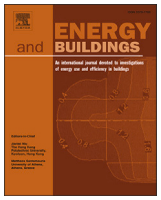
Energy & Buildings 307 (2024) 113936

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enbuild



High-accuracy people counting in large spaces using overhead fisheye cameras ☆

Janusz Konrad [a,*], Mertcan Cokbas [a], Prakash Ishwar [a], Thomas D.C. Little [a], Michael Gevelber [b]

[a] Boston University, Department of Electrical and Computer Engineering, 8 Saint Mary's Street, MA 02215, Boston, USA
[b] Boston University, Department of Mechanical Engineering, 110 Cummington Mall, MA 02215, Boston, USA

ARTICLE INFO

Keywords:
People counting
Occupancy
Energy
HVAC
Spatial analytics
Public safety
Fisheye camera
Deep learning
AI
CNN

ABSTRACT

Quantifying the number of people in various spaces of a commercial building is important for saving energy, optimizing space usage, and assisting with public safety. To accomplish these goals requires obtaining accurate, fine-grained counts in real time. However, existing methodologies are ineffective for covering large areas with high occupancy. We propose an occupancy-sensing system that uses multiple overhead fisheye cameras and state-of-the-art deep-learning algorithms to cover large spaces with high counting accuracy. Tested in a 2,000 ft^2 space, our system shows 54% to 83% reduction in commonly-used error metrics compared to recent people-counting methods proposed for large-space scenarios. Our system is scalable to arbitrarily-large spaces; additional cameras can be integrated with minimal commissioning. We also introduce two new performance metrics for assessing counting accuracy that, unlike common metrics used to date, are independent of occupancy level and can be easily compared across different occupancy scenarios.

1. Introduction

Knowing how many people are in various spaces of a building is critical for saving energy. In 2018, energy consumption in commercial buildings in the United States amounted to $6,787\times10^{15}$ BTUs, with 52% expended on heating, ventilation, and air conditioning (HVAC) [1]. Without adjusting HVAC airflow to match variations in occupancy, a significant fraction of the HVAC energy is wasted due to over-ventilation. There have been a number of studies to estimate the potential energy savings if real-time, fine-grained occupancy data are used to modulate airflow in HVAC systems. Such studies are challenging since they need to account for building size and type (space utilization scenario, hours of occupation, code requirements, and HVAC type), as well as climate zone. Recent estimates of whole-building energy savings utilizing occupancy information range from 10% to 22% [2], while HVAC savings can range from 10% to 40% [3,4].

Beyond energy, occupancy information plays increasingly important role in space management, security and safety. The COVID-19 pandemic has dramatically impacted the office-building market, leading to new office-usage patterns. Some companies opt now for "flexible workspace", where desks are not assigned to employees but can be reserved whenever employees return in-person for work, meetings, etc. Real-time, accurate knowledge of workspace occupancy is essential for an effective implementation of this concept. A similar knowledge of where people are is essential in other industries, such as retail, e.g., the number of people visiting specific locations in a store. Finally, occupancy information is important for security in a building, such as for an emergency situation (e.g., ensuring everyone is accounted for during a fire), or for public safety (e.g., social distancing during a pandemic). Fig. 1 illustrates a potential people-counting deployment scenario.

In the last decade, many approaches have been proposed for occupancy sensing in commercial buildings, from *active* methods that require carrying a cell-phone or swiping an ID card, to *passive-indirect* approaches that use environmental data related to human presence (e.g., CO_2 level, humidity, temperature) and *passive-direct* methods that capture occupants' features such as appearance, movement, body heat, etc.

☆ This research was supported by the Advanced Research Projects Agency - Energy (ARPA-E) through agreement DE-AR0000944 and by Boston University Undergraduate Research Opportunities Program (UROP).

* Corresponding author.
E-mail addresses: jkonrad@bu.edu (J. Konrad), mcokbas@bu.edu (M. Cokbas), pi@bu.edu (P. Ishwar), tdcl@bu.edu (T.D.C. Little), gevelber@bu.edu (M. Gevelber).

https://doi.org/10.1016/j.enbuild.2024.113936
Received 14 June 2023; Received in revised form 29 December 2023; Accepted 20 January 2024
Available online 1 February 2024

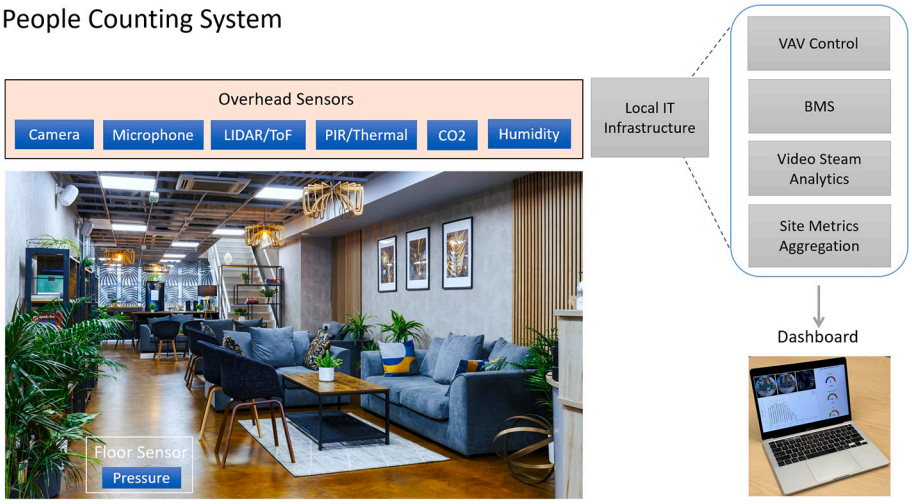


Fig. 1. Overview of indoor people counting using passive sensing with application to HVAC control, spatial analytics and safety/security.

Despite their drawbacks, CO_2-based methods are considered to be the *only* technology today that can deliver HVAC-energy savings in large spaces while maintaining air quality.

In this paper, we offer an alternative to CO_2-based occupancy sensing by leveraging wide-angle, visible-light cameras to detect and count people. We propose COSSY, a *Computational Occupancy Sensing SYstem*, that uses multiple *overhead fisheye* cameras to deliver high-accuracy people counts in large spaces (over 1,000 ft^2). In order to accomplish this, first people need to be detected and then duplicate detections from multiple cameras need to be removed. We leverage state-of-the-art deep-learning algorithms to carry out these two steps. Unlike CO_2-based occupancy sensing, COSSY can simultaneously serve energy-reduction, spatial-analytics and security/safety needs since, in addition to providing accurate counts, it can *localize* people in an open space. COSSY is scalable - it can support spaces of various sizes and shapes. It is easy to deploy using PoE (Power-over-Ethernet) networking; no additional power runs are needed. We assess COSSY performance in various scenarios using two traditional metrics and two new metrics that we propose to facilitate fair comparison across varying levels of occupancy. Finally, we demonstrate experimentally that COSSY significantly outperforms recent people-counting methods proposed for large spaces with high occupancy.

This paper makes the following key contributions:

1. We develop a first-of-its-kind system that collaboratively uses multiple overhead fisheye cameras and state-of-the-art deep-learning algorithms to count people, and we demonstrate its high accuracy in a 2,000 ft^2 space with up to 87 people.
2. We propose two new metrics for validating people-counting methods: Mean-Absolute Error per person and X-Accuracy, that are meaningful across occupancy levels (unlike prior metrics).

2. Related work

As we mentioned in the introduction, *active* occupancy sensing requires carrying a beacon. While it can handle large spaces, it has several shortcomings. Counting people from card swipes is straightforward, but prone to errors when two or more people enter using a single swipe. Such "drift" errors are typical in any tripwire-type system when a *change* in people count rather than the count itself is being estimated; a missed detection can be corrected only by an opposite error (false detection). Counting devices, e.g., cell phones, is not prone to drift errors (the number of devices, not a change in this number, is being estimated), but it relies on each occupant carrying one device only (an unlikely scenario today) and the device being turned on.

