Video Content Analysis-Based Detection of Occupant Presence for Building Energy Modelling

Ipek Gursel Dino[1 2, *], Esat Kalfaoglu[1], Alp Eren Sarı[1], Sahin Akin[1], Orcun Koral Iseri[1], A. Aydın Alatan[1], Sinan Kalkan[1], Bilge Erdogan[2]

[1] Center for Image Analysis (OGAM), Middle East Technical University, Turkey
[2] Department of Architecture, Middle East Technical University, Turkey
[2] Heriot-Watt University, United Kingdom
*email: ipekg@metu.edu.tr

Abstract

The information on occupant presence plays a critical role in building energy modeling for spaces with a high number of occupants. A thorough understanding of occupant behavior is key to precise Building Energy Modeling (BEM) and to increase the precision of the simulation results. Capturing occupant-related information is difficult due to its stochastic and temporally uncertain nature. In this paper, we propose a robust video content analytical approach for the fast and accurate analysis of temporal and spatial video content. This approach counts the number of occupants in a classroom in an existing building by processing the recordings of video cameras. Two novel counting methods were implemented. The first, namely the Average Counting Method, uses cameras installed in the room directed in different angles, this method relies on detecting and counting occupant heads using a deep convolutional network, namely YOLOv2, that we trained on an existing head dataset. The second method, namely the Entrance Counting Method, uses cameras directed towards the room entrance and increments or decrements a counter based on the occupants entering and exiting the classroom. In addition to YOLOv2, the Discriminative Correlation Filter with Channel and Spatial Reliability (CSR-DCF) was used to create temporal relationships. At the same time, the ground truth was established by manual head counting. The analysis of the results of the one-week recordings initially indicate occlusion problems for the videos of door cameras in case of crowded groups. The videos of room cameras also experienced similar difficulties due to occlusions and the detection of occupants located further from the cameras. Based on these observations, an approach to combine the calculations of both methods is developed, wherein the room cameras are considered as the reference in case of local minima, while the rest is calculated using door cameras with respect to these references. Finally, we validate our approach through two experiments. The first experiment concerns the quantitative comparison between the proposed approach and the ground truth acquired through manual counting methods. The second experiment evaluates the results of the proposed approach in an energy model by quantifying the degree of change in terms of different metrics concerning building energy performance. The results are indicative of the critical role of occupancy in energy modeling.

Keywords: computer vision, deep learning, video content analysis, building occupancy, building energy modeling and simulation

1. Introduction

Building retrofit necessitates energy simulation tools to quantify energy performance and occupant comfort measures. Building energy demand is largely determined by two main data categories. The physical data includes aspects related to the climate, building envelope, building services and energy systems, while the occupant-related data relates to the number of occupants, their activities and behavior in building spaces (O'Brien et al., 2017). The second occupant-related category is significant for energy modeling, as occupants contribute to internal heat gains and emit pollutants such as carbon dioxide,

thereby changing the indoor environment (Labeodan et al., 2015). Occupants also adapt the physical building conditions to improve comfort, such as adjusting lighting, heating/cooling set points, curtains, therefore observation becomes complex (Day et al., 2012; Schweiker et al., 2017). Occupant presence is considered as critical, as it has a substantial influence on building resource use and indoor environmental quality, i.e. thermal comfort, ventilation, lighting (Lam, 2015; Toftum, 2010). However, occupant-related information is difficult to capture due to its stochastic and temporal nature (Reinhart C. F. & K., 2003; Yoshino et al., 2017). The most common approach to represent occupancy presence is "diversity profiles". Standard templates using long-term observational data from different buildings and space types are common; however, they run the risk of neglecting the temporal variations, such as seasonal habits, differences in behavior between weekdays and atypical and unpredicted occupant behavior, especially in crowded places (Hong et al., 2017; Kelly Seryak & Kissock, 2003).

Various occupancy sensing technologies have been developed to accurately quantify occupant presence, including in-situ measurements, laboratory experiments (simulating the process) and surveys (Agency, 2018). Between these options, real-time time data collection is effective in order to obtain both regular patterns and extraordinary situations (Gilani et al., 2017). However, choosing the right tool for in-situ measurement is crucial, for instance, Passive Infrared Motion detectors (PIR) are dependent on occupant motion and they only provide presence or absence information rather than the number of people (Amin et al., 2008). On the other hand, the use of proxy measurements through environmental sensor networks is a rather recent research area, but they also require a comprehensive sensor infrastructure (i.e. CO2, CO, TVOC, small particulates, motion, temperature, humidity) and sensor calibrations (Akkaya et al., 2015; Hoes et al., 2009). A robust alternative can be implemented using computer vision methods, such as real-time camera monitoring (Duarte et al., 2013). There is much potential in the use of video cameras for data collection and advanced video analysis for the fast and accurate analysis of temporal and spatial video content (Dziedzic et al., 2017; Sarkar et al., 2008). Related, data can be easily connected with camera registration during the specified period than occupancy information derived by using human detection algorithms (Benezeth et al., 2011; Han & Bhanu, 2007). The collected data of cameras represents a realistic pattern that is also suitable for building energy simulations after converting to a fraction based schedule framework. As in this research aims to obtain, camera monitoring can control occupant presence and occupant activity simultaneously. However, detecting and tracking people remain as challenging problems in complex scenes with multiple people, occlusions and clutter (Andriluka, 2008). This paper presents a method that counts the number of people in a room from video recordings using automated content analysis.

