www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

DEVELOPMENT OF CNN-BASED VISUAL RECOGNITION AIR CONDITIONER FOR SMART BUILDINGS

SUBMITTED: January 2020
REVISED: May 2020
PUBLISHED: July 2020
EDITOR: Žiga Turk
DOI: 10.36680/j.itcon.2020.021

Qian Huang, Assistant Professor
School of Architecture, Southern Illinois University Carbondale, IL, USA
qhuang@siu.edu

Kangli Hao, Software Algorithm Research Scientist
Kneron Inc., San Diego, CA, USA
kangli@kneron.us

SUMMARY: *Demand-driven heating, ventilation, and air conditioning (HVAC) operations have become very attractive in energy-efficient smart buildings. Demand-oriented HVAC control largely relies on accurate detection of building occupancy levels and locations. So far, existing building occupancy detection methods have their disadvantages, and cannot fully meet the expected performance. To address this challenge, this paper proposes a visual recognition method based on convolutional neural networks (CNN), which can intelligently interpret visual contents of surveillance cameras to identify the number of occupants and their locations in buildings. The proposed study can detect the quantity, distance, and angle of indoor human users, which is essential for controlling air-conditioners to adjust the direction and speed of air blow. Compared with the state of the art, the proposed method successfully fulfills the function of building occupant counting, which cannot be realized when using PIR, sound, and carbon dioxide sensors. Our method also achieves higher accuracy in detecting moving or stationary human bodies and can filter out false detections (such as animal pets or moving curtains) that are existed in previous solutions. The proposed idea has been implemented and collaboratively tested with air conditioners in an office environment. The experimental results verify the validity and benefits of our proposed idea.*

KEYWORDS: *Smart Air Conditioner, Occupancy Counting, Convolutional Neural Networks, False Detection*

REFERENCE: *Qian Huang & Kangli Hao (2020). Development of CNN-based visual recognition air conditioner for smart buildings. Journal of Information Technology in Construction (ITcon), Vol. 25, pg. 361-373, DOI: 10.36680/j.itcon.2020.021*

1. INTRODUCTION

According to the U.S. Energy Information Administration (EIA, 2019), building energy consumption accounts for 20% of global energy usage, and a large part of electricity in buildings is used for heating, ventilation, and air conditioning (HVAC). The size of HVAC systems usually meets the full-load heating or cooling requirements, which correspond to the maximum occupancy level of buildings. In traditional buildings, regardless of the actual building occupancy level, HVAC systems are controlled and operated according to the thermostatic settings or handheld remote settings. When building occupants are moving, failure to change HVAC settings in time can result in wasted energy. Therefore, researchers expect future smart buildings to be able to detect indoor occupancy levels and real-time occupant locations in thermal zones of buildings, thereby providing demand-based HVAC service levels (Erickson et al., 2009; Agarwal et al., 2010; Yang et al., 2012; Ekwevugbe et al., 2013; Lu, 2018). Typical thermal zones of buildings can be floors, rooms, or even dining table or sofa areas. In order to take into account the impact of the number of occupants and activities, the researchers of Lim et al. (2016) presented an online HVAC-aware occupancy scheduling scheme. Their experimental study showed that HVAC operation can save up to 12% of energy. Also, researchers in (Jain and Madamopoulos, 2016) proposed an accurate occupant distribution mapping framework for efficient HVAC operation.

Existing building occupancy detection involves various microscale sensors, including passive infrared sensors, acoustic sensors, CO_2 sensors, carbon dioxide sensors, camera-based motion sensors, and so on. Passive infrared (PIR) sensors are widely used to control bathroom and hallway lighting. When a warm object (human or animal) passes by, PIR sensors detect the change in radiated infrared energy, and then turn on/off lighting. However, PIR sensors cannot detect stationary persons or animals, such as sitting on a toilet or sleeping on a bed. Besides, even though PIR sensors can detect the presence of occupants, they are unable to count the number of occupants. Therefore, even if PIR sensors retain good privacy protection, they cannot be used to detect the building occupancy level and distribution for demand-driven HVAC control (Raykov et al., 2016). In 2019, the researchers of (Huang et al., 2019) presented a prototype of active infrared-based occupancy counting systems, which is easy to use and has a higher counting accuracy. Yet, the disadvantage of this method is that it is unable to detect stationary persons.

In the past few years, researchers have attempted to use acoustic sensors along with signal processing algorithms to estimate indoor room occupancy levels (Kelly et al., 2014; Huang et al., 2016). However, limited research has been conducted to reduce the interference of background sounds, such as dog barking or TV/music playback. To deal with this challenge, several signal processing algorithms have been proposed to measure ambient acoustic level and cancel out background noise in buildings (Huang, 2018). Similar to PIR sensors, acoustic sensors cannot detect silent people, and it is also difficult to use acoustic sensors to precisely distinguish the number of people in a thermal zone and their indoor locations. Researchers have investigated using CO_2 sensors to infer the number of building occupants (Nassif, 2012). It is known that indoor CO_2 levels depend on many factors, such as room size, the type and setting of HVAC equipment, the number of building occupants, and the opening status of doors/windows. Since the number of occupants is one of these factors, there is no explicit relationship between the number of occupants and CO_2 levels. Besides, the awareness of indoor CO_2 levels does not reveal the distance and direction between building occupants and air conditioners (Sun et al., 2011). In addition, because the carbon footprint of pets is comparable to humans, indoor CO_2 levels are also largely affected by the presence and activities of dogs and cats. According to the 2015-2016 APPA National PET Owners Survey (NPOS, 2015), 65% of households in the United States have pets, which is equivalent to 79.7 million households. The United States has 54.4 and 42.9 million families with dogs and cats, respectively. Some homeowners are more likely to have multiple pets.

Emerging camera-based motion sensors (such as Google Nest) have reduced false alarms to some extent by analyzing motion patterns. These pattern features help determine differences between objects, such as swaying trees and suddenly opened doors. However, false alarms are present, and stationary persons cannot be detected.

