2.3.2 Dark-Lit Environments

Being able to detect and locate people in dark-lit environments have been previously attempted, usually for security concerns in public spaces. **pa2020PersonDetectionNightTimeFLIR**

developed a system for detecting people in dark-lit environments using a convolutional neural network. They modified the three input channels which usually take RGB to take as input instead (i) the original infrared image, (ii) a difference image from the previous frame, and (iii) a background subtraction mask. Their dataset is vastly different from the setting for this thesis, as the individuals in their photos were far away from the cameras. However, they found that their system was able to detect people in dark-lit environments with an accuracy of 90%. This is a promising result, as it shows that it is possible to detect people in dark-lit environments using infrared imaging and CNNs. They used FLIR cameras, which make images from heat. Doing inferences on pure infrared images may be harder because the infrared radiation may be less prevalent than the heat a FLIR camera may capture. The FLIR cameras are expensive, and thus not considered viable for this project. **pa2020PersonDetectionNightTimeFLIR**. hei

2.3.3 Effectiveness of Training Dataset Specialization

Comparison of various algorithms performance enhancement when data has been optimized for the use case. What is the role of a training dataset in the task of determining the weights in a yolov9 artificial neural network, and how can would special training data optimize the weights for a specific use case?

2.4 Summary of Literature Review

A summarization of the current state of research, and where my thesis aims to contribute.

3 Methodology

This chapter will describe the process of developing the human detection and positioning system, including the hardware and software components used, the dataset collection and preparation (labeling), the training of the object detection model, and the deployment of the system in the museum environment. The chapter will also discuss the ways of assessing the system's performance, before and after deployment (hhv using test dataset og vanskelighetene ved å ikke enkelt kunne vite hva som er korrekt).

Component	Description	Quantity
Raspberry Pi 4	4GB RAM, 1.5GHz quad-core ARM Cortex-A72 CPU	1
Raspberry Pi Camera Module V2	8MP, 1080p video at 30fps	1
MicroSD Card	32GB	1
Power Supply	5V 3A USB-C	1
Modem	Huawei E3372-325 4G LTE	1

Table 1: Hardware setup for the system.

3.1 Project

Two cameras were deployed in a room of acquariums at "Fiskeri- og Søfartsmuseet" in Esbjerg to take images for building a specialized dataset and to evaluate the effects of developing a highly specialized detector rather than using a general.

3.1.1 Dataset Collection

The dataset "acquarium man" was built by capturing images while no unconsenting visitors were present in the acquarium. Due to the constraint to only operate within opening hours when the facility was open to everyone, a way to cancel image capturing was needed in the case if someone entered the room. One of the goals of the dataset was to have the images taken from the same angle as the device will be used in the future. The device was therefore mounted in the corner of the room, and ssh was used to access the device remotely from a pc in the acquarium. Then, a script was ran to capture an image per second and storing it on the SD card in the device. The choice to store the image locally rather than transmitting it was to not have to worry about data transmission costs and issues. The first iteration of image capture was made with non-optimized camera configurations, turning to the brightness setting of the picamera python module to get the images to a sufficient brightness. Still, some images fell short to the auto focus of the camera focusing on the bright acquariums and thus rendering the rest of the image rather dark. However, these images serve as a way of inspecting the impact of captured image quality on inference performance. Camera setup: Hardware camera can be tweaked by screwing the lens with a mechanical tool to modify it's aperture, which influences it's depth focus. Aperture mechanical setting (camera focus adjustment), depth control. Default not found... images well sharp enough... 50cm to infinity...

Due to many technical difficulties the first few times images were being captured for the dataset, only the developer and author of this thesis is present in the images[1]. The acquarium could not provide free tickets for additional participants in the experiments, and the cost of bringing friends to enrichen the dataset would then have to be optimized in terms of time spent at the site.

The detector needs to know the ground truth when training and validating on the obtained data. This can be obtained by manually labeling the data. However, a more scientific, robust and scaleable way of labeling is to have the detector do the heavy lifting. Therefore, the dataset was first inferenced by the detector. The data was then manually validated, and finally the images for which the detector did not find any persons was manually labeled. The time to understand the tools and develop this pipeline was similar to what it would take to manually label all the images, but can now be used for future applications also. The approach to label the images is described in section: 3.2.

After the ground truth was identified for all the dataset images, it was then used to evaluate the general-purpose yolov3 and yolov9. Results are discussed in 4. The data was then used to train the detectors. The training process is described in section 6.1. The trained models were then deployed to the device to evaluate the inference speed and accuracy. The deployment process is described in section 6.2.

To visualize the improvements and highlight the areas in the image benefitting most from detector improvements, heat maps were generated. The process of generating heat maps is described in section 4.

3.2 Labeling

Ground truth values for the dataset must be obtained before improving the model and evaluating it's effect.