2. Methodology

The proposed method aims to estimate the number of people in an indoor environment from video recordings. The results are registered in 10-minute intervals for a duration of one week, which is to be provided as an input to the energy simulation model. The method primarily uses a head detection algorithm. Supportive tools are background modeling and tracking algorithms. There are two different approaches to estimate the number of people considered in this work. The first approach, the average counting method, averages the number of detected heads per frame for the representation of a determined time interval. The second approach, the entrance counting method, counts the heads entering and exiting the room through the door. It is observed that one can outperform another depending on the specific physical conditions in the room. Therefore, a novel hybrid approach is developed, to acquire a more reliable prediction considering these situations.

2.1 Vision Techniques Used in the Algorithm

In this section, three machine vision tools that are used in the proposed methods are explained. The first tool is the head detection algorithm that is used to estimate the count of people at an instant. The second tool is background modeling, which aims to reduce the number of false candidates given by the head detection algorithm. The third tool is tracking, which creates the temporal relationship between the frames and aims to find the missing detections of the head detection algorithm.

Head Detection

Head detection is implemented via the well-known object detection algorithm "You Look Only Once v2" (YOLOv2 or YOLO9000) in the literature (Redmon & Farhadi, 2017). Among the existing methods in the literature, such as Faster R-CNN and Single Shot Detector (SSD), YOLOv2 has proven to outperform others regarding speed and accuracy (Liu et al., 2016; Ren et al., 2017). YOLOv2 algorithm predicts object categories and locations simultaneously. In this algorithm, the image is divided into a grid. Every grid cell is checked for whether an object exists in it or not. For each grid cell, there are five *anchor boxes*, which are specialized for objects of various sizes. Each anchor box is used to estimate the location and the size of the objects. The size of the grid is an important parameter because only one object can be detected per grid. To capture the very distant people and to increase the grid size, we trained the network with 1280x1280 pixel images for the average counting method and 736x736 for the entrance counting method (the original implementation of YOLOv2 was trained with 416x416 pixel images). Moreover, to overcome the problems of overfitting during training, which results in the inability of the model to generalize, a uniform scale between 1.0 and 1.1 is selected and the images are randomly cropped.

Background Modeling (Background Subtraction)

While counting people, head detection is preferable over the face or the whole body. This is because people may not be directly facing the camera, which is problematic for face detection. Another problem is that the whole body is typically susceptible to occlusion in crowded rooms. On the other hand, head detection can be equally challenging since heads are not as distinguishable as faces or whole bodies. Owing to this, objects such as curved chairs, backpacks or even the backs of people can be misclassified as heads.

Background modeling offers a solution to this problem by distinguishing moving objects (humans) against a static background for a given frame. In this study, the Mixture of Gaussian 2 (MOG2) approach is implemented, where every pixel is represented with a Gaussian Mixture Model (GMM) (Zivkovic & Van Der Heijden, 2006). There is a maximum number of N mixtures of Gaussian color clusters and the instantaneous number of clusters are determined adaptively, such that the clusters that are rarely seen are ignored. Since GMM models the background, the components with a higher probability in the data tend to be part of the background. An update coefficient, or the learning rate of the algorithm, is selected as a small value, such that a foreground object is still identified as part of the foreground for approximately one minute even when it is motionless. This is to eliminate the possibility that motionless people are identified as background by the algorithm. The flowchart showing the combined implementation of head detection with background modeling can be found in Figure 1. The head candidates are eliminated if the area of boxes of head candidates is less than one percent of the total image area. To clarify the head candidate elimination, every head candidate has a bounding box to represent the head which is shown in Figure 1. Background modeling block produces a mask in which white pixels represent the foreground regions and black pixels represent the background regions which is also shown in Figure 1. For the corresponding pixels of the bounding box in the mask, if the ratio of a number of foreground pixels to the number of total pixels is less than one percent, the head candidate which has that bounding box is eliminated. These filtered head candidates are called as foreground head candidates.

Tracking

Tracking is used to increase the probability of a head being detected in consecutive frames. In the proposed algorithm, it is only used in the entrance counting method. The reason for not using tracking in average counting method is that there are too many heads to track in this method and the complexity of good trackers is linearly increasing with the number of objects tracked.