The most common *passive-indirect* methods track changes in the environment due to fluctuations in occupancy, for example by measuring CO_2 level, temperature, humidity, etc. Jin et al. [5] developed a physics-based model that relates CO_2 concentration and temperature to occupancy, and demonstrated its excellent performance in a 140 ft^2 room with up to 7 people. It is unclear how the model would perform in a large space with high occupancy. Elkhoukhi et al. [6] have combined humidity measurements with those of CO_2 and temperature to jointly estimate occupancy using machine-learning algorithms, but have tested the approach only in scenarios with at most 8 occupants. Szczurek et al. [7] applied a machine-learning classifier jointly to CO_2, temperature and humidity measurements, and obtained high accuracy in a 70 m^2 space with up to 43 people. However, applying a classifier, where each class corresponds to an occupancy range, rather than regression, poses a huge challenge in practice (training data is needed for many classes). In contrast, Zou et al. [8] exploited disruptions in WiFi-signal spectrum caused by the human body. Using two carefully-placed, custom WiFi routers, they showed good performance with 11 occupants in a 50 m^2 space. However, locations of routers installed commercially are not optimized for people counting, and the approach requires data annotation for new router topologies.

The most common *passive-direct* approach is through the analysis of human-body appearance captured by a camera equipped with rectilinear (conventional) lens, e.g., surveillance camera. While early work focused on *model-based* approaches, performance could not be scaled to higher levels of occupancy. All recent methods have been largely *data-driven*, where large, annotated datasets are used to train machine-learning algorithms. Ryan et al. [9] applied scene-invariant perspective normalization and showed that when trained on images from 6 different scenes it works well on a new scene with up to 32 occupants. Liu et al. [10] applied Support Vector Machine (SVM) classification to image-gradient features, but the method was demonstrated on scenes with 5-8 occupants only. Erickson et al. [11] developed a tripwire-type system that uses above-door cameras and k-Nearest Neighbor (kNN) classification, but the method was demonstrated only on 25 people and required periodic resets due to drift errors. The most recent work leverages deep learning for people detection, such as You-Only-Look-Once (YOLO) [12] and Faster Region-based Convolutional Neural Network (R-CNN) [13]. In each case, validation was performed using a high-definition camera mounted close to the ceiling in scenarios with quite low occupancy of 14 and 8, respectively.

As for high-occupancy scenarios, Conti et al. [14] proposed two methods, one based on head detection using LeNet-5 neural network and another one based on crowd-density estimation using a multi-resolution CNN followed by regression. The methods were tested in two large classrooms (170 m^2 and 250 m^2) with up to 95 people. In another work from the same group, Paci et al. [15] proposed to extract features from temporal differences between spatial image gradients and count people using Support Vector Regression (SVR). Yang et al. [16] applied SVM-based face detection to images from a pan-tilt-zoom (PTZ) camera in patrol mode. The method was tested in a 204 m^2 classroom with up to 150 students, but required precise manual delineation of 6 patrol zones to avoid overlapping detections.

Recently Wang et al. [17] proposed localization of people indoors using up to 8 time-synchronized rectilinear cameras. Their method estimates 2-D skeletons in each camera view and projects them to 3-D space for skeleton clustering. It was demonstrated in a 100 m^2 classroom with 11 people. The use of multiple conventional cameras with overlapping fields of view for occupant localization is analogous to the use of multiple fisheye cameras for occupant detection that we propose.

Beyond conventional cameras, thermal/IR sensors have been used for occupancy sensing, either to detect entry/exit [18,19] or to detect human presence in a small area [20,21]. 3D/depth cameras have been proposed as well [22,23], but were tested in small areas only.

To summarize, while the environmental sensors are inexpensive and often part of a BMS, the people-counting accuracy is sensitive to sensor placement and model parameterization. The use of WiFi signals is not generalizable as it requires custom routers and specific topology. Today's thermal/IR/depth sensors do not have sufficient resolution for accurate people counting in large spaces, while their use as tripwires suffers from drift errors. In contrast, occupancy sensing using conventional cameras and modern machine learning is a promising approach for fine-grained occupancy sensing. However, in large spaces many rectilinear cameras are needed since their field of view (FOV) is typically limited to about 90°. We propose to use fisheye cameras instead.

3. Fisheye cameras: benefits and challenges

To maximize the monitored-area coverage, COSSY uses ceiling-mounted fisheye cameras. The overhead mounting significantly reduces the severity of occlusions (by other people or furniture) and fisheye lens delivers a very wide FOV (360° parallel and 180° orthogonal to the floor). This reduces the number of cameras needed to cover a large space compared to rectilinear cameras. However, the camera installation height must be carefully selected to avoid FOV obstruction by light fixtures, sprinklers, etc. Fig. 2 shows two spaces where COSSY was deployed and validated: 500 ft^2 standard classroom (Fig. 2(a)) and 2,000 ft^2 studio-style classroom (Fig. 2(b)). Fig. 3 shows overhead fisheye views of these spaces in a typical test scenario. It is clear that people detection in such images faces several challenges.

Circular field of view: In rectilinear cameras, a sensor captures only the central rectangular portion of the FOV and the lens is carefully designed to project straight lines in the physical world onto straight lines in the rectilinear image. However, in fisheye cameras a straight line in the physical world, that does not belong to a plane orthogonal to the sensor and passing through its center, becomes curved in the fisheye image (e.g., horizontal edges of whiteboards in Fig. 3(a)). This geometric distortion introduced by the fisheye lens also affects human-body shape, especially if not in an upright position.

Non-linear foreshortening: With its hemispherical design to capture wide FOV, a fisheye lens introduces radial distortion (non-linearity) in the captured images. In particular, the projection of a person standing under the camera (e.g., certain shoulder width), becomes smaller at 8 ft away from the camera and even smaller at 16 ft away. While this size reduction is *linear* in standard cameras on account of rectilinear lens, in fisheye cameras the doubling of the distance from the camera results in size compression by more than 2, especially pronounced close to FOV periphery. This is particularly visible in Fig. 4 where the concentric green circles are equispaced in room coordinates, but in the fisheye image appear closer to each other with increasing radius. This non-linear mapping of physical distances (and of body sizes) poses challenges for people detection in fisheye images.

Human-body orientation: The overhead viewpoint is unusual with people in the image center seen directly from above (Fig. 3(b)) and those farther away seen from a side-view perspective. This dramatic person-viewpoint variability is not encountered in images captured by standard side-mounted cameras. Related, and critically important, is the fact that *standing* people appear in radial directions in a fisheye image, including horizontal and "upside-down" orientations. In general, in overhead fisheye images people can appear at *any* orientation. This is unlike in images captured by side-mounted rectilinear cameras (standing people appear upright) for which the vast majority of people-detection algorithms have been developed.

4. Performance metrics

Several performance metrics have been used in the occupancy-sensing literature, each with its own deficiency. No single metric has been universally adopted, which makes it difficult to compare performance of different occupancy-sensing systems. Below, we review key metrics used to date, discuss their deficiencies and propose two new ones.

(a) (b)

Fig. 2. (a) 500 ft^2 classroom monitored by a single camera mounted 8-1/2 ft above the floor; (b) 2,000 ft^2 studio-style classroom monitored by 3 cameras mounted 10 ft above the floor. The distance between cameras A and B is 121 in, and between cameras B and C it is 248 in. All cameras shown are Axis M3057-PLVE (3,072×2,048 pixels, 185° lens).

(a) Standard classroom (b) Studio-style classroom

Fig. 3. Overhead fisheye views of two rooms from Fig. 2 in a typical testing scenario.

Most papers treat people counting as a *regression* problem. Let η_i be the true people count in frame number i, and let $\hat{\eta}_i$ be the corresponding people-count estimate. Also, let N be the total number of occupancy estimates. The two most often used performance metrics are the Mean-Absolute Error (MAE) and the Root Mean-Squared Error (RMSE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{\eta}_i - \eta_i|, \quad RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\eta}_i - \eta_i)^2}. \tag{1}$$

Both are commonly used in occupancy-sensing literature, but MAE is more robust to outliers (large errors) than RMSE. To allow comparison with other methods, we provide MAE values in this paper.

However, MAE can be misleading when comparing results across different ground-truth occupancy levels (e.g., 10 versus 100 people). To account for this, Mean-Absolute Relative Error (MARE) has been used in some papers (e.g., in [24]), defined as follows:

$$MARE = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{\eta}_i - \eta_i|}{\eta_i},$$

but it is undefined for frames with $\eta_i = 0$ (empty room) which is a major deficiency. Therefore, error normalization by the dynamic range of occupancy has been also proposed [12], which for MAE is defined as follows:

$$NMAE = \frac{\frac{1}{N}\sum_{i=1}^{N}|\hat{\eta}_i - \eta_i|}{\eta_{max} - \eta_{min}},$$

where η_{min} and η_{max} are, respectively, the minimum and maximum of the true people count across all instances considered. However, $NMAE$ is undefined when the occupancy is constant for $i = 1, ..., N$, not an unlikely scenario.