"Label Studio" was used to label the images. First, the dataset was predicted with a yolo v9 model trained on the COCO dataset. Predictions were mostly decent, but some needed small tweakings and in some cases the persons were not discovered. The detector had close to zero hallucinations due to a sufficiently high confidence rate of 0.5 (: for), but in some cases the fish were identified

[1]Initially, an attempt was made to pass MQTT messages as a way to initialize image capture so multiple cameras could be deployed in several locations, thus speeding up and simplifying the image capturing process. This was discarded due to technical difficulties related to efficiently stopping the image capturing. For a different solution, a recommendation would be to implement a way of communicating to multiple devices however, so one could obtain all images from deployment locations in one shooting. For this single-deployment angle and area project, however, the approach with ssh-ing into the device worked fine.

as human.

3.3 Ethical Considerations

In the deployment of advanced machine learning technologies for visitor localization and engagement analysis, this research proactively addresses privacy concerns through the implementation of image obscuration techniques. These methods ensure that no personally identifiable information is captured or communicated, thus significantly reducing privacy risks associated with visitor tracking in cultural spaces such as museums and aquariums.

3.3.1 Privacy by Design

At the forefront of our ethical approach is the principle of "privacy by design." This concept involves integrating privacy into the development and operation of our tracking technologies from the outset, rather than as an afterthought. By employing image obscuration techniques, such as real-time pixelation or silhouette generation, we ensure that the visual data processed by our system remains anonymous. This method effectively eliminates the possibility of identifying individual visitors from the captured data, thereby safeguarding their privacy.

The application of these privacy-preserving techniques negates the need for explicit consent from visitors for two primary reasons. First, the anonymization process occurs instantaneously as the data is captured, meaning no identifiable information is ever stored or analyzed. Second, the focus of the research is on aggregate behavior patterns rather than individual actions, further distancing the study from privacy concerns.

3.3.2 Ethical Use and Data Protection

Ensuring the ethical use of technology extends beyond privacy considerations to include the responsible handling and protection of any data generated by the system. Although the data is anonymized, we are committed to maintaining high standards of data protection. This includes secure data storage, limiting access to authorized personnel, and employing robust data management policies that comply with relevant data protection laws and guidelines.

The utilization of anonymization techniques also reflects our commitment to minimizing any potential impact on visitor behavior and the overall museum or aquarium experience. By ensuring that the tracking system is unobtrusive and does not compromise privacy, we aim to maintain the integrity of the visitor experience, allowing individuals to engage with exhibits without concern for their privacy.

3.3.3 Transparency and Accountability

While the technical approach effectively addresses privacy concerns, maintaining transparency about the use and purpose of tracking technologies is still essential. Information about the tracking system and its privacy-preserving nature will be made available to visitors, ensuring they are informed about how data is used to enhance the visitor experience.

Furthermore, the project will adhere to an ongoing ethical review process, ensuring that all aspects of the research remain aligned with ethical best practices and respond to evolving technological and societal standards.

In summary, by prioritizing privacy through the use of image obscuration techniques and adopting a comprehensive ethical framework, this research aims to advance the understanding of visitor engagement in a manner that is both innovative and respectful of individual privacy rights. This approach sets a precedent for the ethical application of machine learning technologies in cultural

institutions, balancing the benefits of visitor behavior analysis with the imperative of protecting privacy.

Notes

Tried to download/use model from Roboflow, but either image has to be sent to an API which would not retain privacy, or the device has to host an API itself to run the inference... Seems unlikely to be the most preferable solution, as the device would have to set up the service and run it locally. Possibly an interesting solution would be to do this with multiple devices. This supports the master-slave pattern of having multiple weaker computers and have them send to the stronger unit. Setting up private TCP connection between the weaker units and the strong unit and have the images sent to the stronger, so it can detect on them and send information etc... How many weak units do we need in order to make it profitable to have a strong GPU unit to do the processing? This whole systems sounds to be complicating processes, not making the product modular and easy-to-use. Includes a lot of connection/networking to make the weaker units find and connect to strong, physically close device. This task would mean setting up a strong device to host a network to which the weak units might connect to, and send images to. The issue is whenever images are sent, a lot of transmission is used... But the model takes image input size of 416x416. Would it be similar to just downscale the image before sending, or would this give the model less detail to work with?

Will now run several models on datasets from the web, i.e. the CrowdHuman dataset to see their accuracies. Will then deploy the models to device in acquarium to see if the best-performing model is an option in terms of size and inference speed. If it is preferable, I will attempt to increase it's accuracy by accumulating and annotating a specialized dataset for that setting, and training the final layers on the data. Can this be done with a

4 Results

This chapter will present the results of the human detection and tracking system, including the system's performance in the museum environment, the effects of adding labeled images from the museum environment to the training dataset, and the system's ability to detect and track humans in real-time.

5 Discussion

This chapter will discuss the implications of the results, summarizing the results and their significance for the development of similar systems, and whether or not the approach in this thesis is a viable solution for the presented problem.

6 Conclusion

Summarization of the thesis and its contributions to the field.

6.1 Recommendations

If I were to do it all over, what would I do differently and why? What are the key takeaways from this project, and what recommendations do I have for anyone looking to do similar work?

6.2 Ethical Implications

What are the ethical implications of the development of automated visual tracking?

6.3 Future Work

Recommendations for future work.