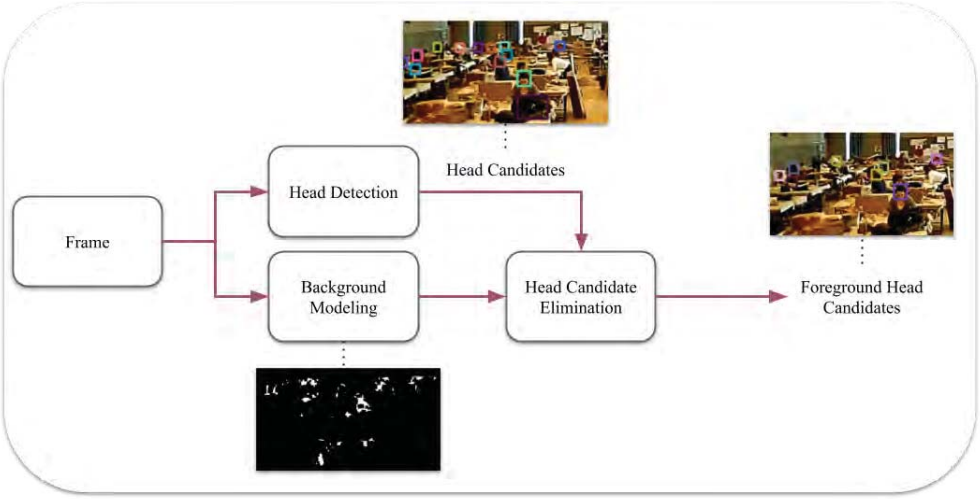


Figure 1: Head Detection with Background Modeling

Recently, the correlation-based trackers have been receiving attention because of the joint consideration of speed and performance. For our algorithm, we consider three types of correlation-based trackers, that are Kernelized Correlation Filters (KCF), Background-Aware Correlation Filters (BACF) and Discriminative Correlation Filter Tracker with Channel and Spatial Reliability (CSR-DCF) (Galoogahi et al., 2017; Henriques et al., 2015; Lukežič et al., 2018). While KFC is the fastest among all, it has the lowest performance and is susceptible to the change in the head's scale, and therefore it loses the head track very easily in our case. According to TrackingNet, CSR-DCF shows higher performance while BACF is faster (Müller et al., 2018). However, the OpenCV 3.4.2 implementation of CSR-DCF seems to be much faster from the original BACF. For the head detection problem, several scenarios are investigated and it is observed that there is not a significant performance difference between them. Therefore, CSR-DCF is selected due to its speed and ease of implementation.

The Intersection Over Union (IOU) measure is used to establish the connection between the frames. However, to tackle the problems of the change of head sizes in the head detection algorithm, we propose a modification to IOU, in that the area of intersection of the two boxes is divided by the area of the smaller box, instead of the area of union. The boxes of foreground head candidates (see Figure 1) from the previous frame are tracked and compared with the boxes of foreground head candidates of the upcoming frame. If the IOU values of the box pairs exceed a certain threshold, they are assumed to belong the same head. If an IOU value is below the threshold, this means that, in the upcoming frame, the head detection failed to detect the head. Therefore, the tracked missing heads are also added to the list of found heads of the upcoming frame.

2.2 Proposed Person Counting Approaches

In this section, three different approaches have been proposed to estimate the number of people in the classroom by using the tools explained above. These are (i) the average counting method that takes the average of the people count from the related frames, (ii) the entrance counting method that counts the entering and exciting people from the entrance of the classroom and (iii) the combined counting method, which is a hybrid of the average counting method and entrance counting method.

The Average Counting Method

The flow diagram of the average counting method can be seen in Figure 2. The algorithm counts the number of heads frame by frame. The average of all the counts estimates the number of people in the environment.

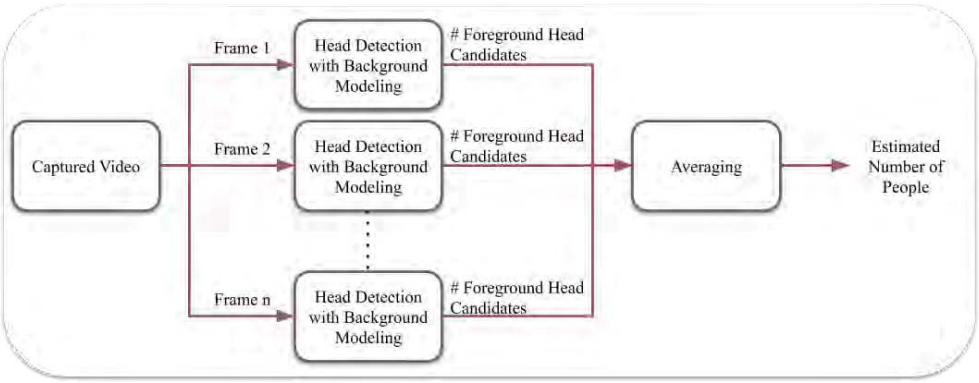


Figure 2: Average Counting Method

For this method, we experimented with the videos of the cameras installed in the classroom (see Section 3). To be able to visually cover the whole room, four cameras are used such that every camera is responsible for a fixed region of the room (see Figure 4). This results in some undesired intersections between the regions because of the projection of the world to the camera scene. From the videos, only every n^{th} frames are considered. This n value is changed to 1, 3, 5 and 10 and it is observed that there is not a significant change in performance.