To address these issues, we propose a new metric. Rather than scaling MAE (or $RMSE$) by the dynamic range of true occupancy, we propose to scale it by the *average* of true occupancy as follows:

$$MAE_{pp} = \frac{\frac{1}{N}\sum_{i=1}^{N}|\hat{\eta}_i - \eta_i|}{\frac{1}{N}\sum_{i=1}^{N}\eta_i}. \tag{2}$$

MAE_{pp} expresses MAE value *per person*, so different occupancy scenarios (e.g., 10 versus 100 people) can be fairly compared. The only scenario when it is undefined occurs when a space remains completely empty throughout the whole period of interest, a very special case that can be handled separately.

Works that consider people counting as a *classification* problem (e.g., [7]) report accuracy defined as the fraction of total number of instants (expressed in percent) for which $\hat{\eta}_i = \eta_i$. Since this requires an exact match between the true and estimated accuracy, this fraction is usually low in high-occupancy scenarios, and not very useful. Therefore, we extend the definition of "accuracy with slack of 1" [25] and propose X-Accuracy, defined as follows:

$$Acc_X = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(|\hat{\eta}_i - \eta_i| \le X), \qquad \mathbb{1}(\mathcal{E}) := \begin{cases} 1 & \text{if } \mathcal{E} \text{ is true} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

For $X = 0$ this definition reverts to the original definition of accuracy, but for larger values of X it tolerates the departure of $\hat{\eta}_i$ from η_i by up to X. In the context of people counting, for example Acc_5 gives the percentage of frames in which the estimated count is within 5 off the true count.

5. Single-camera COSSY

The unusual FOV and geometric distortions in fisheye images are key obstacles to reliable people detection and, therefore, people counting. To date, many model-based and deep-learning algorithms have been proposed for the detection of objects, including people. However, they have been designed under the assumption of a side view from standard rectilinear-lens camera, e.g., Fast Feature Pyramids [26], YOLO [27], Single-Shot Detector (SSD) [28], R-CNN [29]. By design, these algorithms can only find bounding boxes (around objects/people) that are aligned with the image axes, and this works well on standard images where people typically appear upright. However, they perform poorly on overhead, fisheye images by mis-detecting non-upright bodies (e.g., upside-down) and uncommon viewpoints (e.g., directly from above).

To accommodate the unusual body orientation, several SVM- or YOLO-based people-detection algorithms have been recently proposed [30–33], each dealing differently with the radial geometry. However, these algorithms are either very complex computationally [30,32] or do not accurately account for arbitrary body orientation [33]. To address these issues, we proposed Rotation-Aware People Detection (RAPiD) [34], a novel end-to-end people-detection algorithm for overhead, fisheye images. RAPiD is a convolutional neural network that predicts *arbitrarily-rotated* bounding boxes of people in a fisheye image. Its source code is publicly available [35].

RAPiD handles unusual body viewpoints by training on a variety of fisheye images. For this purpose, we have collected and curated several fisheye-image datasets, namely HABBOF, MW-R, CEPDOF, WEPDTOF and FRIDA. They have been recorded in a variety of indoor spaces and occupancy scenarios, and are publicly available as well [36].

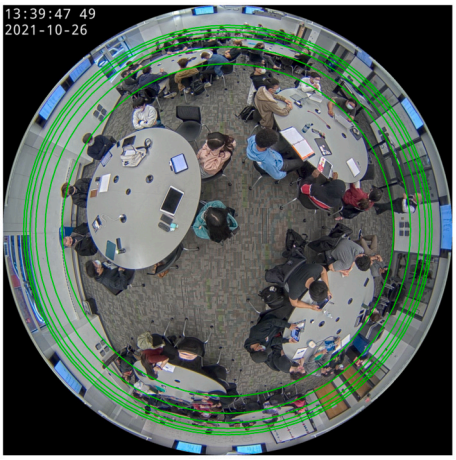


Fig. 4. Overhead fisheye view of the studio-style classroom (Fig. 2(b)) with superimposed concentric circles corresponding to true physical distances from the camera ranging from 10 ft to 35 ft in 5 ft increments. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

In all experiments below, we used RAPiD [34] with network weights obtained by initial training on the COCO 2017 dataset's training images [37,38] for 100,000 iterations and fine-tuning on MW-R and HABBOF datasets for 4,000 iterations. When training on COCO images, we updated the network weights by Stochastic Gradient Descent (SGD) with the following parameters: step size 0.001, momentum 0.9, weight decay 0.0005. When training on MW-R and HABBOF images, we used SGD with default parameters except for step size of 0.0001. We applied random rotation and color augmentation in both training stages. More details can be found in [34].

RAPiD produces bounding boxes and confidence values that tell how likely a bounding box is to contain a person (0 = "impossible", 1 = "certain"). By selecting a confidence threshold γ, the algorithm can be tuned to specific scenarios. Unless otherwise stated, we used $\gamma = 0.05$.

5.1. Performance for different fields of view

In order to understand the limitations of RAPiD, we first quantify its people-counting performance as a function of the monitored area. Fig. 4 shows a video frame from camera B (Fig. 2(b)) with concentric green circles superimposed, each circle corresponding to true physical distance of 10-35 ft from the camera in 5 ft increments. As discussed earlier, the circles are not equally spaced out in the image, confirming radial non-linearity of the lens.

We evaluated RAPiD's performance on a video segment with little movement of occupants (the number of people inside of each circle is constant) to ease annotation. The segment consists of 50 frames, each with 62 occupants. Table 1 lists the area of each circle (FOV) and of a square-shaped room that can be covered by this FOV. As the FOV radius increases from 10 ft (with 23 people visible) to 35 ft (62 people visible), MAE increases from 0.64 to 16.66 while MAE_{pp} increases from 0.028 to 0.269 (or from 2.8% to 26.9%).

This increase in the error was to be expected and is due to the reduction of projected body size and increasing likelihood of occlusions with distance. Up to about 20 ft, the algorithm can count people with less than 0.07 of MAE_{pp} (7% error); it accurately counts the vast majority of the 44 people located inside this circle, despite severe occlusions by other people, tables, chairs, etc. The 20 ft radius corresponds to effective coverage of a square room with 800 ft^2 area. For larger FOVs, the error rapidly increases, so to achieve more accurate counting additional fisheye cameras are needed.

Table 1
People-counting performance of RAPiD for increasing camera field of view. "FOV area" is the area inside the corresponding green circle in Fig. 4, while "Square-room coverage" is the area of a square inscribed in this circle.

FOV radius	FOV area	Square-room coverage	Number of people	MAE	MAE_{pp}
10 ft	314 ft^2	200 ft^2	23	0.64	0.028
15 ft	707 ft^2	450 ft^2	34	2.36	0.069
20 ft	1,257 ft^2	800 ft^2	44	2.88	0.065
25 ft	1,963 ft^2	1,250 ft^2	52	6.92	0.133
30 ft	2,827 ft^2	1,800 ft^2	60	14.78	0.246
35 ft	3,848 ft^2	2,450 ft^2	62	16.66	0.269

Table 2
People-counting performance (MAE, MAE_{pp} and X-Accuracy) of RAPiD in 3 selected scenarios and cumulatively over all 8 scenarios in the CEPDOF dataset [36].

	Lunch meeting 2	Edge cases	IRfilter	Cumulative
No. of frames	3,000	4,201	3,000	25,504
MAE	0.436	0.420	1.582	0.827
MAE_{pp}	0.040	0.076	0.236	0.122
Acc_X [%] $X=0/1/2$	59/98/100	64/95/99	19/53/79	44/81/94

5.2. Performance in different occupancy scenarios

To assess performance in more challenging conditions, we used CEPDOF [36], an annotated dataset with 8 scenarios differing in human poses, movement, occlusions, illumination, etc. Sample images with RAPiD's detections in 3 scenarios are shown in Fig. 5. We selected "Lunch meeting 2" since it has the highest occupancy (up to 13). "Edge cases" includes significant movement and unusual poses (e.g., crouching, stretching) to challenge the detector. In "IRfilter", lights are turned off further challenging the detector by low image contrast and detail. While in the well-lit images all people are correctly detected, in "IRfilter" (Fig. 5(c)) there are 2 misses and 1 false positive, not surprising as RAPiD was not trained on low-light images.

