Cerberus: Privacy-Preserving Crowd Counting and Localisation using Face Detection in Edge Devices

Justas Brazauskas, Chris Jensen, Matthew Danish, Ian Lewis, Richard Mortier
University of Cambridge first.last@cl.cam.ac.uk

Abstract
Cerberus uses face detection in edge devices to perform privacy-preserving crowd counting and localisation. We describe its deployment in a university setting where ceiling-mounted cameras perform real-time face detection to report occupied seats without storing or transmitting images. Cerberus’ aim is ultimately to integrate with digital twins over a LoRa network enabling data visualisation and support applications in building informatics, while balancing data accuracy and individual privacy. The paper describes the system’s design, deployment, and potential for broader urban informatics applications, highlighting its effectiveness in privacy-preserving crowd monitoring.

ACM Reference Format:

1 Introduction
There is growing demand for systems that monitor building space utilisation and track occupancy, due both to the desire for improved indoor air quality monitoring following the COVID-19 pandemic, and more traditional requirements such as managing emergency building evacuation. Object detection on video streams is a commonly proposed technique but increasing sensitivity to privacy concerns makes such systems problematic to deploy widely. Cerberus is a privacy-preserving ensemble face detection system targeting edge devices that addresses these concerns. It offers crowd counting and localisation far more cheaply than traditional methods such as under-desk mounted PIR sensors, while avoiding the need to manage hundreds of deployed sensors and having far fewer failure points.

In this paper we describe a Cerberus deployment into a university lecture theatre of 264 seats, comprising three edge devices each controlling a ceiling-mounted camera angled towards the audience, and performance local face detection on the video stream from that camera using ensemble machine learning methods. Following face detection, device transmits the list of occupied seats to a central node, the Cerberus Crowd Count Node (C3Node) that asynchronously relays this data to our server. The system is privacy preserving as captured images never leave the system; the only output from each device is a list of strings encoding seat occupancy in the lecture theatre.

Cerberus is designed for integration with a digital twin, creating a real-time virtual copy of the physical lecture theatre supporting more complex in-building informatics applications and better data visualisation, improving legibility for room occupants (Figure 2).

In this paper we provide an overview of Cerberus, detailing its design, development, and the face detection techniques it employs. We further discuss its broader applications and potential in the evolving landscape of urban informatics and privacy-preserving data collection.

2 Related Work
Gao et al. [12] categorize crowd counting technologies as either computer vision-based (CV) or Wi-Fi/BLE-based via (e.g.) Channel State Information (CSI) measurement or MAC address tracking by packet sniffing. Computer vision-based approaches are further...
subdivided into density, object, and people estimation techniques. We focus on density and people estimation as we are primarily concerned with people localisation.

CV methods can also be segmented based whether data is processed remotely at cloud servers, or locally in edge devices. The former rely on centralised computation and offer robust computational resources but increase latency and raise privacy concerns when used for localisation existing [2, 15]. Edge device processing use lower-powered computers (e.g., Raspberry Pi) to analyse data directly at the source, reducing latency and network bandwidth requirements and privacy concerns, while potentially increasing the time taken to process each individual frame [3, 4, 6, 16].

Singh et al. [21] review privacy preservation in crowd monitoring systems, highlighting significant gaps in current research regarding CV-based privacy-sensitive systems. Among the 28 papers reviewed that utilise CV for crowd monitoring, only 10% consider privacy implications. While Wi-Fi or BLE packet sniffing methods are suggested as more privacy-compliant [10, 13], privacy challenges remain due to regulations like the GDPR [11]. Several papers similarly underscore the necessity for privacy-conscious CV solutions [17, 19] but they lack the localisation aspect essential for many in-building applications. Similarly, projects like the privacy-preserving person-tracking DeepDish camera [7, 8] or the “Occupancy-as-a-Service” Databox implementation [24] offer privacy-preserving measures on the edge but also lack localisation.