From the above discussion and Table 1, it is obvious that these existing indoor occupancy detection and counting methods cannot fully meet the rigid requirements of next-generation smart air conditioners (Labeodan et al., 2015). To address these challenges, it is interesting to collaborate on the design of smart building systems and information technology for energy savings in buildings (Ortega et al., 2015; Huang et al., 2017). In this work, we explore the design of next-generation CNN-based visual recognition air conditioner. With the help of artificial intelligence technology and surveillance cameras, our equipped air conditioners can adjust their operation based on the perceived user movements, patterns, and surrounding environments. This smart air conditioner can accurately

detect spaces that are currently occupied by human users, and then provide cooling services to these occupied areas only with the appropriate breeze strength and direction, rather than the entire thermal space.

Table 1: Comparison of Existing Building Occupancy Detection and Counting Methods

Detection Mechanisms	Advantages	Disadvantages
Passive infrared (PIR)	Low-cost; privacy protection	Unable to detect stationary building occupants; Unable to count the number of occupants; Unable to distinguish other moving objects
Sound level	Low-cost; privacy protection	Limited accuracy; Subject to nearby noise interference; Unable to detect silent people
CO_2 level	Privacy protection	Limited accuracy of occupancy counting; Unable to distinguish animal interferences
Camera-based motion	Improved accuracy	Poor privacy protection; Unable to detect stationary people
Active infrared	High accuracy; Low-cost;	False alerts when people stay in infrared pathways;

Regarding the contribution of the body of knowledge, this paper makes the following contributions: (1) based on rapid advances of convolutional neural networks, we utilize a deep-learning visual recognition technique to perform occupancy detection. In addition to accurately sensing the number of persons in the sight of surveillance cameras, this technique can also obtain the distance and direction of each person in the range, regardless the person is moving or stationary, making sounds or keeping silent. The proposed technique is non-intrusive and highly accurate for room occupancy counting and localization towards energy-efficient buildings. Furthermore, this method can filter out false alarms caused by animal interference. (2) In order to implement and test the proposed idea, we have realized it in a standalone computing system (i.e., Rockchip RK3399 hardware platform) and tested its performance in an office environment. The experimental results validate our proposed idea. The high detection accuracy of 98% validates the trained YOLO neural network architecture. The average estimation errors for distances and angles are 17% and 10%, respectively. These results indicate that the use of the YOLO neural network in the Rockchip RK3399 computing platform is an appropriate choice for building occupancy detection and positioning applications. This method successfully fulfills the function of building occupant counting, which cannot be realized when using PIR, sound, and carbon dioxide sensors.

2. SYSTEM DESIGN OF SMART AIR CONDITIONERS

2.1 Introduction of CNN-based Visual Recognition

As a basic research area in the field of computer vision, object recognition tries to identify what objects exist in images, and report the position and orientation of these objects in the images. These objects may be human bodies, faces, cars, or animals, etc. In the past few years, human face and body detection have received extensive attention. A type of object usually has its own special features, which help to classify itself in computer vision. For example, the special feature of all circles is a circular shape. When looking for circles, people are seeking objects at a certain distance from a point (i.e., the center). Similarly, when looking for squares, people are seeking objects that are perpendicular at four corners and have equal side lengths. Special features for heads, arms, legs, torso, skin color, and arm distance can be used for human body recognition.

Traditional machine learning methods must first define special features, and then use support vector machine (SVM) or other techniques to classify objects. With the recent availability of high-performance computing platforms (such as GPU cards) and large image databases such as ImageNet (Deng et al., 2009), it is feasible to run deep learning approaches to improve the accuracy of the object classification. So far, deep learning has greatly surpassed these traditional machine learning approaches, and deep learning can perform end-to-end object detection without defining special features in advance.

The concept of neural networks was initially inspired by human brains. Then, various neural network architectures have been developed for many years (Sze et al., 2017; Gu et al., 2018; Zheng et al., 2020). In 1998, researchers introduced convolutional neural networks to classify hand-writing digit numbers. Convolutional neural network (CNN) is a kind of deep neural networks. Compared with traditional machine learning methods (such as SVMs),

CNN greatly improves the accuracy of object classification. The CNN technique has been used to automatically detect workers and heavy equipment on construction sites (Fang et al., 2018) and detect hardhats worn by construction personnel (Wu et al., 2019).

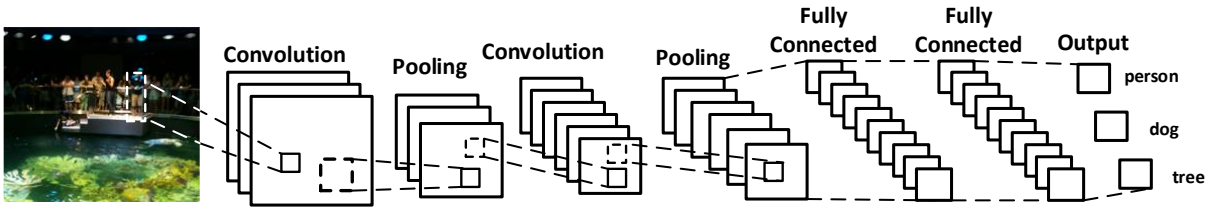

Figure 1. Processing Flow of Generic Machine Learning Techniques for Room Occupancy Detection