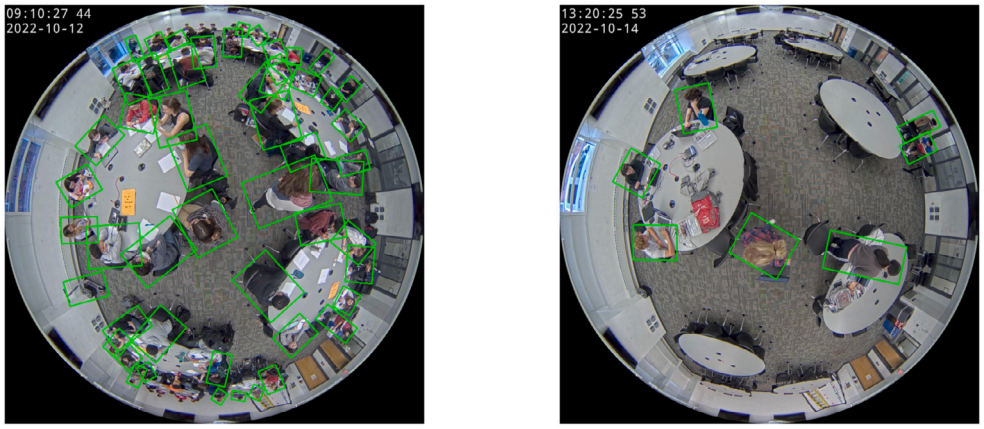
Table 2 summarizes RAPiD's performance on CEPDOF. While MAE is relatively low for well-lit scenarios, unsurprisingly it is quadrupled for the low-light scene. The cumulative MAE over all 8 scenarios is reasonable at 0.827. MAE per person (MAE_{pp}) is small for "Lunch meeting 2" (0.04 or 4%) and for "Edge cases" (0.076 or 7.6%) indicating that RAPiD can handle unusual poses, movement, and occlusions well. However, for "IRfilter" it is much higher at 0.236 (or 23.6%) suggesting that improvements are needed in low light. Cumulatively, MAE_{pp} of 0.122 (or 12.2%) is quite high since 35% of CEPDOF images have been captured in low light. We show X-Accuracy only for $X = 0, 1, 2$ since the occupancy is quite low (up to 13). While perfect counting ($X = 0$) is accomplished in about 60% of well-lit frames, in low light it happens in only 19% of frames. A slack of 1 or 2 increases X-Accuracy to 95-100% in well-lit scenes suggesting that RAPiD is accurate, but only to 53-79% in low light. Cumulatively, X-Accuracy is high for $X = 1, 2$ but only 44% for $X = 0$ due to a large fraction of low-light images in CEPDOF.

5.3. Performance in large space with high occupancy

Thus far, we showed that RAPiD performs well in small-to-mid-sized spaces in various scenarios (except for low light) in short-term tests (75 min in total). To assess RAPiD's performance over longer time in a highly-dynamic occupancy scenario, we recorded a 3-day video in the studio-style classroom. On the first day, there were 11 high-occupancy periods (lectures) with up to 87 students (Fig. 6). On the second day there were 4 high-occupancy periods with up to 65 students, while on the third day the classroom was mostly empty with only one period when up to 9 students were present.

(a) Lunch meeting 2 (b) Edge cases (c) IRfilter

Fig. 5. Examples of detections by RAPiD in the standard classroom (Fig. 2(a)) in 3 scenarios selected from the CEPDOF dataset [36].



(a) Day 1 – crowded lecture (b) Day 3 – low occupancy

Fig. 6. Sample frames from camera B with RAPID's people detections in the 3-day test.

Notably, the classroom was empty during more than half of the test time (Fig. 7) with lights turned off at night, the worst case scenario for RAPiD. To adapt to these adverse conditions, we changed RAPiD's threshold γ to 0.6 in low light to reduce false positives, while keeping the original value of 0.05 in normal light. An alternative would be to train two versions of RAPiD for the two cases, a potential project in the future.

Fig. 6 shows two sample frames: from day 1 with 87 people and from day 3 with 8 people. Clearly, not all of the 87 people were detected by RAPiD, especially at the top and bottom of the frame, since they are too small and/or occluded. Fig. 7 shows the people-count estimates produced by RAPiD for camera B against the ground-truth count (obtained by carefully counting people at each time instant). Plots for cameras A and C look very similar and are omitted. While the occupancy estimate is highly correlated with the ground truth, it is not accurate. The true count is underestimated on days 1 and 2 by as much as 20-30. This is due to the size of the room and people present at the FOV periphery where RAPiD is unreliable. On day 3, the estimate accurately tracks the ground truth since the room is largely empty or the few occupants sit in the room's center under camera B (Fig. 6(b)).

Table 3 shows performance metrics for RAPiD in the 3-day test. For all 3 cameras, MAE is largest (at about 12) on day 1 when the occupancy is very high during daytime. It drops to about 7-8 on day 2 when the classroom is occupied in daytime, but with fewer people. It is the smallest (at about 0.4-0.6) on day 3 when the room is mostly empty. Clearly, MAE values approximately scale with average occupancy on each day. This is confirmed by MAE_{pp} values that are relatively constant around 0.4 except for camera C on day 3 when the occupants sit under camera B (Fig. 6(b)) and camera C is farther from room's center than camera A (Fig. 2(b)). The X-Accuracy for $X = 0$ is low on days 1 and 2 (29-47%) but quite a bit higher on day 3 (55-67%). This is not surprising as it is easier to declare zero occupancy than to get high occupancy exactly. At $X = 5$ (the count estimate may be up to 5 off), the X-Accuracy increases to 53-64% on days 1-2, and to 100% on day 3. This is consistent with day 3 having mostly zero occupancy. At $X = 10$, the X-Accuracy is in 59-70% range on days 1-2, and 100% on day 3.

Cumulatively across the 3 days, MAE is around 6 and MAE_{pp} is around 0.4 for all three cameras. Also, the X-Accuracy is fairly consistent between the cameras. Note, that MAE_{pp} of 0.4 is higher than 0.269 reported in Table 1 for 35 ft FOV. This is due to higher occupancy (87 instead of 62) and significant movement of people, which causes occlusions, especially at the start and end of each class. The cumulative MAE_{pp} of 0.373 for camera B is slightly lower than that for the other cameras, which is not surprising since it is mounted close to classroom's center where students tend to congregate.

These results indicate that while single-camera COSSY is suitable for small-to-medium size spaces (up to about 800 ft^2), where MAE_{pp} does not exceed 0.07 (or 7%), as reported in Table 1, the system underper-

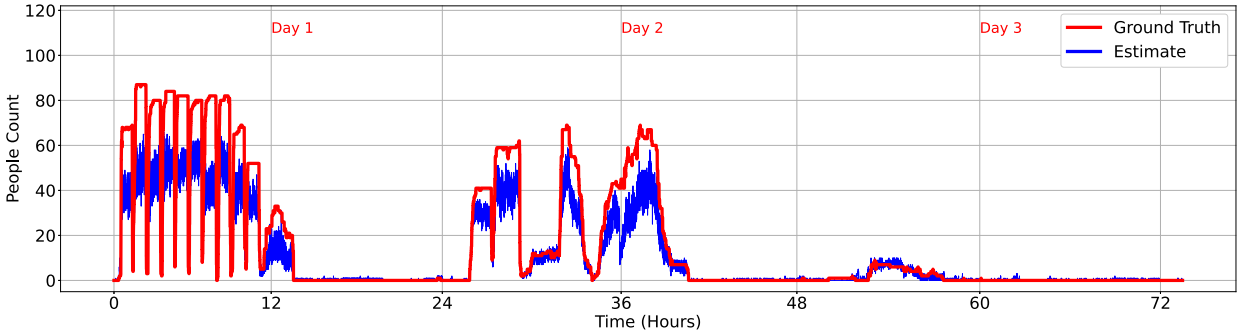


Fig. 7. True number of people (red) and RAPiD estimate (blue) over 3-day test (starting at 7:30 am) in the studio-style classroom for camera B (see camera layout in Fig. 2(b)).