3 System Design and Implementation

Cerberus combines low-power hardware with lightweight machine-learning software for privacy-preserving data collection indoors. This section emphasises design choices, deployment, and face detection.

3.1 System Design

3.1.1 Hardware. Cerberus utilises Raspberry Pi 4 computers paired with High-Quality cameras, chosen for their resolution of 4056 x 3040 pixels. This high resolution is advantageous for detecting faces over distances exceeding twelve meters. Additionally, an optional Coral [1] Raspberry Pi accelerator can be attached to enhance TensorFlow-based face detection models, significantly improving inference speed beyond what is ordinarily achievable on a Raspberry Pi 4 without overclocking. The Cerberus devices are powered using PoE (Power over Ethernet) dongles (Figure 1).

3.1.2 Software. Our software, designed for devices running Pi OS, leverages the OpenCV library [18] for image data processing, utilizing models such as YuNet [23] and HaarCascade [22] for face detection. Additionally, it incorporates our proprietary model, DiffBoxes, for high-speed seat occupancy detection. DiffBoxes exploits the contrast difference between the seat colour and a person occupying the seat to determine occupancy.

The system’s core functionality includes real-time image analysis from a camera feed and running inference using several models in sequence, allowing to balance the speed of some models, with the accuracy of others. Occupancy is calculated by matching detected faces with seat locations, based on the camera’s field of view and the floorplan.

3.1.3 Node. The Cerberus triple camera system operates as a single-node device, with data being shared across the three devices using the MQTT protocol, which facilitates asynchronous data transmission. One device, which we chose to be the middle camera, serves as the central node. It aggregates data from all cameras, focusing on identifying which seats are occupied and reducing system noise, which results from people in the lecture moving their heads.

3.2 Deployment

In our deployment, cameras are placed on the ceiling, angled at 30 degrees downward (Figure 3). Each camera’s FOV covers approximately one-third of the lecture theatre, minimising blind spots. Furthermore, edge regions for each camera have slight overlaps, ensuring higher accuracy when the data is aggregated at the C3Node level. In our testing, fixed exposure and focus optimised image quality. Metadata for the lecture theatre was created so that every seat was given a unique three-character identifier, similar to those seen in cinemas (Figure 5).
3.2.1 System Calibration. Post-deployment, we used photographs from each camera for system calibration. This process involved mapping the pixel coordinates of detected faces in these photographs to corresponding seat IDs on the theatre’s floorplan. Such calibration was crucial for the face detection models to accurately associate each pixel coordinate in the image with the appropriate seat ID (Figure 5). This mapping method focuses on the location of individuals rather than their identities, aligning with privacy and ethical standards.

Additionally, our DiffBoxes model required precise calibration to function correctly, as it involved creating bounding boxes around the distinct bright blue seats, which seated individuals would later obscure. The model’s inference mechanism relies on detecting colour differences between a reference image of the empty seats and an image taken with people present. This assumes most individuals won’t wear bright blue, thus enabling detection. When attire matches seat colours, face detection algorithms such as Yunet can still ensure detection, though at a cost to inference speed (Figure 8).

![Figure 5. Before (L) and after (R) system calibration. Overexposure was an issue. Mapping the middle camera seats was done after data for face locations was collected and clusters were drawn.](image)

3.3 Ensemble Face Detection

Cerberus’ distinctiveness lies in its ensemble face detection capability and ease of interchanging models. This approach allows the system to balance between detection speed and accuracy, ensuring overall optimal performance.

In total, 13 models were tested for their accuracy and inference time. Some models, like YuNet[23] or RetinaFace [9] demonstrated promising results with minimal adjustments. Other models, like HaarCascade, would have benefited from deeper customisation. The results for these tests are shown in the Results section.