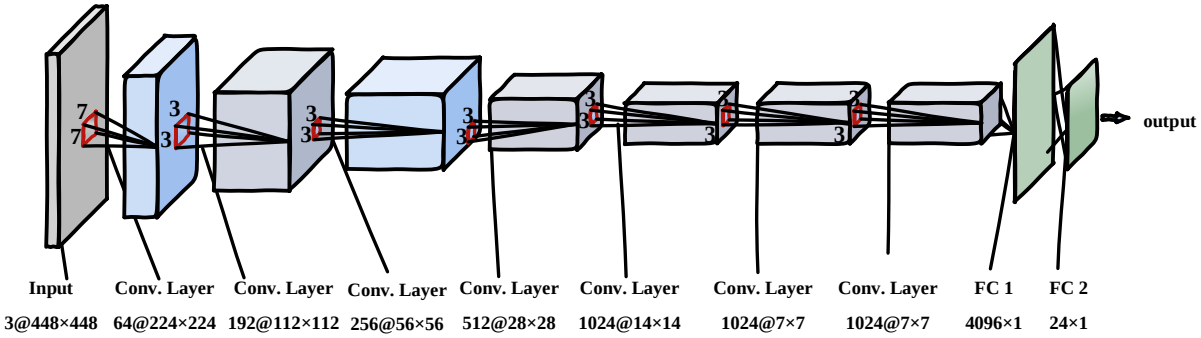

Figure 2. A general architecture of convolutional neural network

As shown in Figure 2, CNN comprises an input layer, an output layer, and multiple hidden layers. A large number of parameters in these layers will be determined by training. A CNN architecture involves three main types of hidden layers: convolutional layers, pooling layers, and fully-connected layers. The convolutional layer is a key building block of CNN because it performs most computations. The function of a convolutional layer is to extract features from its preceding layer through the mathematical operation of convolution. In the pooling layer, features are down-sampled along with spatial width and height to reduce the dimension of special features. In addition to reducing the number of parameters and computation complexity in neural networks, pooling can improve the network robustness to small shifts and distortions. The fully-connected layer calculates the classification scores to help finalize object recognition decisions. In this way, CNN transforms input images layer by layer from the original pixel values to the final classification scores. If the human body scores are the highest among all types of objects or exceed a certain threshold, the CNN-based visual recognition algorithm will output the human bodies detected in the input images. In order to introduce nonlinear effects into neural networks, people usually apply a nonlinear activation layer after the convolution or fully-connected layers. Overall, CNN is a clever combination of linear and nonlinear layers.

Training CNN is the same way as training traditional neural networks. Through backpropagation, stochastic gradient descent algorithms (Ruder, 2017) are adopted to adjust parameter values in all network layers. Once the training process is completed, the parameter values are finalized for object classification.

2.2 CNN Model Training and YOLO Architecture

In this study, we selected publicly available human body datasets (Dalal and Triggs, 2005; Wang et al., 2007; Dollar et al., 2012) as training and evaluation samples. In order to make up for the lack of human sitting and lying down in these datasets, we used cameras to collect about 15,000 human images in various poses in office and home environments.

In order to enable CNN-based visual detection algorithms to run on resource-constrained edge devices (such as surveillance cameras and HVAC controllers), the human detection algorithm we developed should be optimized so that it has low requirements for computing resources, memory footprint, and power consumption. Besides, the developed algorithm is supposed to recognize multiple human bodies in real time, so HVAC equipment can adjust its control in time. Based on the above considerations, we chose the YOLO neural network architecture (Redmon et al., 2015; Redmon and Farhadi, 2018), because of its huge potential to run extremely fast on edge devices. Prior to the YOLO architecture, all object detection models (Ren et al., 2015) need to perform initial detection first to find regions of interest. Then, classification is applied only to these regions of interest. As YOLO wisely regards

object detection as a regression problem, it attempts to use a single neural network for detection and classification. Therefore, the YOLO neural network architecture can run much faster than previous methods.

General YOLO neural network architecture is illustrated in Figure 3. Behind the convolutional layers of the YOLO architecture are two fully connected layers. To reduce the depth size of special features from preceding layers, some convolutional layers use 1×1 convolutions. As illustrated in Figure 4, the YOLO neural network architecture divides an input image into many grid cells. A grid cell can only detect one object and a fixed number of bounding boxes. Each bounding box is associated with a confidence score. To eliminate repeated detection of the same object, YOLO uses non-maximum suppression to remove prediction with the lowest confidence score. YOLO pre-trains the convolutional layers in the ImageNet classification tasks, and then completes human detection training.

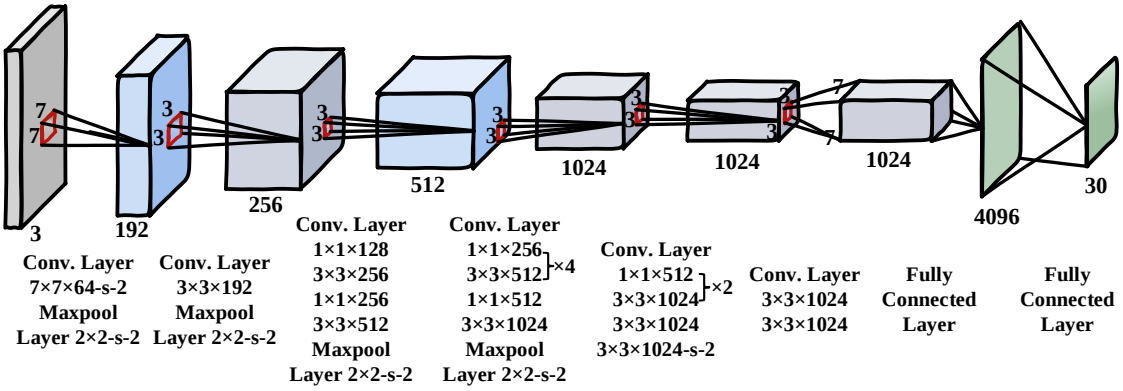

Figure 3. The Network Layers of a YOLO Architecture (Redmon et al., 2015)

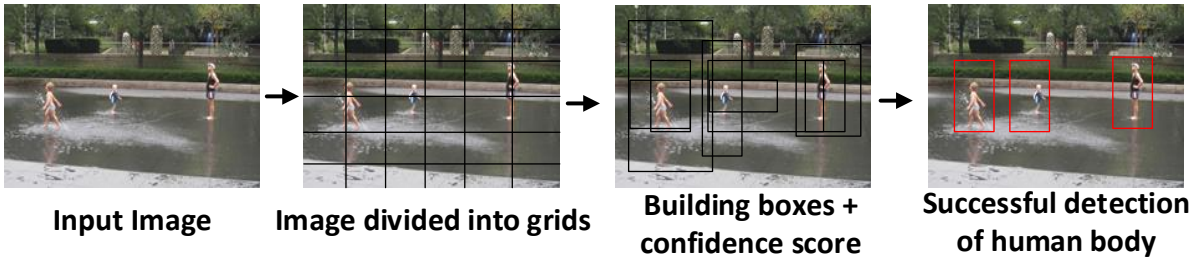

Figure 4. Illustration of How the YOLO Neural Network Architecture Detects Human Bodies

2.3 Distance and Direction Estimation from a Single Camera

So far, human body recognition has been achieved by using the YOLO neural network architecture. The remaining challenge is to predict the distance and direction between an air conditioner and a human body. For a system cost point of view, it is best to use only one camera to achieve this function. In order to obtain distance estimates using a single camera, we must use perspective knowledge (Stein et al., 2003), where the size of a human body in an image is a hint. Due to the unknown body type (slim, fat, adult, children, male, female, etc.), the width of human bodies varies from 40 cm to 70 cm. As a result, the distance based on the width of human bodies is estimated to have an accuracy of 70%-80%. In typical air conditioning applications, this level of accuracy is acceptable to customers.