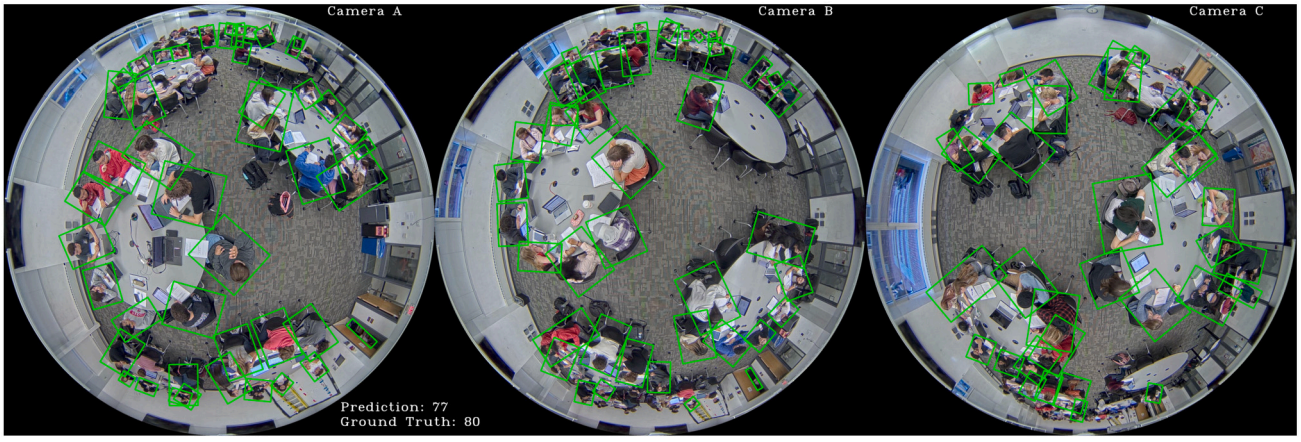


Fig. 8. Overhead fisheye views from cameras A, B and C (Fig. 2(b)) installed in a 2,000 ft^2 studio-style classroom with overlaid people detections produced by RAPiD.

Table 3
People-counting performance of RAPiD in 3-day test in the studio-style classroom for cameras A, B, C. The last column shows cumulative metrics computed across 3 days.

Metric	Camera	Day 1	Day 2	Day 3	Cumulative
MAE	A	11.60	8.55	0.48	6.61
	B	11.82	6.92	0.38	6.11
	C	12.32	7.36	0.62	6.49
MAE_{pp}	A	0.393	0.415	0.480	0.404
	B	0.400	0.336	0.384	0.373
	C	0.417	0.358	0.623	0.397
Acc_X [%] $X=0/5/10$	A	47/54/60	29/60/65	67/100/100	48/72/76
	B	45/53/59	43/64/70	74/100/100	55/73/77
	C	38/59/64	42/64/70	55/100/100	46/75/79

forms in large spaces with MAE_{pp} reaching 0.4 (or 40%) and must be redesigned.

6. Multi-camera COSSY

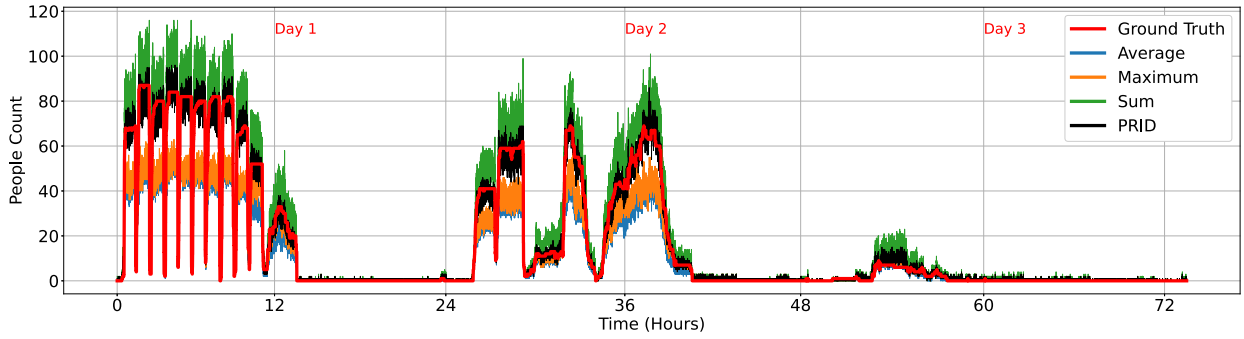
To improve the single-camera COSSY performance, we need to understand its failure modes. Fig. 8 shows views from 3 cameras on day 1 of the 3-day test. With 80 students in the classroom, single camera is insufficient to detect all of them since far-away students appear tiny and are often occluded. While camera B is installed in the room's center (best location for a single camera), cameras A and C are located away from the center, each covering one end of the room better than camera B.

The idea is to use cameras A and C jointly to count all occupants. However, people may appear in the views of *both* cameras, so adding the counts from the two cameras would cause overcounting. One could average these counts or take the maximum, but neither guarantees correct count in general. The correct approach is to identify who is visible in both cameras' FOVs, and count them once – occupants need to be re-identified between different camera views, a problem generally known as *person re-identification* (PRID).

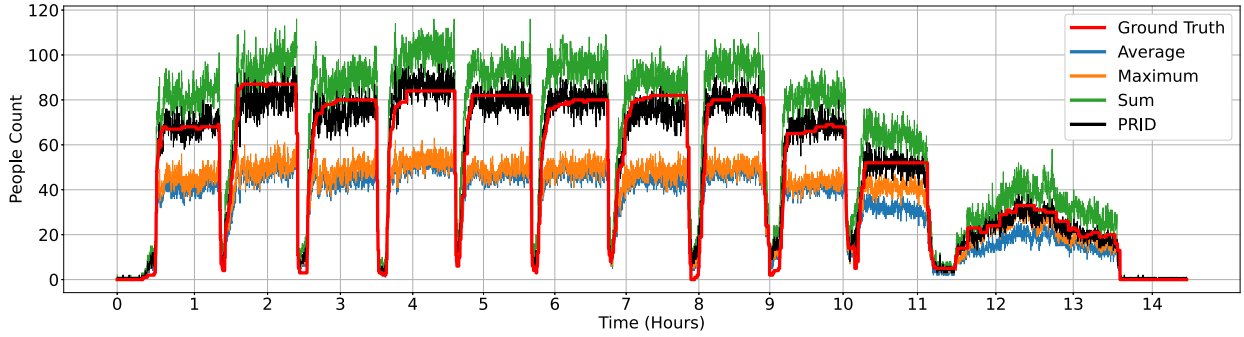
While many PRID algorithms have been developed based on appearance (e.g., color, texture, edges [39,40] or neural-network embedding [41–43]), in our case due to tiny images of far-away people, partial body occlusions and non-linear fisheye-lens distortions, appearance-based methods often fail to accurately match identities [44]. The partial-occlusion problem is particularly severe in high-density scenarios such as the one shown in Fig. 8.

Therefore, we have developed an approach that is based on a person's location rather than appearance [45]. The key insight is that only one person can occupy a given location in a room (3D space) at a given moment. Since all cameras capture images at the same time, this person appears at a *specific* pixel-location in each camera's image. If we know the location of a person in a camera's view (e.g., detected by RAPiD), we can map their location to another camera's view and check if a person is detected nearby.

This mapping of locations between two cameras depends on a number of parameters. Before estimating these parameters, using a precision laser measure we made sure that both cameras' sensor planes are parallel to the floor. Then, using the same device we precisely measured two *extrinsic* parameters, the distance between cameras and camera installation height. Finally, we estimated other parameters: one *extrinsic* parameter, namely rotation angle between the cameras in the plane par-



(a) 3-day test



(b) Day 1

Fig. 9. (a) True number of people (red) and 4 estimates obtained in the 3-day test by aggregating RAPiD's detections in frames from cameras A and C by means of: average (blue), maximum (orange) or sum (green) of people-count estimates, or person re-identification (black) between detections from these cameras. (b) Expanded plot of day 1.

allel to the floor, and 5 *intrinsic* parameters characterizing the specific camera model used. This camera-calibration process was automated: a spherical, distinct-color LED light source was moved around the test classroom at a fixed known height on a dolly and the collected images were used to find optimal camera parameters by minimizing a camera-to-camera mapping error using SGD over 5,000 iterations with a learning rate of 10^{-5}. More details can be found in [45]. Note that the parameter calibration process needs to be performed only once for each new camera model (intrinsic parameters). When deploying an already calibrated camera unit in new space, the only measurements needed are its installation height, distance to a reference point and rotation angle (after ensuring that all cameras are level).