Given the constraints and compatibility with Raspberry Pi 4, Yunet was selected for its suitability, alongside DiffBoxes for its swift performance, enabling a comparison of the device’s inference capabilities at varied speeds. Additionally, HaarCascade was integrated into our final software deployment. Despite its reduced accuracy, it demonstrates the feasibility of interchanging models on the Cerberus cameras.

In our experimentation with images of resolutions lower than the native resolution across various models, it became apparent that the trade-off of enhanced inference speed for significantly reduced accuracy was not justified, considering the distance between the cameras and the faces. Consequently, to achieve faster processing times without compromising accuracy, models like DiffBoxes were developed, which are inherently faster due to their simpler architecture (Figure 8).

4 Data Collection

This section outlines the specific data collection methodologies employed by Cerberus and the analytical techniques used to process and interpret the collected data.

4.1 Data Collection Strategy

The Cerberus data collection strategy aims to accurately and privately track seating positions in a lecture theatre, incorporating several key considerations.

**Face Detection Models:** The system’s architecture incorporates a modular design, featuring an ensemble of interchangeable face detection models. In instances where a specific model underperforms in a given context, it can be promptly replaced by loading an alternative model. Furthermore, the software is designed so that the system can run multiple models in sequence and aggregate their outputs. This approach potentially enhances overall accuracy by strategically utilising each model’s strengths. By mitigating the limitations of individual algorithms, it ensures robust performance across a wide range of scenarios.

**Message types:** Devices operate at their maximum inference capacity and generate two primary types of messages. The first type involves periodic readings sent at set intervals, covering both device “health” reports and updates on the lecture theatre’s status. The second type includes messages triggered by specific events, such as the start or end of lectures and breaks, or other significant changes in crowd activity. This approach helps reduce network traffic, which in the case of having LoRa backhaul is essential.

**Integration with Digital Twins:** The streaming sensor data can be integrated with real-time digital twins, providing a digital representation of the physical lecture theatre. This not only aids in data visualisation but also offers a framework for advanced data analysis and sensor fusion. In this scenario, the lecture theatre has been extensively equipped with a comprehensive range of environmental data sensors, which is significantly enhanced by the addition of crowd localisation.

4.2 C3Node Processing

The raw inference data from face detection models on each camera exhibits noise, primarily due to movements and head tilts of people taking notes. To address this challenge and enhance overall accuracy, particularly in areas with significant overlap among cameras, we developed an additional architecture for preprocessing the data from all three cameras, consolidating the information into a unified output. This methodology capitalises on the ensemble model concept, as well as leveraging data from multiple cameras to achieve improved accuracy.

The difference in raw readings and C3Node readings (Figure 9) illustrates how C3Node can provide less noisy, more easily interpretable data. C3Node records the occupancy status of each seat in the lecture theatre, attributing confidence values to evaluate the probability of occupancy. This is achieved by retaining a memory of the seat’s status over a fixed period, such as 5 minutes. For instance, if the average inference cycle for the YuNet model is 13 seconds (Figure 8), this results in approximately 23 inference cycles within 300 seconds, with each cycle’s outcome for each seat position stored as either 0 or 1. When running multiple models in sequence, such as Yunet and DiffBoxes, we effectively double the number of recorded states for each seat in the retained memory. Alternatively,
employing two slower yet more accurate models, such as Yunet and Yoloface V5, was considered. However, the final model selection aimed to balance the trade-offs between speed and accuracy. The assessment at the node level is made by calculating the mean of these states; if it exceeds a predefined threshold (e.g. 15%), the seat is considered occupied. In our experiments, we implemented a memory duration of 300 seconds and a threshold level of 15%, as it effectively reduced the total fluctuation error in crowd count during lectures to 3%-5%. Ultimately, the duration for which the seat memory array (Figure 4) can be maintained is constrained by the edge system’s RAM capacity and processing speed.

The system is designed modularly, allowing for flexible model selection and parameter configuration, making it suitable for deployment in environments such as classrooms or conference venues for real-time occupancy tracking. The system was designed to be deployed as part of a larger system, with C3Node combining the different angles into a unified localisation system.