The Entrance Counting Method

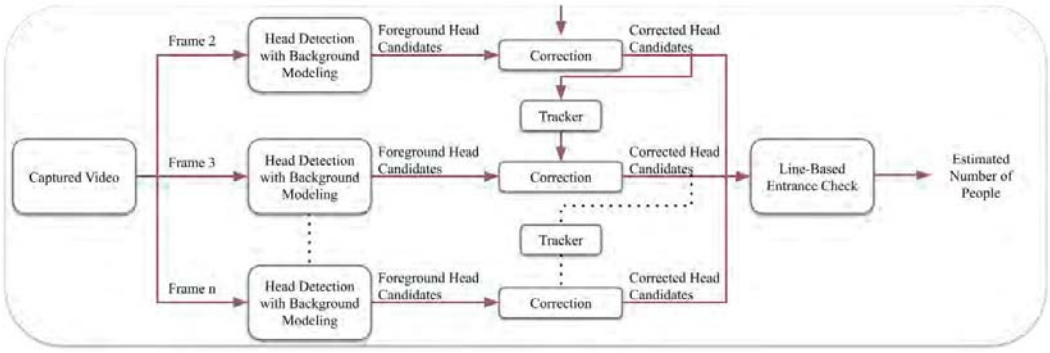


Figure 3: Entrance Counting Method

The entrance counting method counts the people entering and exiting the room through the door. An important difference between this method from the average counting method is that it includes a tracker in order to introduce temporal information into the algorithm. The flow diagram of the method can be seen in Figure 3. The correction process seen in the figure is the method that combines the head candidates of the upcoming frame with the tracked heads of the previous frame (see Section 2.1 for details). The entering and exiting people are counted with a basic *line* approach. If the center of a head passes from one side of the line to the other side, the head is assumed to enter or exit the environment.

There are three cameras used in this method. The average counts of these three cameras yield the estimate of the number of people in the room. As a side note, in order to deal with the detection of backs of the people which may yield to a double count for a person, thresholds on the size of the boxes of the head candidates are applied differently for every door camera and these are determined by visual experiments.

The Combined Counting Method (CCM)

From the tests, it is observed that the average counting method and the entrance counting method can outperform one another in different circumstances. The average counting method is prone to occlusion and image resolution due to long camera distances whereas the entrance counting method

works better when people enter the classroom one person at a time. However, its performance is drastically affected by severe occlusions, especially in cases of large numbers of people exiting the classroom at once. Therefore, we combine these two methods in CCM in such a way that they are alternately used when required. In this combined method, the algorithm starts with the entrance counting method. In case of severe occlusions at the entrance, the average counting method takes over. When the method detects an increasing trend in the people count, the combined method switches back to the entrance counting method and takes the reference of people to count from the average counting method. As such, the algorithm switches back and forth between the two methods to complement each other.

3. Experiment Results

This section presents the results concerning the proposed occupant counting method. In particular, we propose two different experiments. Experiment 1 concerns a quantitative comparison between the proposed approach, the standard occupancy templates and the ground truth. In Experiment 2, we evaluate the results of the proposed approach in an energy model and quantify the degree of change in terms of different metrics concerning building energy performance.

3.1 Experiment setup and dataset

The cameras used in this work are IP cameras with 1280x720 resolution at 10 fps, 130° camera angle, and H.264 video encoding. The cameras record videos only when motion is detected and are directly uploaded to the Amazon cloud services. Seven cameras were installed in a classroom of 331 m^2, three of which are pointed at the door (door cameras), while the other four are pointed at the classroom (room cameras) as can be seen in Figure 4. The door cameras are used for the entrance counting method. The classroom is used as a design studio for an architecture department, where a high number of people and unusual occupancy patterns are expected due to the students' prolonged study hours. The proposed method was used for person counting during the same week. Videos were recorded for a duration of a week (24-30 December 2018). The results were registered in a resolution of 10 minutes, and are timestamped in data sheets.

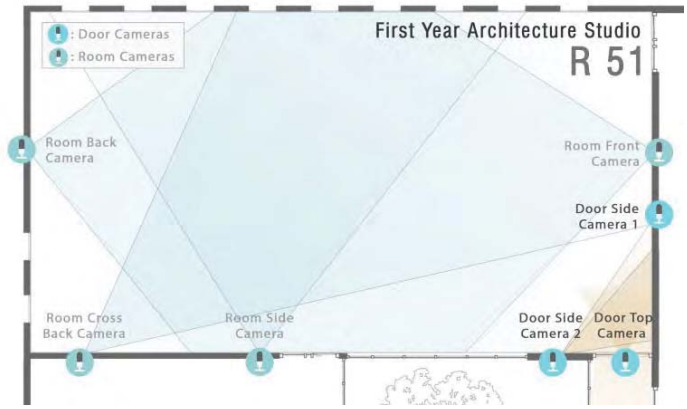


Figure 4: The plan layout of the classroom showing the cameras

The dataset used in this project is SCUT-HEAD which is a large-scale head detection dataset and contains 2000 images with 67321 annotated heads from the same class environment and same camera angle and 2405 images with 43930 annotated heads which are crawled from the Internet (Peng et al., 2018). In addition, 535 images from the test environment in this work were added to the dataset. Approximately 200 of 535 images that do not contain people or heads were added as negative samples. This aims to make sure that the system performs better at ignoring objects that appear like heads, such as curvy-shaped chairs and backpacks.