Figure 5 shows a diagram of an imaging geometry including a pinhole (P) camera mounted at a height (H) parallel to the floor surface and an imaging plane (I) placed at a focal distance (f) from the pinhole. The distance between the human body and the camera is D. The position of the human foot is projected at the position Y on the image plane. The formula for calculating the distance (D) is: $D=f\times H/Y$, which can be derived from the similarity of triangles.

Next, let us discuss how to estimate the angular direction between the human body and the air conditioner. Figure 6 shows a diagram of an imaging geometry to detect the angular direction. This figure consists of a pinhole (P) camera, whose horizontal field angle (Afov_H) is mounted parallel to the floor surface, and an imaging plane (I) is placed at the focal distance (f) from the pinhole. The distance between the human body and the camera is D, and it is projected onto the image plane at the position (X). The angle direction of the person is $A = \arctan(x/f)$.

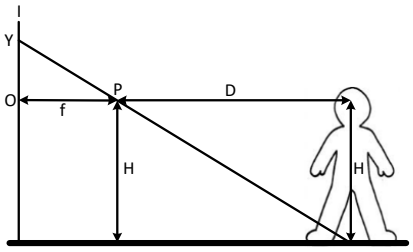

Figure 5. A Diagram of the Imaging Geometry for Distance Estimation

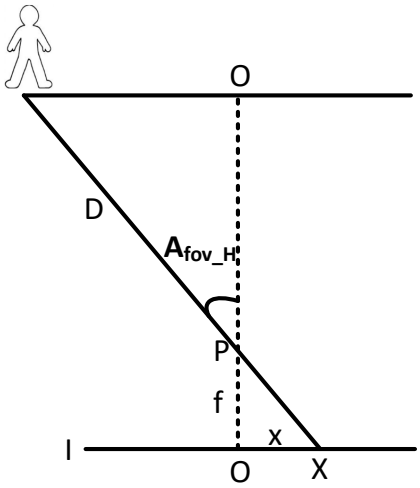

Figure 6. A Diagram of the Imaging Geometry for Angular Direction Estimation

Therefore, information about the number of occupants, their distances and angles from the air conditioner helps HVAC equipment serve areas that require automatic control of the strength and direction of the breeze.

2.4 Proper Image Camera Selection

If human bodies in the captured images or videos are too small, it is difficult to detect them. According to our test results in indoor environments, our visual recognition algorithm for human detection requires the minimum number of pixels of a human body in images to be 80 pixels. Although the viewing angle of image cameras is usually kept constant, different sizes in the field of view (FOV) are obtained by focusing the lens at different working distances.

Figure 7 shows the relationship between horizontal FOV (in length), focal length, camera size, working distance (WD), and angular FOV (AFOV). A_{FOV} is specified as the full angle (in degrees) related to the horizontal size (width) of the camera used with the lens.

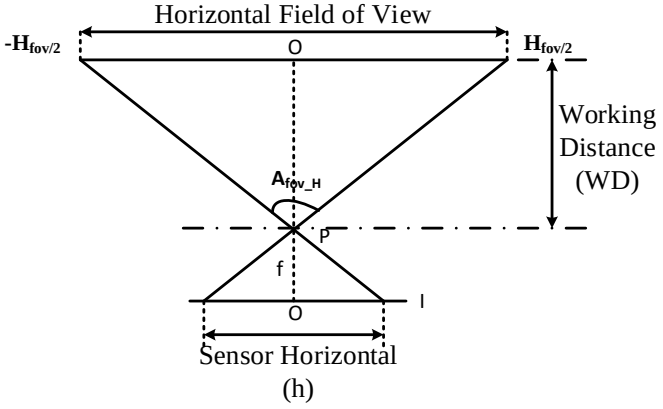

Figure 7. A Diagram of the Imaging Geometry for Angular FOV Calculation

For surveillance cameras in vision-based intelligent buildings, system designers focus on the range and distance that the surveillance cameras can monitor. If image cameras are installed in room corners, a horizontal FOV of 80-90 degrees is usually required. If image cameras are installed in hallways, a wide horizontal FOV is not required, but it is better to use a longer working distance. As shown in Figure 7, when the horizontal dimension (h) and focal length (f) are known, the calculation formula of angular FOV is $AFOV_{\_H}=2*\arctan(h/2f)=2*\arctan(\text{Horizontal FOV in length}/2/WD)$.

As shown in Figure 8, the most common aspect ratio for cameras is 4:3 (width: height = 4:3). The number marked in the middle of each yellow rectangle is the diagonal length. The number above each yellow rectangle is the type of image, and they are the size descriptors for the equivalent camera tube size. Table 2 lists the most commonly used 4:3 cameras on the market. For cameras installed in room corners, $AFOV_{\_H}$ >= 80 degrees and $AFOV_{\_V}$ > 50 degrees are usually required. Since cameras become very expensive as the size increases, the economical choice for cameras and lens is the 1/3`` type (width=4.8 mm, height=3.6 mm) and the focal length is 2.8 mm.