![Density plot for long-term lecture theatre usage.](image)

### Figure 6

#### 5 Privacy

Indoor environments pose particular privacy challenges with camera deployment, often eliciting greater concern compared to outdoor settings. Reflecting on insights from the DeepDish paper [7], while an adversary could ascertain the occupancy within a public building’s space, such information is deemed non-sensitive. This claim is based on the assumption that building occupants are pre-informed of ongoing events, such as lectures, which are also inferable from publicly available timetables or alternative data sources like environmental sensor readings (e.g. CO$_2$) in lecture theatres.

In our research, explicit measures were taken to uphold privacy norms. Participants frequenting the lecture theatre were informed about the operational camera system, which was implemented to monitor presence discreetly. Throughout the year-long deployment, no privacy concerns were reported.

A pivotal aspect of employing face detection technology lies in distinguishing it from facial recognition to mitigate privacy fears. Our approach emphasises that individuals are anonymised in the system’s output, appearing as coloured dots on a digital floor plan. This differentiation was further reinforced through visualisations to demonstrate the non-identifiable nature of the data.

Regarding data handling, captured images were transiently stored in RAM before being discarded, ensuring no permanent storage on any device or server. The system’s architecture facilitates secure communication between cameras via encrypted local Wi-Fi, transmitting occupancy data without retaining identifiable imagery. Optional LoRa connectivity was introduced for enhanced security, making it physically infeasible to access the images, thus maintaining privacy through obfuscation.

Cerberus leverages sophisticated data structures for efficient information transmission to the Adaptive City Platform [14], optimizing data processing and facilitating seamless integration into practical applications. The designed structures (Listing 1) support the system’s privacy-preserving objectives while ensuring adaptability for real-world deployment, and are based on the data management techniques described in [5].

#### Listing 1

C3Node’s JSON output. Seat confidence gauges face-seat distance in a photo; model confidence measures prediction certainty.

```json
{
  "acp_id": "cerberus-node-lt1",
  "filepath": "2024/02/07",
  "acp_ts": "1707314444.066",
  "msg_type": "event_lecture_start",
  "acp_type_id": "cerberus-node",
  "payload_cooked": {
    "received_ts": {
      "L": "1707314443.9989824",
      "M": "1707314438.8186474",
      "R": "1707314437.2528498"
    },
    "crowdcount": 3,
    "seat_capacity": 264,
    "seats_occupied": {
      "MF9": {
        "seat_confidence": 0.83,
        "model_confidence": 0.71,
        "seat_id": "MF9"
      },
      "LB7": {
        "seat_confidence": 0.93,
        "model_confidence": 0.86,
        "seat_id": "LB7"
      },
      "RG5": {
        "seat_confidence": 0.97,
        "model_confidence": 0.95,
        "seat_id": "RG5"
      }
    },
    "occupancy_filled": 0.011
  },
  "mqtt_topic": "csn/cerberus-node-lt1"
}
```

### 6 Results

During its year of operation, Cerberus demonstrated robust performance, with its reliability momentarily affected by two hardware malfunctions, quickly resolved upon identification. These issues were attributed to defective PoE dongles.
6.1 Technical Performance

Model Performance and Accuracy: The results show that neural net-based models are vastly superior to traditional face detection approaches like HaarCascade classifiers. Specifically, in our testing environment, models like Yoloface v5 [20] and YuNet [23], showcased high accuracy, adapting seamlessly to various lighting and focus levels, without the need to adjust metaparameters.

![Model Normalized Error Deviation Distribution](image)

**Figure 7.** Normalised error, calculated as the absolute difference between ground truth and detected counts divided by the ground truth, indicates model accuracy; lower values are better. However, all models showed high standard deviation, thus the need for the C3Node data aggregation. HC refers to HaarCascade, FF means Frontal Face, PF means Profile Face.