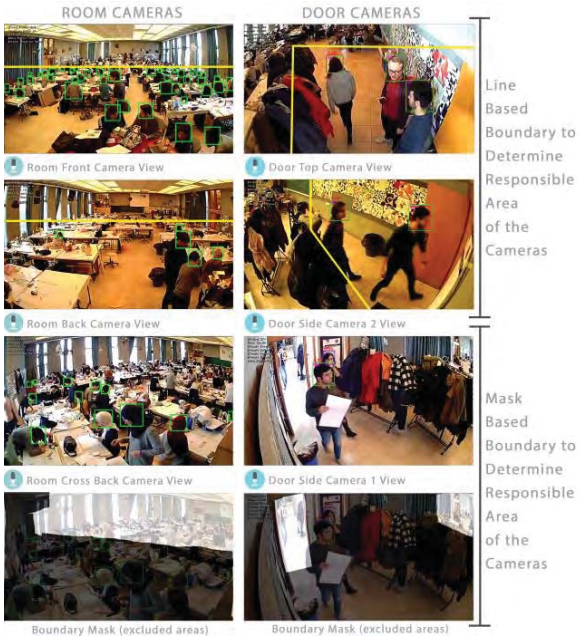


Figure 5: Views of the classroom under different camera angles

3.2 Results for Experiment 1

In Experiment 1, the proposed approach is benchmarked against two ground truth measurements manually recorded during the first day of the video recordings (24 December, from 9:30 AM to 6:30 PM) every 30 minutes. The first was In-Situ Measurement (ISM), where the people in the classroom were manually counted by a person in the room. ISM isn't completely reliable, as it was not possible to control the human movement or new people entering the classroom during the counting process. The second was Video Recording Measurements (VRM), where people were manually counted from the video recordings for the same day, therefore inheriting the occlusion and human recognition problems in the proposed algorithm.

The results can be found in Figure 6 and Table 1. We notice that although the trends are similar, there are some discrepancies between the measurement results, largely due to the fact that the reference point is taken from the average counting method. Therefore, even if the entrance counting method were very precise, the error in the reference point would still exist in the system and result in an offset. For both ground truth measurements in Table 1, it is observed that some of the high percentage errors are caused by mass exiting from the door at the end of the class, i.e. at 12.30PM and 5.30PM. This is observed to be due to the fact that the algorithm does not switch to the average counting method on time. Another reason for high percentage errors stems from the cases with a few numbers of people being present in the environment. For example, at 9.30AM, while the error is very low (3 people), the percentage error appears to be 30%.

Table 1. Comparative analysis between the two ground truth measurements and the proposed approach

Time	In-situ measurement (ISM)	Video recording measurement (VRM)	Combined Counting Method (CCM)	Percentage Error (ISM vs. CCM)	Percentage Error (VRM vs. CCM)
9:30	9	10	11.66	29.56	16.60
10:00	10	11	11.86	18.60	7.82
10:30	19	25	23.00	21.05	8.00
11:00	116	92	111.00	4.31	20.65

11:30	125	95	110.00	12.00	15.79
12:00	115	105	119.00	3.48	13.33
12:30	30	28	34.60	15.33	23.57
13:00	35	33	36.00	2.86	9.09
13:30	58	62	56.53	2.53	8.82
14:00	101	86	97.00	3.96	12.79
14:30	98	86	98.66	0.67	14.72
15:00	103	95	108.00	4.85	13.68
15:35	106	92	105.40	0.57	14.57
16:00	111	98	111.40	0.36	13.67
16:30	114	108	116.00	1.75	7.41
17:00	118	103	120.00	1.69	16.50
17:30	35	32	21.60	38.29	32.50
18:00	19	24	18.50	2.63	22.92
18:30	16	15	16.86	5.38	12.40
Average	70.42	63.16	69.85	0.82	10.59