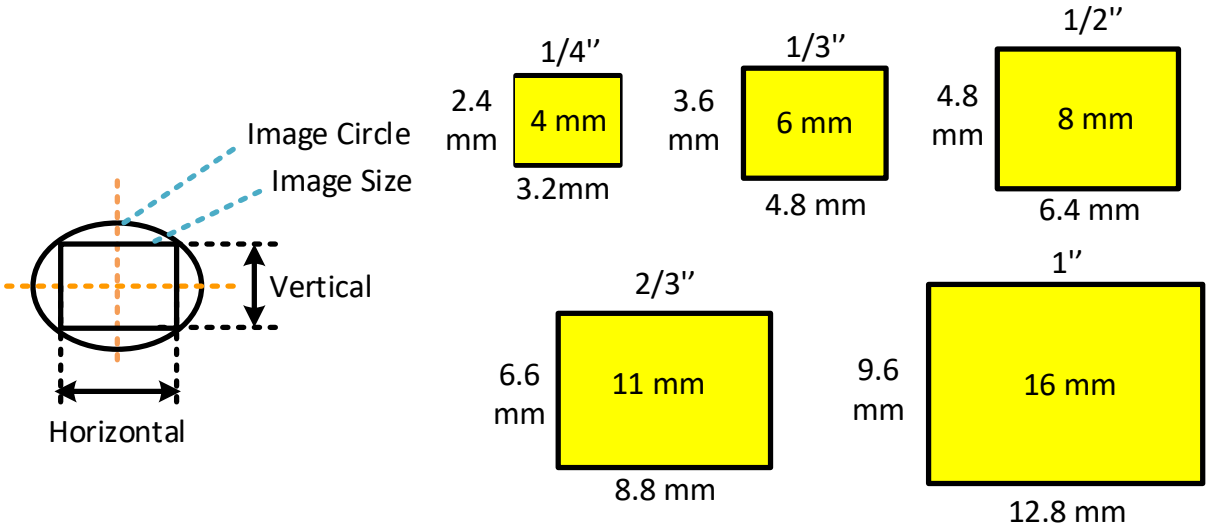

Figure 8. Illustration of Image Camera Size and Common Examples

Table 2: Summary of Existing Cameras with an Aspect Ratio of 4:3

	A focal length of 2.8 mm			A focal length of 3.6 mm			A focal length of 6.0 mm		
Type of 4:3 image camera	AFOV_D (degrees)	AFOV_H (degrees)	AFOV_V (degrees)	AFOV_D (degrees)	AFOV_H (degrees)	AFOV_V (degrees)	AFOV_D (degrees)	AFOV_H (degrees)	AFOV_V (degrees)
1/4''	71.1	59.5	46.4	58.1	47.9	36.9	36.9	29.9	22.6
1/3''	93.9	81.2	65.5	79.6	67.4	53.1	53.1	43.6	33.4
1/2''	110.0	97.6	81.2	96.0	83.3	67.4	67.4	56.1	43.6
2/3''	126.0	115.1	99.4	113.6	101.4	85.0	85.0	72.5	57.6
1''	141.4	132.7	119.5	131.5	121.3	106.3	106.3	93.7	77.3

Assuming that a human body of 40 cm or wider has at least 80 horizontal pixels in images, Table 3 shows the horizontal, vertical, and the total number of pixels required by the 4:3 cameras and the normalized computation time for different working distances, when the horizontal and vertical AFOVs are 80 and 50 degrees, respectively. For example, when the working distance is 5 meters, the required number of pixels is 1.57 million, which means 5 times the computation time for VGA images (640 × 480).

Table 3: Corresponding Pixel Resolutions and Normalized Processing Time

Working distance	Required horizontal pixels	Corresponding vertical pixels	Corresponding pixel resolutions	Normalized processing time with respect to VGA images (640×480)
3 meters	1,007	560	0.56 million	1.8 ×
4 meters	1,343	746	1 million	3.3 ×
5 meters	1,678	933	1.57 million	5 ×

3. EXPERIMENTS AND DISCUSSION

3.1 System Implementation

We have implemented and optimized the proposed visual-based human detection algorithm on several hardware computing platforms in Table 4. For example, Rockchip RK3399 SoC is a popular choice for home multimedia and face authentication. This platform has dual 2.0 GHz CPU clusters. One cluster contains two Cortex-A72 high-performance cores and the other cluster contains four Cortex-A53 cores. Rockchip RK3399 can also be embedded with Mali-T860 MP4 GPUs for computing acceleration. In this study, we have implemented a pure CPU version with multithreading and Arm NEON optimization, and a GPU version with OpenCL 1.2 for parallel computing. For server platforms, we have implemented the proposed algorithm on Intel Xeon CPUs, and tested the CPU version with OpenCL 1.2 on NVidia GeForce GTX 1080Ti, which is popular for affordable server platforms.

Table 4: Hardware Computing Platforms for Running our CNN Architectures

Option of Hardware Computing Platforms	Operating System	Speed (FPS) assuming the 1080p image/video input
Rockchip RK3399 (pure CPU)	32bit Debian	0.6
Rockchip RK3399 (with Mali T860 GPU)	32bit Debian	5
Intel Xeon W-2133 CPU @ 3.60 GHz	64bit Ubuntu	3.5
Intel Xeon W-2133 CPU @ 3.60GHz with NVidia GeForce GTX 1080Ti	64bit Ubuntu	195

Assuming 1080p image/video inputs, the performance is listed in Table 4. For example, it shows that Rockchip RK3399 (GPU version) can process 5 frames per second (FPS), which is equivalent to 0.2 seconds per frame. For edge-side processing (that is, using a SoC to process video streams), HVAC equipment adjusts the frequency of power inverters quickly and controls the speed and strength of the breeze. The PCB board in Figure 9 shows our Rockchip RK3399 platform, which contains several key components: USB port, memory, CPU, Wi-Fi, HDMI port. We can see that a camera is placed on top of the air conditioner to simulate the actual installation height and detect the presence and location of building occupants. For large buildings with many air conditioners, it is appropriate to use a server platform to handle all camera video streams. As shown in Table 4, the Intel W-2133 CPU @ 3.60GHz with NVidia GeForce GTX 1080Ti server can process 195 frames per second. The total power consumption including the air conditioning panel and display is only 20 watts.

Figure 9. Rockchip RK3399 Hardware Computing Platform and its Joint Operation with an Air Conditioner

3.2 Experimental Results

Figure 10 shows the human body recognition by a surveillance camera in an office environment. For accuracy measurement, we test the implemented system (i.e., Rockchip RK3399 platform) in an indoor office environment. The width of the office is greater than 6 meters, and the lighting intensity is greater than 80 lux.

As shown in Figure 11, we draw 5 arcs at radii of 1, 2, 3, 4, and 5 meters from the camera position, respectively. Four lines were drawn on the floor, marking 40 degrees to the right, 20 degrees to the right, the center, 20 degrees to the left, and 40 degrees to the left, respectively. One to eight testers stand or sit on one or more of the 25 points in Figure 11. These testers faced the camera in different ways (front, back, left, and right sides of the human body facing the camera). Then, we compare the results of the visual human body detection system with ground truth.