For a calibrated camera pair, this approach maps the location of a person (bounding-box center) from one camera's frame to 3D space and then from 3D space to the other camera's frame. Since the 2D-to-3D projection has one degree of freedom (scale factor), we constrain it using the average height of a person in the US (66 in). Distances are computed between all pairs of *predicted* locations (of people from the first camera) and *detected* locations (of people in the second camera), and the best matching between them is found using a greedy algorithm. Out of 4 distance measures proposed in [45], we selected Count-Based Distance (CBD) since it outperforms other metrics. A *predicted* location far away from all *detected* locations in a frame likely corresponds to a person visible in one camera only. To separate "matches" from "non-matches", we use a distance threshold $\tau = 0.2$ (see [46] for details).

If $\hat{\eta}_i^A$ and $\hat{\eta}_i^C$ are the estimated people counts by RAPiD in frame number i from cameras A and C, respectively, and $\hat{\eta}_i^{AC}$ is the number of people successfully *re-identified* between these two frames, then the people-count estimate for this pair of frames is computed as follows:

$$\hat{\eta}_i = \hat{\eta}_i^A + \hat{\eta}_i^C - \hat{\eta}_i^{AC}. \tag{4}$$

The subtraction of $\hat{\eta}_i^{AC}$ removes double counts discovered by PRID.

6.1. Performance in large space with high occupancy

Fig. 9 shows the ground truth and 4 estimates obtained in the studio-style classroom in the 3-day test by aggregating RAPiD's detections from cameras A and C in different ways. While the sum of counts produced by RAPiD from the two cameras significantly overestimates the true count, the average and the maximum of counts severely underestimate the true count (note that the blue line is mostly covered by the orange line). The application of PRID via (4) results in quite accurate tracking of the true count.

Table 4 shows quantitative performance of RAPiD for the 3-day test (Fig. 9). Similarly to single-camera COSSY, MAE is largest on day 1 and smallest on day 3 for all algorithms. While the MAE values for average and maximum aggregation roughly scale with occupancy on each day, this scaling is much less pronounced for the summation of counts and even less so for PRID. The MAE_{pp} values stay relatively unchanged between days 1 and 2, with those for PRID being by far the lowest, but significantly increase on day 3 for some algorithms (for the maximum it about doubles to 0.670, for the summation it increases 5-fold to 1.436 and for PRID it increases almost 10-fold to 0.752). This is due to occasional false positives that RAPiD produces (e.g., double detection of a person, spurious detection in low light). While fairly insignificant when occupancy is high, such false positives are very detrimental when the space is empty (most of day 3) or almost empty (e.g., one false positive in addition to one true positive creates a 100% absolute error per person). The summation of counts is very sensitive to false positives as they directly impact the final count. PRID is a little less sensitive than summation since a false positive may be matched with a detection in the other camera's frame and not counted. The averaging of counts is even less sensitive to false positives due to the division by 2 and rounding down, as is the maximum since when a false positive happens in a frame with lower count it has no impact on their maximum. As for X-Accuracy, PRID significantly outperforms other algorithms for $X = 5$

Table 4
People-counting performance (MAE, MAE_{pp} and X-Accuracy) of RAPiD in a 72-hour test in the studio-style classroom using cameras A and C.

Metric	Algorithm	Day 1	Day 2	Day 3	Cumulative
MAE	Average	11.99	7.97	0.36	6.50
	Maximum	10.36	6.35	0.66	5.57
	Sum	6.70	5.31	1.42	4.35
	PRID	2.43	1.71	0.75	1.59
MAE_{pp}	Average	0.406	0.387	0.365	0.397
	Maximum	0.351	0.309	0.670	0.340
	Sum	0.227	0.258	1.436	0.266
	PRID	0.082	0.083	0.752	0.097
Acc_X [%] $X=0/5/10$	Average	47/55/62	37/62/66	72/100/100	53/73/77
	Maximum	38/59/66	32/65/71	52/100/100	41/76/80
	Sum	37/56/70	27/61/80	48/91/98	38/70/83
	PRID	41/84/96	37/93/99	51/99/100	43/92/98

and $X = 10$, but is in the middle for $X = 0$ (see an explanation in the next paragraph).

Cumulatively across 3 days, PRID significantly outperforms other algorithms in MAE, MAE_{pp} and X-Accuracy except for $X = 0$ when the average of counts performs better. Looking at the plot in Fig. 9 this seems surprising for the blue line of the average (largely covered by the orange line) is mostly well below the true count (red line). However, the perfect people counting in 53% of frames by means of averaging compared to 43% by PRID is due to the *unoccupied* periods that constitute more than half of the 72-hour test period. If in an empty scenario RAPiD falsely detects one person in a frame, averaging *always* produces zero-occupancy estimate due to rounding down while PRID, maximum and sum might not (as explained earlier).

Comparing the results in Table 4 to those in Table 3 it is clear that a two-camera COSSY significantly outperforms a single-camera COSSY in this 2,000 ft^2 space. MAE_{pp} is reduced from 0.373 to 0.097 (or from 37.3% to 9.7%) and the X-Accuracy is increased from 75% to 92% for $X = 5$ and from 79% to 98% for $X = 10$, although it drops from 55% (camera B) to 43% for $X = 0$ (due to the unoccupied periods, as explained above).

6.2. Temporal post-processing

The plots in Figs. 7 and 9 show that even if the true occupancy does not change, the raw people-count estimates vary significantly in time (due to RAPiD's detection errors). These variations are detrimental to HVAC control because when the BMS periodically intercepts people counts it expects representative values of recent occupancy rather than random fluctuations. Since, in reality, occupancy does not dramatically change in seconds, temporal smoothing of estimates should help. A simple and outlier-robust smoothing can be accomplished by *causal* median filtering as follows:

$$\hat{\eta}_i^{med} = \text{median}\{\hat{\eta}_{i-W+1}, \hat{\eta}_{i-W+2}, ..., \hat{\eta}_i\}, \quad (5)$$

where W is the median window size. The causality is needed for real-time implementation, but causes a delay by $W - 1$ frames.

Fig. 10 shows MAE_{pp} and X-Accuracy as functions of window size W for the best-performing multi-camera COSSY algorithm that uses PRID-based people-counts aggregation. It achieves the minimum MAE_{pp} of 0.082 for $W = 11$ compared to 0.097 for $W = 1$ (no median filtering). Clearly, there is a modest performance benefit to median smoothing, but as W increases beyond 11, MAE_{pp} increases as well. The X-Accuracy at $X = 0$ achieves the maximum of 47% for $W = 33$ compared to 43% with no median filtering, again a slight improvement. Similarly, the X-Accuracy maxima of 95% at $X = 5$ for $W = 19$ and of 99% at $X = 10$ for $W = 13$ are slight improvements over the results with no smoothing (92% and 98%, respectively).

The curves in Fig. 10 are slowly varying and any W between 10 and 30 guarantees improved performance over no median filtering while assuring reduced estimate variations as shown in Fig. 11. More specifically, for $W = 11$ there are smaller estimate variations (Fig. 11(a)) than in Fig. 9 and for $W = 99$ (Fig. 11(b)) the plots are quite smooth. Note that PRID tracks the ground truth very well – it is often occluded by the red line in Fig. 11(b).

The reduction of people-count variations is important for HVAC control. When the BMS receives counts, they reflect recent occupancy rather than random fluctuations. However, there is a compromise – the larger the W, the larger the delay (estimates lag behind the ground truth, but this is not visible in Fig. 11 due to compressed time scale). Our current implementation of two-camera COSSY requires about 3 sec on average to complete people counting from a pair of fisheye frames. For $W = 11$ (Fig. 11(a)) this implies a delay of about 30 sec and for $W = 99$ (Fig. 11(b)), about 5 min. Depending on the building and HVAC scenario, W needs to be judiciously selected.