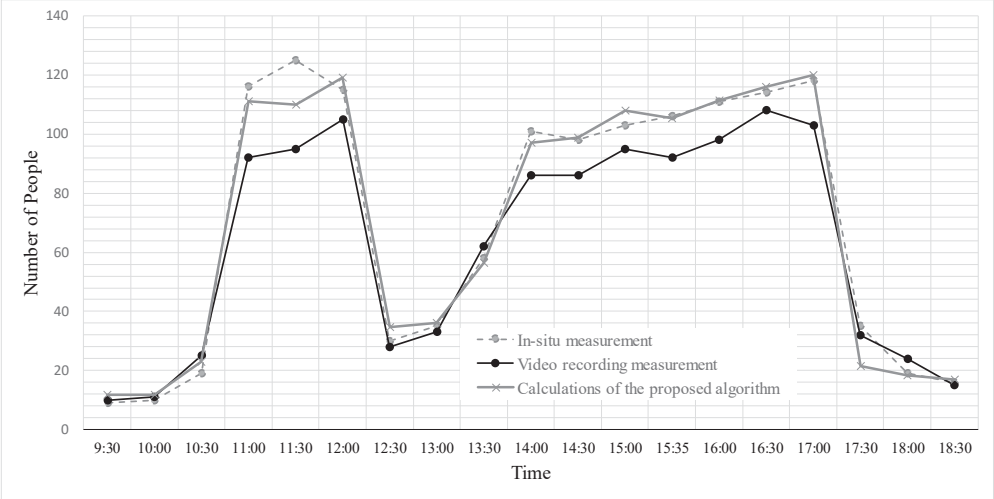


Figure 6: The results of ISM, VRM and the proposed algorithm (24 December, 9:30AM-6: 30 PM)

3.3 Results for Experiment 2

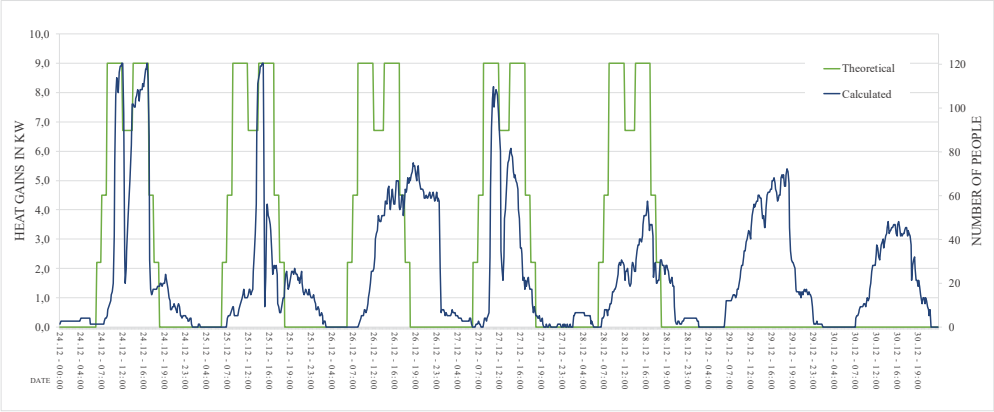
For Experiment 2, building energy simulations were used to quantify the impact of occupancy count on building performance metrics. Two separate energy models were generated for the same classroom. The first directly used the occupancy data calculated using the proposed approach. For the second model, namely the theoretical model, an existing standard dataset for occupant density fraction schedule for classrooms was used by multiplying the hourly schedule values by the maximum number of people expected in the classroom (set to 120). Occupant activity level is set to 144 W/person, which corresponds to standard office work. The following building setup is used for both energy models. Heating setpoint and setback temperatures are 22 C° (5:00AM-6: 00 PM during weekdays) and 18 C° respectively, while natural ventilation is set to activate after 25 C°. The thermal transmittance values of the room surfaces are 3.316, 0.577, 2.84 and 2.6 W/m2-K for the walls, ceiling, floor, and window respectively. Simulations were run for one whole week in mid-spring (15-21 May) and winter (24-30

December), with a frequency of 10 minutes. The resulting performance data is used for benchmarking. The first benchmark metric is total occupant heat gain (kW), which is calculated as the activity level times the number of occupants. The second metric is total heating energy use (kW), calculated only for winter simulations. The last metric is indoor air temperature (C°), calculated only for mid-spring simulations.

It must be noted first that that total number of people are 1803,7 and 2424 by the proposed algorithm and the theoretical dataset respectively. This amounts to a -%25.59 difference between the two datasets. The simulation results indicate that, due to the difference between the total number of people used in the two energy models, the total heat gain, which is independent of seasonal climatic conditions, is subject to decrease with the same amount as occupant count (-%25.59) from theoretical to calculated (Table 2). This places stress on the building energy balance, such that critical building performance criteria change as well. To compensate for the reduced internal load, the heating energy use in winter increases by %12.77 in the energy model that uses data from the proposed approach. In summer, the proposed method estimates the average and maximum indoor air temperature values 0,32 C° and 1,50 C° lower than the theoretical approach respectively. Imprecise performance predictions due to faulty occupancy estimation can give way to uninformed decision-making in buildings. Incorrect estimation of heating loads or summer indoor temperatures can misinform HVAC sizing, or lead to incorrect decision-making for building improvement by over- or under-estimating performance problems.