The test results show that the system we proposed can achieve an accuracy of 98% when counting the number of human bodies. Such high detection accuracy successfully validates the trained YOLO neural network architecture. This indicates that the use of the YOLO neural network in the Rockchip RK3399 computing platform is an appropriate choice for building occupancy detection and positioning applications. This method successfully fulfills the function of building occupant counting, which cannot be realized when using PIR, sound, and carbon dioxide sensors.

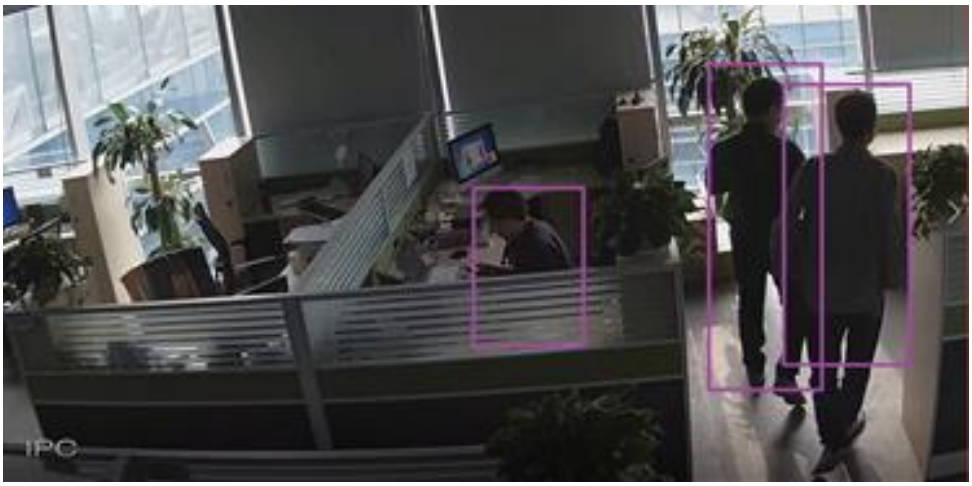

Figure 10. Human Body Recognition using Our Proposed Design through a Surveillance Camera

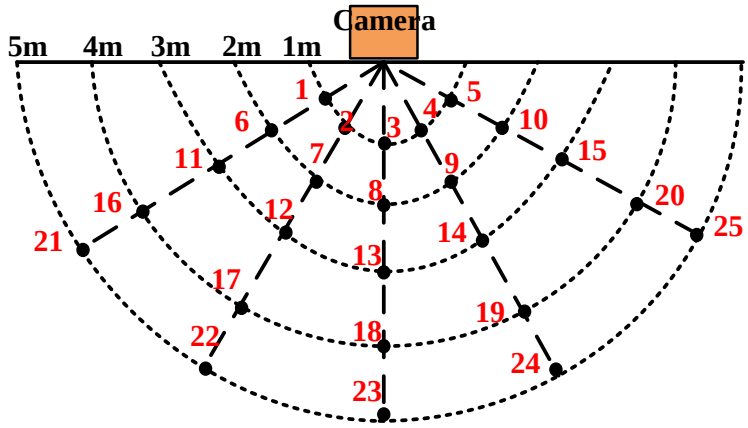

Figure 11. Illustration of the Five Arcs from the Camera Location in our Test Environment

Regarding distance and direction estimation, Table 5 lists the actual distance and angle values of these 25 points in Figure 11.

Table 5: Real Distance and Angle Values of 25 Points in Figure 11

Point	Real distance (meter)	Real angle (degree)	Point	Real distance (meter)	Real angle (degree)
#1	1	50	#14	3	110
#2	1	70	#15	3	130
#3	1	90	#16	4	50
#4	1	110	#17	4	70
#5	1	130	#18	4	90
#6	2	50	#19	4	110
#7	2	70	#20	4	130
#8	2	90	#21	5	50
#9	2	110	#22	5	70
#10	2	130	#23	5	90
#11	3	50	#24	5	110
#12	3	70	#25	5	130
#13	3	90			

Experiments are conducted to test the estimated distance and angle values, when testers have different directions towards the camera. The estimation results are summarized in Table 6, where the average estimation errors of distance and angle are 17% and 10%, respectively. In addition, we plot the relative errors of distance and angle estimates in Figure 12. We can see that when the distance between the human body and camera is shorter, a higher distance estimation accuracy can be found. For example, when the distance does not exceed 3 meters, the relative error is less than 10% regardless of the face direction. On the other hand, it is found in Figure 12 that at a distance of 3-4 meters, the relative error of angle estimation is very small (<5%).

Table 6: Estimated Distance and Angle Values of 25 Points in Figure 11

Point	Face toward the camera	Estimated distance (meter)	Estimated angle (degree)	Point	Face toward the camera	Estimated distance (meter)	Estimated angle (degree)
#1	front	1.2	54	#14	front	2.7	112
	back	1.1	53		back	2.7	112
	left/right side	1.2	52		left/right side	3.6	108
#2	front	0.9	80	#15	front	1.8	124
	back	0.8	80		back	1.8	124
	left/right side	0.8	84		left/right side	2.4	130
#3	front	0.9	92	#16	front	4.4	64
	back	0.8	89		back	4.8	71
	left/right side	0.8	80		left/right side	5.2	74
#4	front	0.9	111	#17	front	4.8	79
	back	1.0	114		back	4.8	79
	left/right side	1.2	112		left/right side	4.8	80
#5	front	0.8	120	#18	front	4.8	92
	back	0.9	119		back	4.8	92
	left/right side	0.8	118		left/right side	5.2	94
#6	front	1.6	56	#19	front	4.4	109
	back	1.8	60		back	4.4	107
	left/right side	1.8	65		left/right side	4.4	107
#7	front	2.0	74	#20	front	3.2	124
	back	2.0	74		back	3.6	124
	left/right side	1.8	73		left/right side	4.0	124
#8	front	2.3	93	#21	front	7.5	64
	back	2.2	95		back	7.0	71
	left/right side	2.2	94		left/right side	7.0	74
#9	front	1.5	109	#22	front	9.0	80
	back	1.4	110		back	9.0	80
	left/right side	1.4	110		left/right side	9.5	78
#10	front	1.9	114	#23	front	7.5	91
	back	1.9	110		back	7.5	91
	left/right side	2.1	117		left/right side	9.5	91
#11	front	2.7	64	#24	front	5.5	108
	back	2.7	71		back	9.5	108
	left/right side	2.7	64		left/right side	9.5	108
#12	front	3.0	77	#25	front	5	122
	back	3.0	79		back	5	122
	left/right side	3.9	78		left/right side	6.0	124
#13	front	3.0	92				
	back	3.0	92				
	left/right side	2.7	89				