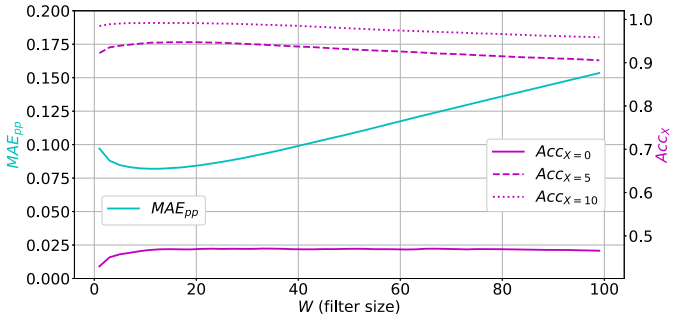
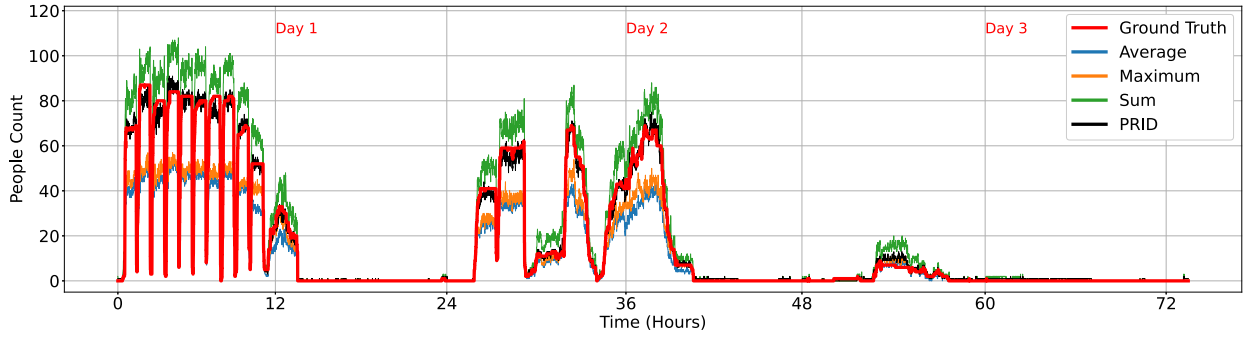
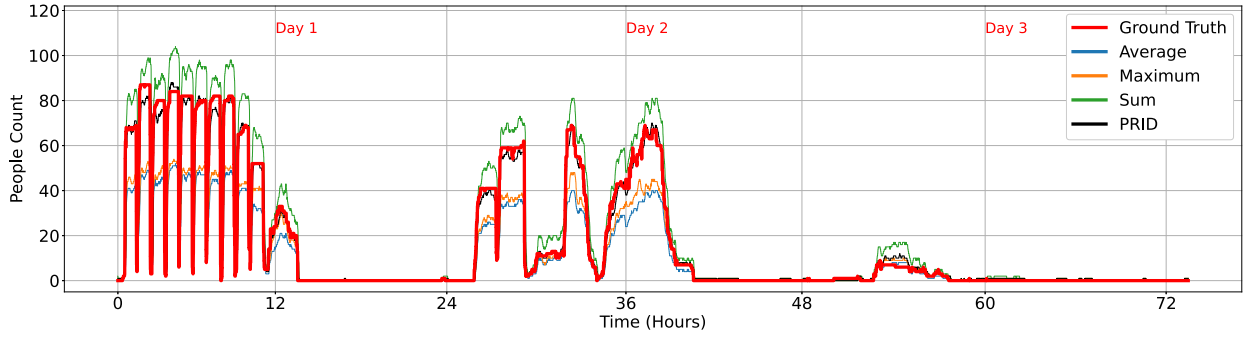


Fig. 10. Dependence of MAE_{pp} and X-Accuracy on the size of median filtering window W for the PRID-based multi-camera COSSY.



(a) $W = 11$



(b) $W = 99$

Fig. 11. True and estimated number of people from Fig. 9 after causal median filtering with (a) $W = 11$, and (b) $W = 99$.

7. Discussion

Few works in the literature report occupancy-sensing results in a large, high-occupancy space. We found 3 studies that fit this scenario and compare them with COSSY in Table 5. Since two methods report only RMSE (1), we also provide COSSY's RMSE values although they were not reported above.

All methods in Table 5 were evaluated in similarly-sized spaces with similar highest occupancy (except for Yang et al. [16] who tested with more occupants). Conti et al. [14] report results for two different classrooms, whereas Paci et al. [15] report a range of errors for different training/testing splits averaged between the two classrooms. While the performance of single-camera COSSY is on-par with the other methods, its two-camera version employing PRID significantly outperforms the competition. Its MAE of 1.59 is between 67% and 83% lower than MAE reported by Paci et al. for different splits. Similarly, RMSE of 2.99 for two-camera COSSY is 54% lower than the lowest RMSE of 6.46 reported by Conti *et al.* and as much as 81% lower than the worst-case result of Paci *et al.* While two-camera COSSY also outperforms Yang et al., the latter was tested in a larger space with more occupants. However, it uses a PTZ camera in patrol mode and requires labor-intensive commissioning to avoid overlapping detections between adjacent patrol zones.

A crucial drawback, however, is that all three systems use images from the test classroom during training of either a CNN or SVM. This is not the case with COSSY for its person detector (RAPiD) was trained on unrelated fisheye data (different rooms and occupancy levels). Therefore, unlike the competition, COSSY is expected to deliver very similar performance regardless of the test space.

8. Conclusions

We have developed COSSY, an occupancy-sensing system that uses overhead fisheye cameras and advanced computational algorithms. Tested in a large space with high occupancy, COSSY delivers accurate people counts in seconds on a modern CPU [47]. It is also straightforward to deploy. The single-camera system is plug-and-play and has been validated to work well with cameras installed at 8-12 ft height. The multi-camera version requires only 3 measurements for each additional camera, assuming the specific camera model is already-calibrated for intrinsic parameters. While the calibration of a new camera model is time-consuming, it needs to be performed only once. The current

Table 5
Performance comparison of COSSY against state-of-the-art methods tested in large spaces. Results in [14] and [15] are reported for two classrooms. Room area reported in corresponding paper is in roman font, while the one in italic font results from conversion.

	1-camera COSSY	2-camera COSSY	Conti et al. [14]	Paci et al. [15]	Yang et al. [16]
Room area	2,000 ft^2 *186 m^2*	2,000 ft^2 *186 m^2*	*1,830/2,690 ft^2* 170/250 m^2	*1,830/2,690 ft^2* 170/250 m^2	*2,196 ft^2* 204 m^2
Max. occup.	87	87	80/95	95/70	150
MAE	6.11	1.59	–	4.86-9.44	–
RMSE	12.34	2.99	6.46/8.55	7.12-15.71	7

version of RAPiD (the people detector) could benefit from fine-tuning in low-light conditions, but to some degree this can be addressed by parameter adjustment. In this regard, COSSY requires selection of 3 parameters only: threshold γ to control false positives/negatives in people detection, threshold τ to control the sensitivity to double-counting of people, and window size W to control variations of people counts in time. We also proposed two new performance metrics that are independent of occupancy level and should be very useful when comparing performance in different occupancy scenarios.

COSSY has far exceeded our initial expectations in terms of performance and flexibility. With additional fine-tuning it can be ready for deployment in large commercial spaces with dynamic occupancy to help realize energy savings and provide actionable information to spatial-analytics platforms.

CRediT authorship contribution statement

Janusz Konrad: Conceptualization, Formal analysis, Investigation, Funding acquisition, Supervision, Writing – original draft, Writing – review & editing. **Mertcan Cokbas:** Conceptualization, Formal analysis, Investigation, Methodology, Software, Writing – review & editing. **Prakash Ishwar:** Conceptualization, Formal analysis, Investigation, Funding acquisition, Supervision, Writing – review & editing. **Thomas D.C. Little:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing. **Michael Gevelber:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] 2018 Commercial Buildings Energy Consumption Survey (final results), https://www.eia.gov/consumption/commercial. (Accessed 9 June 2023).

[2] W. Wang, J. Zhang, M.R. Brambley, B. Futrell, Performance simulation and analysis of occupancy-based control for office buildings with variable-air-volume systems, Energies 13 (15) (2020), https://doi.org/10.3390/en13153756.

[3] Z. O'Neill, Y. Li, H. Cheng, X. Zhou, S. Taylor, Energy savings and ventilation performance from CO2-based demand controlled ventilation: simulation results from ASHRAE RP-1747 (ASHRAE RP-1747), Sci. Technol. Built Environ. 26 (2019) 1–20, https://doi.org/10.1080/23744731.2019.1620575.

[4] S. Gunsteinsson, R. Kahn, M. Gevelber, Airflow based model to estimate commercial building HVAC energy use: analysis to determine principal factors for different climate zones, in: International High Performance Buildings Conference, Purdue, 2016.