On-Device Performance: Despite hardware setbacks, the system maintained efficient real-time operations, underlining software resilience and uninterrupted data collection capabilities. The operating temperatures on the Raspberry Pi devices were around 80°C. Nevertheless, in the next iteration of the system we plan to add active cooling.

6.2 Data Analysis

For the Data Analysis section, we present three sample applications of our collected data, focusing on key areas: crowd count, localisation, and density estimation.

1. Crowd Count Plot: Offers quantitative insights into occupancy trends, demonstrating the models’ crowd estimation accuracy. Figure 2 (R) shows the crowd count plot per day.

2. Crowd Localisation Plot: Provides spatial occupancy distributions, revealing movement patterns and seating preferences within the theatre. Figure 2 (L) shows a crowd localisation plot for a particular moment in time that is denoted by the red vertical line on the crowd count plot (R). Dragging the red line to the left and right updates the crowd localisation plot.

3. Long-term KDE Plots: Data collected over longer terms can be visualised as density plots using Kernel Density Estimation (KDE) methods. Such plots are important because they facilitate a nuanced understanding of theatre usage patterns at various time scales (Figure 6).

![Inference Duration Comparison](image)

**Figure 8.** Inference times for each of the cameras. We use a combination of slow and fast models to increase the overall system accuracy.

We believe that such tools highlight the system’s operational effectiveness, as well as its potential use-cases when managing and interpreting complex data for space utilisation and lecture theatre management.

![Cerberus readings collected over a day](image)

**Figure 9.** Cerberus readings collected over a day.

7 Conclusions

Experience with a year-long deployment of Cerberus showed it provides insight into both short and long-term lecture theatre use, and enables enhancement of safety and emergency response strategies.

Its potential use cases extend beyond lecture theatres to sports arenas, cinemas, theatres, and other places where crowds typically face in one direction, and the C3Node architecture can be adapted to work in such places. This enables facility managers to understand how spaces are used, informing decisions on infrastructure improvements, and via analysis of historical data, to identify anomalies that signal potential problems.

A particularly beneficial application of Cerberus would be to enhance safety and emergency response by providing real-time data on space occupancy. This allows tracking of current use state of spaces and subsequent evacuation progress, crucial during emergency planning and execution.

Smart buildings have historically focused on tracking environmental data and energy consumption. Cerberus introduces a low-cost privacy-preserving option to track people in shared spaces, enabling deeper integration with digital twins and broader sensor fusion. We illustrate this in our deployment by demonstrating how changes in environmental temperature data in the lecture theatre correlate with the locations of people seated (Figure 10).
Cerberus achieves all this while preserving privacy, using unique seat identifiers to dissociate from individual occupants and transiently processing all image data locally on-device; images and messages are stored only temporarily in RAM, ensuring that all data is immediately wiped in the event of a power outage, providing an extra layer of privacy protection against adversarial attacks.

There are several avenues for further exploration. We have not yet evaluated various alternate models or the use of the Coral TPU accelerator, which may provide opportunities to enhance inference speed. Cerberus could be deployed into larger spaces which would explore its adaptability and scalability, particularly in understanding the maximum capacity a single node can handle under the current architecture. Although Cerberus prioritises privacy, exploring advanced encryption or differential privacy methods could further protect data. Finally, adding real-time feedback such as overcrowding alerts could increase user interactivity.

In summary, Cerberus is effective at in-building data collection, balancing detailed data collection with user privacy. It strategically places cameras to minimise blind spots and maximise accuracy. Its use of face detection models and clustering techniques provides actionable insights on seating patterns and space utilisation while prioritising privacy by anonymising data. Its ability to integrate with digital twins and other platforms suggests potential for real-time alerts and broader applications in building operational use. Cerberus exemplifies a responsible, efficient approach to in-building informatics, offering valuable lessons for future data collection.

References