Table 2. Building performance metrics for benchmarking

	Total Heating Energy Use (kW)	**Total Occupant Heat Gain (kW)**	**Average Indoor Temperature (C°)**
Theoretical occupancy	5216.5	2424.0	26.98
Calculated occupancy	5882.6	1803.7	26.67
Difference	%12.77	-%25.59	-%1.15



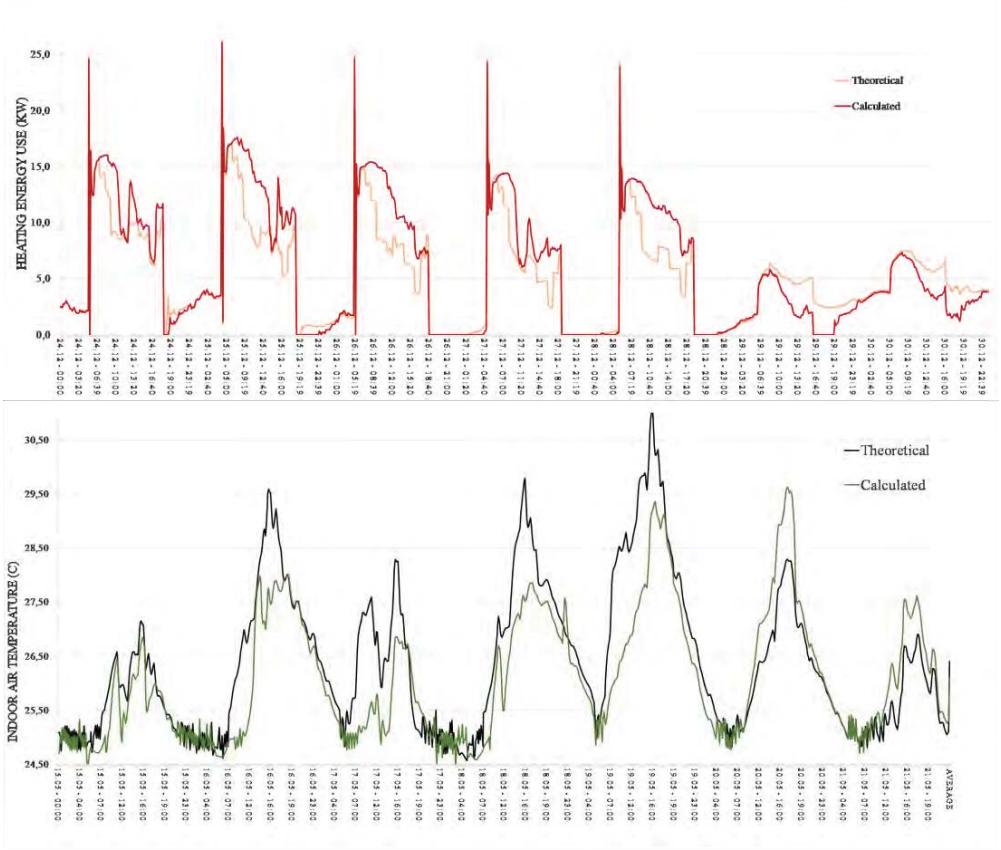


Figure 7. People heat gain, building heating energy use (winter) and indoor air temperature (mid-spring) values

4. Discussion and Conclusion

This paper proposed an approach that counts the people in a room from video content through head detection and tracking. A hybrid method was developed that uses the head detection algorithm for estimating the number of people. The method was evaluated in a classroom environment for a week. Despite the occlusion and image resolution problems due to the high number of people and the size of the room, it is observed that the highest average percentage error was 11%. In the future, we plan to improve the proposed approach by increasing the number of samples of the dataset for increased performance. Moreover, the resulting occupant count was used in the energy simulations of the same room, and the acquired results were benchmarked against an existing standard occupancy schedule. The results show that the difference in the two data sets has a critical influence on building performance metrics of heating energy use and indoor air temperature.

Acknowledgments

The authors would like to acknowledge the support by the TUBITAK – British Council Newton – Katip Celebi Fund, Grant No. 217M519.

References

Agency, I. E. (2018). *EBC Annex 66 Definition and Simulation of Occupant Behavior in Buildings*.

Akkaya, K., Guvenc, I., Aygun, R., Pala, N., & Kadri, A. (2015). *IoT-based Occupancy Monitoring Techniques for Energy-Efficient Smart Buildings*. 58–63.

Amin, I. J., Taylor, A. J., Junejo, F., Al-Habaibeh, A., & Parkin, R. M. (2008). Automated people-counting by using low-resolution infrared and visual cameras. *Measurement: Journal of the International Measurement Confederation*, *41*(6), 589–599. https://doi.org/10.1016/j.measurement.2007.02.010

Andriluka, M. (2008). People-Tracking-by-Detection and People-Detection-by-Tracking. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. https://doi.org/10.1109/CVPR.2008.4587583

Benezeth, Y., Laurent, H., Emile, B., & Rosenberger, C. (2011). Towards a sensor for detecting human presence and characterizing activity. *Energy and Buildings*, *43*(2–3), 305–314. https://doi.org/10.1016/j.enbuild.2010.09.014

Day, J., Theodorson, J., & Van Den Wymelenberg, K. (2012). Understanding controls, behaviors, and satisfaction in the daylit perimeter office: A daylight design case study. *Journal of Interior Design*, *37*(1), 17–34. https://doi.org/10.1111/j.1939-1668.2011.01068.x