[5] M. Jin, N. Bekiaris-Liberis, K. Weekly, C.J. Spanos, A.M. Bayen, Occupancy detection via environmental sensing, IEEE Trans. Autom. Sci. Eng. 15 (2) (2018) 443–455, https://doi.org/10.1109/TASE.2016.2619720.

[6] H. Elkhoukhi, M. Bakhouya, D. El Ouadghiri, M. Hanifi, Using stream data processing for real-time occupancy detection in smart buildings, Sensors 22 (6) (2022), https://doi.org/10.3390/s22062371.

[7] A. Szczurek, M. Maciejewska, T. Pietrucha, Occupancy determination based on time series of CO2 concentration, temperature and relative humidity, Energy Build. 147 (2017) 142–154, https://doi.org/10.1016/j.enbuild.2017.04.080.

[8] H. Zou, Y. Zhou, J. Yang, C.J. Spanos, Device-free occupancy detection and crowd counting in smart buildings with WiFi-enabled IoT, Energy Build. 174 (2018) 309–322, https://doi.org/10.1016/j.enbuild.2018.06.040.

[9] D. Ryan, S. Denman, S. Sridharan, C. Fookes, Scene invariant crowd counting, in: 2011 International Conference on Digital Image Computing: Techniques and Applications, 2011, pp. 237–242.

[10] D. Liu, X. Guan, Y. Du, Q. Zhao, Measuring indoor occupancy in intelligent buildings using the fusion of vision sensors, Meas. Sci. Technol. 24 (7) (2013) 074023, https://doi.org/10.1088/0957-0233/24/7/074023.

[11] V.L. Erickson, S. Achleitner, A.E. Cerpa, Poem: power-efficient occupancy-based energy management system, in: 2013 ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), 2013, pp. 203–216.

[12] H. Choi, C.Y. Um, K. Kang, H. Kim, T. Kim, Application of vision-based occupancy counting method using deep learning and performance analysis, Energy Build. 252 (2021) 111389, https://doi.org/10.1016/j.enbuild.2021.111389.

[13] S. Wei, P.W. Tien, T.W. Chow, Y. Wu, J.K. Calautit, Deep learning and computer vision based occupancy CO2 level prediction for demand-controlled ventilation (DCV), J. Build. Eng. 56 (2022) 104715, https://doi.org/10.1016/j.jobe.2022.104715.

[14] F. Conti, A. Pullini, L. Benini, Brain-inspired classroom occupancy monitoring on a low-power mobile platform, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 624–629.

[15] F. Paci, D. Brunelli, L. Benini, 0, 1, 2, many — A classroom occupancy monitoring system for smart public buildings, in: Proceedings of the 2014 Conference on Design and Architectures for Signal and Image Processing, 2014, pp. 1–6.

[16] J. Yang, A. Pantazaras, K.A. Chaturvedi, A.K. Chandran, M. Santamouris, S.E. Lee, K.W. Tham, Comparison of different occupancy counting methods for single system-single zone applications, Energy Build. 172 (2018) 221–234, https://doi.org/10.1016/j.enbuild.2018.04.051.

[17] H. Wang, G. Wang, X. Li, Image-based occupancy positioning system using pose-estimation model for demand-oriented ventilation, J. Build. Eng. 39 (2021) 102220, https://doi.org/10.1016/j.jobe.2021.102220.

[18] A.K. Mikkilineni, J. Dong, T. Kuruganti, D. Fugate, A novel occupancy detection solution using low-power IR-FPA based wireless occupancy sensor, Energy Build. 192 (2019) 63–74, https://doi.org/10.1016/j.enbuild.2019.03.022.

[19] M. Cokbas, P. Ishwar, J. Konrad, Low-resolution overhead thermal tripwire for occupancy estimation, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 398–406.

[20] A. Tyndall, R. Cardell-Oliver, A. Keating, Occupancy estimation using a low-pixel count thermal imager, IEEE Sens. J. 16 (10) (2016) 3784–3791, https://doi.org/10.1109/JSEN.2016.2530824.

[21] M. Piechocki, M. Kraft, T. Pajchrowski, P. Aszkowski, D. Pieczynski, Efficient people counting in thermal images: the benchmark of resource-constrained hardware, IEEE Access 10 (2022) 124835–124847, https://doi.org/10.1109/ACCESS.2022.3225233.

[22] G. Diraco, A. Leone, P. Siciliano, People occupancy detection and profiling with 3D depth sensors for building energy management, Energy Build. 92 (2015) 246–266, https://doi.org/10.1016/j.enbuild.2015.01.043.

[23] H. Lu, A. Tuzikas, R.J. Radke, A zone-level occupancy counting system for commercial office spaces using low-resolution time-of-flight sensors, Energy Build. 252 (2021) 111390, https://doi.org/10.1016/j.enbuild.2021.111390.

[24] S. Kim, S. Kang, K.R. Ryu, G. Song, Real-time occupancy prediction in a large exhibition hall using deep learning approach, Energy Build. 199 (2019) 216–222, https://doi.org/10.1016/j.enbuild.2019.06.043.

[25] M.O. Tezcan, J. Konrad, J. Muroff, Automatic assessment of hoarding clutter from images using convolutional neural networks, in: 2018 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), 2018, pp. 1–4.

[26] P. Dollár, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (8) (2014) 1532–1545.

[27] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proc. IEEE Conf. Computer Vision Pattern Recognition, 2016.

[28] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: single shot multibox detector, arXiv:1512.02325, 2015.

[29] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (Eds.), Proc. Conf. Neural Inf. Proc. Systems, Curran Associates, Inc., 2015, pp. 91–99.

[30] A.-T. Chiang, Y. Wang, Human detection in fish-eye images using hog-based detectors over rotated windows, in: Proc. IEEE Intern. Conf. on Multimedia and Expo Workshops, 2014, pp. 1–6.

[31] T. Wang, C. Chang, Y. Wu, Template-based people detection using a single downward-viewing fisheye camera, in: Intern. Symp. on Intell. Signal Process. and Comm. Systems, 2017, pp. 719–723.

[32] S. Li, M.O. Tezcan, P. Ishwar, J. Konrad, Supervised people counting using an overhead fisheye camera, in: Proc. IEEE Int. Conf. Advanced Video and Signal-Based Surveillance, 2019, pp. 1–8.

[33] M. Tamura, S. Horiguchi, T. Murakami, Omnidirectional pedestrian detection by rotation invariant training, in: Proc. IEEE Winter Conf. on Appl. of Computer Vision, IEEE, 2019, pp. 1989–1998.

[34] Z. Duan, M.O. Tezcan, H. Nakamura, P. Ishwar, J. Konrad, RAPiD: Rotation-aware people detection in overhead fisheye images, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020.

[35] https://vip.bu.edu/rapid.

[36] https://vip.bu.edu/projects/vsns/cossy/datasets.

[37] https://cocodataset.org.

[38] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, P. Dollár, Microsoft COCO: Common objects in context, arXiv:1405.0312, 2014.

[39] T. Matsukawa, T. Okabe, E. Suzuki, Y. Sato, Hierarchical gaussian descriptor for person re-identification, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[40] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, S.Z. Li, Salient color names for person re-identification, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer Int. Publishing, Cham, 2014, pp. 536–551.

[41] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, Z. Wang, ABD-net: attentive but diverse person re-identification, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8350–8360.

[42] M. Wieczorek, B. Rychalska, J. Dabrowski, On the unreasonable effectiveness of centroids in image retrieval, preprint, arXiv:2104.13643, 2021, https://doi.org/10.48550/ARXIV.2104.13643.

[43] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, R. Ji, Pyramidal person re-identification via multi-loss dynamic training, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8506–8514.

[44] M. Cokbas, J. Bolognino, J. Konrad, P. Ishwar, FRIDA: fisheye re-identification dataset with annotations, in: 18th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS), 2022.

[45] J. Bone, M. Cokbas, O. Tezcan, J. Konrad, P. Ishwar, Geometry-based person reidentification in fisheye stereo, in: 17th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS), 2021.

[46] M. Cokbas, P. Ishwar, J. Konrad, Spatio-visual fusion-based person re-identification for overhead fisheye images, IEEE Access 11 (2023) 46095–46106, https://doi.org/10.1109/ACCESS.2023.3274600.

[47] D. Konrad, Z. Duan, M. Cokbas, P. Ishwar, Complexity evaluation of parallel execution of the RAPiD deep-learning algorithm on Intel CPU, arXiv:2312.06544, 2023.