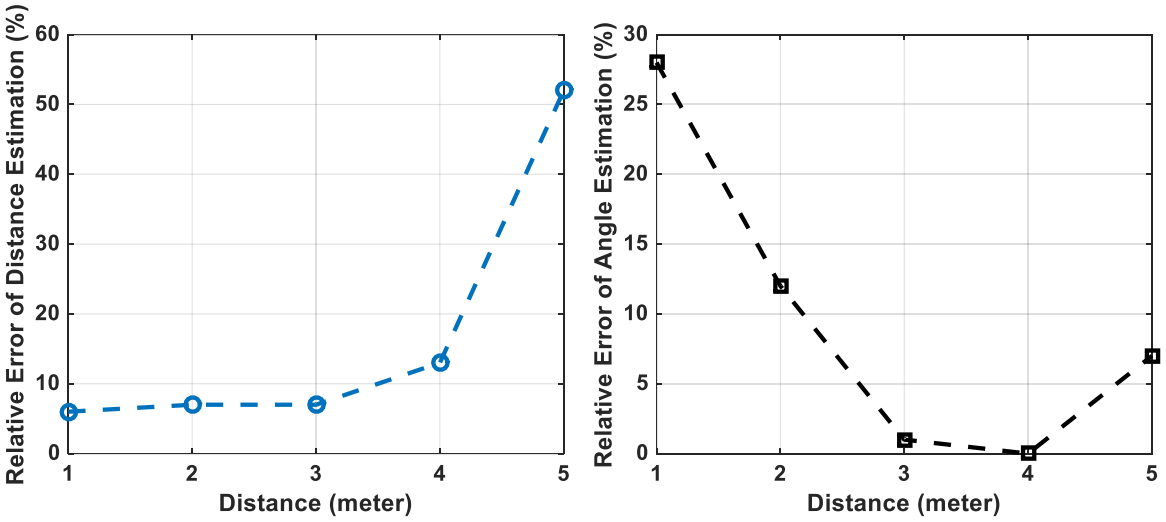

Figure 12. Relative Errors versus Distance Based on the Data in Table 6

3.3 Comparison

Table 7 summarizes the existing building occupancy detection methods in this work and literature. These existing approaches include passive infrared (PIR) sensors, sound sensors, CO_2 sensors, and camera-based motion sensors. We compare several aspects of occupancy count, stationary/silent human detection, animal interference, false alarm, distance detection, direction detection, implementation cost, and privacy protection. PIR sensors fail to count the occupancy quantity. Sound sensors cannot perceive silent people in buildings. Although CO_2 sensors and camera-based motion sensors can perform occupancy counting (even for stationary or silent persons), they are plagued by animal interference and high false alarms. This proposed work leads to high detection precision without animal interference. It supports the detection of stationary or silent persons, as well as distance and direction estimation. Its relatively high implementation cost is due to the use of a hardware computing platform to run the proposed YOLO neural network. Among all these mechanisms in Table 7, our proposed method is the most powerful solution, and its implementation cost is comparable to CO_2 sensors or camera-based motion sensors. This is because the nature of the CNN-based visual recognition method requires high computation, therefore, it is impractical to run visual recognition tasks on hardware devices with limited resources, such as micro-controllers. Regarding the privacy protection of building users, PIR, sound, and carbon dioxide sensors are superior to the use of camera-based motion sensors or CNN-based visual recognition. As a result, privacy protection is the main limitation of this study. Measures need to be taken to prevent adversaries or attackers from accessing visual data on the local camera side.

Table 7. Comparison with Existing Building Occupancy Detection Methods

Occupancy Detection Mechanisms	PIR Sensor	Sound Sensor	CO_2 Sensor	Camera-based Motion Sensor	CNN-based Visual Recognition (This Work)
Occupancy counting	Not capable	Not accurate	Not accurate	Limited	Accurate (98%)
Static/silent person detection	Not capable	Not capable	Capable	Capable	Capable
Animal interference	Cause false alert	Cause false alert	Cause false alert	Cause false alert	Can filter out
False alarms	High	High	High	High	Very low
Distance detection	Not capable	Not capable	Not capable	Capable	Capable
Direction detection	Not capable	Not capable	Not capable	Capable	Capable
Implementation Cost	Low (< $10)	Low (<$10)	Medium (around $100)	Medium (around $100)	Medium ($65 for Rockchip RK3399)
Privacy Protection	Good	Good	Good	Poor	Poor

4. CONCLUSION

This paper applies the state-of-the-art convolutional neural network technology to visual recognition in intelligent building applications. Specifically, we equip air conditioners with neural network algorithms to create smart air conditioners, which can detect the number, distance, and angle of indoor human occupants. Compared with previous methods, the proposed approach leads to an accuracy of 98% when detecting human bodies in various states: moving, stationary, vocal, silent, standing, sitting, front, side, and back. The average estimation errors for distances and angles of our method are 17% and 10%, respectively. Note the function of building occupant counting is hard to realize when using PIR, sound, and carbon dioxide sensors. Moreover, this method can filter out false alarms caused by animal interference. As a result, depending on the number of occupants and their locations, air conditioners can adjust the direction and speed of air blowing to achieve occupancy-driven energy-efficient HVAC operation.

The primary limitation of visual recognition is privacy protection. The challenge is how to prevent adversaries or attackers from accessing visual data on the local camera side, while allowing air conditioners to identify the number and recognize the location of building occupants. In future work, we will strive to study privacy-preserving techniques to avoid leakage of visual privacy. One basic idea is to transform visual images captured from air-conditioning cameras into a substitute image, which does not display sensitive human information such as faces. As a result, the risk of leaking sensitive visual information is reduced. In the future work, we will also develop new features for artificial intelligence air conditioners, including temperature control and wind speed control using gesture recognition, preference mode adaptation through face recognition, and house intrusion reminder, etc.