Duarte, C., Van Den Wymelenberg, K., & Rieger, C. (2013). Revealing occupancy patterns in an office building through the use of occupancy sensor data. *Energy and Buildings*, *67*, 587–595. https://doi.org/10.1016/j.enbuild.2013.08.062

Dziedzic, J., Yan, D., & Novakovic, V. (2017). Occupant migration monitoring in residential buildings with the use of a depth registration camera. *Procedia Engineering*, *205*(1877), 1193–1200. https://doi.org/10.1016/j.proeng.2017.10.352

Galoogahi, H. K., Fagg, A., & Lucey, S. (2017). Learning Background-Aware Correlation Filters for Visual Tracking. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*er, 1144–1152. https://doi.org/10.1109/ICCV.2017.129

Gilani, S., Brien, W. O., Gilani, S., & Brien, W. O. (2017). *Review of current methods, opportunities, and challenges for in-situ monitoring to support occupant modeling in office spaces occupant modeling in office spaces*. *1493*(July). https://doi.org/10.1080/19401493.2016.1255258

Han, J., & Bhanu, B. (2007). A fusion of color and infrared video for moving human detection. *Pattern Recognition*, *40*(6), 1771–1784. https://doi.org/10.1016/j.patcog.2006.11.010

Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2015). High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(3), 583–596. https://doi.org/10.1109/TPAMI.2014.2345390

Hoes, P., Hensen, J. L. M., Loomans, M. G. L. C., Vries, B. De, & Bourgeois, D. (2009). *User behavior in whole building simulation*. *41*, 295–302. https://doi.org/10.1016/j.enbuild.2008.09.008

Hong, T., Yan, D., D'Oca, S., & Chen, C. fee. (2017). Ten questions concerning occupant behavior in buildings: The big picture. *Building and Environment*, *114*, 518–530. https://doi.org/10.1016/j.buildenv.2016.12.006

Kelly Seryak, J., & Kissock, K. (2003). Occupancy and behavioral effects on residential energy use. In *American Solar Energy Society*.

Labeodan, T., Zeiler, W., Boxem, G., & Zhao, Y. (2015). Occupancy measurement in commercial office buildings for demand-driven control applications - A survey and detection system evaluation. *Energy and Buildings*, *93*, 303–314. https://doi.org/10.1016/j.enbuild.2015.02.028

Lam, K. P. (2015). International Energy Agency Energy in Buildings & Communities Programme Annex 66: Definition and simulation of occupant behavior in buildings. *In 9th Energy Forum*.

Karpacz, Poland.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multi-box detector. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9905 LNCS*, 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

Lukežič, A., Vojíř, T., Čehovin Zajc, L., Matas, J., & Kristan, M. (2018). Discriminative Correlation Filter Tracker with Channel and Spatial Reliability. *International Journal of Computer Vision*, *126*(7), 671–688. https://doi.org/10.1007/s11263-017-1061-3

Müller, M., Bibi, A., Giancola, S., Alsubaihi, S., & Ghanem, B. (2018). TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11205 LNCS*, 310–327. https://doi.org/10.1007/978-3-030-01246-5_19

O'Brien, W., Gunay, B., Tahmasebi, F., & Mahdavi, A. (2017). Special issue on the fundamentals of occupant behavior research. *Journal of Building Performance Simulation*, *10*(5–6), 439–443. https://doi.org/10.1080/19401493.2017.1383025

Peng, D., Sun, Z., Chen, Z., Cai, Z., Xie, L., & Jin, L. (2018). *Detecting Heads using Feature Refine Net and Cascaded Multi-scale Architecture*. 2528–2533. https://doi.org/10.1109/ICPR.2018.8545068

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-Janua*, 6517–6525. https://doi.org/10.1109/CVPR.2017.690

Reinhart C. F., & K., V. (2003). Monitoring manual control of electric lighting and blinds Reinhart, C.F.; Voss, K. NRCC-45701. *Lighting Research and Technology*, *35*(3), 243–260.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Sarkar, A., Fairchild, M., Salvaggio, C., Color, M., & Imaging, D. (2008). *Integrated Daylight Harvesting and Occupancy Detection Using Digital Imaging*. *6816*, 1–12.

Schweiker, M., Kingma, B. R. M., & Wagner, A. (2017). Evaluating the performance of thermal sensation prediction with a biophysical model. *Indoor Air*, *27*(5), 1012–1021. https://doi.org/10.1111/ina.12372

Toftum, J. (2010). Central automatic control or distributed occupant control for better indoor environment quality in the future. *Building and Environment*, *45*(1), 23–28. https://doi.org/10.1016/j.buildenv.2009.03.011

Yoshino, H., Hong, T., & Nord, N. (2017). IEA EBC annex 53: Total energy use in buildings—Analysis and evaluation methods. *Energy and Buildings*, *152*(July), 124–136. https://doi.org/10.1016/j.enbuild.2017.07.038

Zivkovic, Z., & Van Der Heijden, F. (2006). Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, *27*(7), 773–780. https://doi.org/10.1016/j.patrec.2005.11.005