REFERENCES

APPA National PET Owners Survey (NPOS), available at https://petleadershipcouncil.org/pet-industry-news/2015-2016-appa-national-pet-owners-survey-generational-report

Agarwal, Y., Balaji, B., Gupta, R., Lyles, J., Wei M., & Weng, T. (2010). Occupancy-Driven Energy Management for Smart Building Automation. *Embedded Sensing Systems for Energy-Efficiency in Building*, pp. 1-6.

Dalal, N. & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 886-893.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Li, F. (2009). ImageNet: A Large-Scale Hierarchical Image Database. IEEE Conference on Computer Vision and Pattern Recognition, pp. 248-255.

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian Detection: An Evaluation of the State of the Art. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 34, No. 4, pp. 743-761.

Ekwevugbe, T., Brown, N., & Pakka, V. (2013). Real-Time Building Occupancy Sensing for Supporting Demand Driven HVAC Operations.

Energy Information Administration (EIA), available at https://www.eia.gov/todayinenergy/

Erickson, V., Lin, Y., Kamthe, A., Brahme, R., Surana, A., Cerpa, A., Sohn, M., & Narayanan, S. (2009). Energy Efficient Building Environment Control Strategies using Real-time Occupancy Measurements. *ACM Workshop on Embedded Sensing Systems for Energy Efficiency in Buildings*, pp. 19-24.

Fang, W., Ding, L., Zhong, B., Love, P., and Luo, H. (2018). Automated Detection of Workers and Heavy Equipment on Construction Sites: A Convolutional Neural Network Approach. *Advanced Engineering Informatics*, Vol. 37, pp. 139-149.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). Recent Advances in Convolutional Neural Networks. *Pattern Recognition*, Vol. 77, pp. 354-377.

Huang, Q., Ge, Z., & Lu, C. (2016). Occupancy Estimation in Smart Buildings using Audio-Processing Techniques. *ASCE International Conference on Computing in Civil and Building Engineering*, pp. 1413-1420.

Huang, Q., Lu, C., & Chen, K. (2017). Smart Building Applications and Information System Hardware Co-Design. *Big Data Analytics for Sensor-Network Collected Intelligence*, pp. 225-240.

Huang, Q. (2018). Occupancy-Driven Energy-Efficient Buildings using Audio Processing with Background Sound Cancellation. *Buildings*, Vol. 8, No. 6, pp. 1-16.

Huang, Q., Rodriguez, K., Whetstone, N., & Habel, S. (2019). Rapid Internet of Things (IoT) Prototype for Accurate People Counting Towards Energy Efficient Buildings. *Journal of Information Technology in Construction*, vol. 24, pp. 1-13.

Jain, S. & Madamopoulos, N. (2016). Ahorrar: Indoor Occupancy Counting to Enable Smart Energy Efficient Office Buildings. *IEEE International Conferences on Big Data and Cloud Computing*, pp. 469-476.

Jin, M., Liberis, N., Weekly, K., Spanos, C., & Bayen, A. (2015). Sensing by Proxy: Occupancy Detection Based on Indoor CO_2 Concentration. *IARIA Ubicomm*.

Kelly, B., Hollosi, D., Cousin, P., Leal, S., Lglar, B., & Cavallaro, A. (2014). Application of Acoustic Sensing Technology for Improving Building Energy Efficiency. *Procedia Computer Science*, pp. 661-664.

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*, Vol. 93, pp. 303-314.

Lim, B., Hijazi, H., Thiebaux, S., & Briel, M. (2016). Online HVAC-Aware Occupancy Scheduling with Adaptive Temperature Control. International Conference on Principles and Practice of Constraint Programming, pp. 683-700, 2016.

Lu, S. (2018). An Integrative HVAC System Featuring Adaptive Personalized Cooling with Non-Intrusive Sensing Techniques. Thesis for the Degree of Doctor of Philosophy, Carnegie Mellon University, USA.

Nassif, N. (2012). A Robust CO_2-based Demand-Controlled Ventilation Control Strategy for Multi-Zone HVAC Systems. *Energy and Buildings*, Vol. 45, pp. 72-81.

Ortega, J., Han, L., Whittacker, N., & Bowring, N. (2015). A Machine Learning based Approach to Model User Occupancy and Activity Patterns for Energy Saving in Buildings. *Science and Information Conference*, pp. 474–482.

Raykov, Y., Ozer, E., Dasika, G., Boukouvalas, A., & Little, M. (2016). Predicting Room Occupancy with a Single Passive Infrared (PIR) Sensor through Behavior Extraction. *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 1016-1027.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2015). You Only Look Once: Unified, Real-Time Object Detection. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779-788.

Redmon, J. & Farhadi, A. (2018). Yolov3: An Incremental Improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *International Conference on Neural Information Processing Systems*, pp. 91-99.

Ruder, S. (2017). An Overview of Gradient Descent Optimization Algorithms. Available at https://arxiv.org /pdf/1609.04747.pdf

Stein, G., Mano, O., & Shashua, A. (2003). Vision-based ACC with a Single Camera: Bounds on Range and Range Rate Accuracy. *IEEE Intelligent Vehicles Symposium*, pp. 120-125.

Sun, Z., Wang, S., & Ma, Z. (2011). In-Situ Implementation and Validation of a CO_2-based Adaptive Demand-Controlled Ventilation Strategy in Multi-Zone Office Building. *Building and Environment*, Vol. 46, pp. 124-133.

Sze, V., Chen, Y., Yang, Y., & Emer, J. (2017). Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *Proceedings of the IEEE*, Vol. 105, No. 12, pp. 2295-2329.

Wang, L., Shi, J., Song, G., & Shen, I. (2007). Object Detection Combining Recognition and Segmentation. 8th Asian Conference on Computer Vision, pp. 189-199.

Wu, J., Cai, N., Chen, W., Wang, H., & Wang. G. (2019). Automatic Detection of Hardhats Worn by Construction Personnel: A Deep Learning Approach and Benchmark Dataset. *Automation in Construction*, Vol. 106, 102894, 2019.

Yang, Z., Li, N., &Gerber, B. (2012). A Non-Intrusive Occupancy Monitoring System for Demand Driven HVAC Operations. *Construction Research Congress*, pp. 828-837.

Zheng, J., Lu, C., Hao, C., Chen, D., & Guo, D. (2020). Improving the Generalization Ability of Deep Neural Networks for Cross-Domain Visual Recognition. *IEEE Transactions on Cognitive and Developmental Systems*, Early Access.