IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 5, October, 2024, pp. 2711-2719
Journal homepage: http://iieta.org/journals/ts

An End-to-End Object Detection System in Indoor Environments Using Lightweight Neural Network

Mouna Afif[1], Yahia Said[2*], Riadh Ayachi[3], Manel Hleili[4]

[1] ISITCom, University of Sousse, Sousse 4011, Tunisia
[2] Department of Electrical Engineering, College of Engineering, Northern Border University, Arar 91431, Saudi Arabia
[3] Laboratory of Electronics and Microelectronics (LR99ES30), Faculty of Sciences of Monastir, University of Monastir, Monastir 5019, Tunisia
[4] Department of Mathematics, Faculty of Science, University of Tabuk, Tabuk 71491, Saudi Arabia

Corresponding Author Email: Yahia.said@nbu.edu.sa

https://doi.org/10.18280/ts.410544

Received: 19 April 2024
Revised: 18 August 2024
Accepted: 28 August 2024
Available online: 31 October 2024

Keywords:
indoor object detection, indoor navigation, visually impaired people (VIP), deep convolutional neural network (DCNN), deep learning

ABSTRACT

Vision plays a pivotal role in how humans interact with and understand their surroundings. Despite significant advancements in assistive technology, the specific challenges of indoor navigation for the visually impaired and blind (VIP) community remain inadequately addressed. Recognizing and locating objects within indoor environments is crucial for enhancing the independence and safety of VIP individuals. This paper introduces an innovative deep-learning framework designed for superior object detection in indoor settings, alongside a comprehensive new dataset tailored for this purpose. At the heart of the proposed system is an optimized version of the YOLOv3 architecture, re-engineered to significantly reduce computational demands while maintaining high accuracy. The newly developed indoor dataset encompasses over 11,000 images featuring 25 essential object categories, curated to represent a variety of lighting conditions and complex indoor scenarios that pose navigational hazards. This dataset not only challenges the model with real-world complexities but also facilitates the training of more robust, efficient neural networks. Experimental results underscore the proposed model's exceptional performance, demonstrating marked improvements in detection precision and system efficiency, thereby offering a promising new direction for assistive technologies in indoor navigation. The obtained results of the proposed indoor objects detection system are highly promising, showcasing a mean average precision (mAP) of 89.78% across all 25 object classes. This impressive performance indicates the system's robust capability in accurately detecting a diverse range of objects commonly found in indoor environments.

1. INTRODUCTION

Blindness and other visual impairments constitute a significant challenge that is experienced by a large number of people all across the nation and the world. There are approximately 1.3 billion people who are affected by vision impairments, of which 36 million are completely blind, according to estimates provided by the World Health Organization (WHO) [1] in the year 2018.

A challenge that is significant in the fields of computer vision and artificial intelligence is the identification and recognition of objects that are found within buildings. A human being is more dependent on their sight sense than any of their other senses in order to move around and investigate their surroundings. This is because the sight sense offers exact information about the size, texture, form, color, and distance to the item. The construction of new indoor object identification systems is attracting a growing amount of interest for the time being. Individuals who are blind or visually impaired (VIP) require a set of talents in order to travel and engage with the world that surround them.

One may argue that vision is the most important sense that humans possess in order to move around and interact with the environment in a secure manner. The problem of object detection is a particularly essential one since it may be used to a broad variety of situations. One of the most difficult jobs in computer vision is object identification and recognition in real-world scenes. This is because it involves dealing with a wide range of challenges, including variations in lighting and occlusion, changes in viewpoint, and more complicated backgrounds, among other things.

The elderly population all around the world faces the challenge of living with a vision impairment on a seemingly daily basis. Individuals who have high visual abilities have an easier time walking and exploring in private, locating items inside the building, and avoiding obstructions and challenges. This job, on the other hand, poses a particularly difficult challenge for those who are blind or have other impairments. This task becomes more difficult to do, particularly in congested indoor scenes, because these kinds of surroundings provide a high level of occlusion and variable lighting conditions.

In order to get a comprehensive understanding of the input image, we need not only focus on categorizing the things that are presented in the photographs, but we should also make an effort to find objects that are located inside the image. This particular work is a part of the object detection task [2], which has the potential to make contributions to a wide variety of other domains, including face identification [3], pedestrian detection [4], skeleton detection [5], road sign detection [6], indoor item detection [7-9], and indoor scene recognition [10]. As a result of the fact that it offers blind and visually impaired individuals extremely helpful information about their surroundings, the topic of computer vision that deals with indoor object recognition is a particularly attractive one. In the context of human-robot interaction, augmented reality, and root manipulation activities, object detection and recognition constitute an essential component. On the other hand, occlusion and deformable indoor objects continue to make the problem an even more difficult one to solve in inside congested situations. Furthermore, depending on the angle and calibration of the camera, the same interior item may look in a variety of various shapes, textures, and sizes. In light of this, the development of an accurate indoor object identification system has the potential to assist blind and sighted individuals in avoiding obstacles and risks and improving their ability to navigate inside environments efficiently.

Object identification algorithms have reached a level of maturity in today's applications that allows them to address a wide variety of real-world computer vision problems. The issue of object detection is a well-known topic that is recognized to be particularly difficult since it requires a huge number of labeled datasets in order to train models and to generalize the performances of these detection models.

Utilizing deep learning strategies for the purpose of detecting objects inside of buildings is a well-known and effective approach to solving this challenge. However, in order to get satisfactory outcomes through the use of algorithms that are based on deep learning, it is essential to train these models by making use of an enormous quantity of data. The performance of object identification tasks in a considerable variety of domains has been significantly improved by deep learning models in recent times. This improvement is particularly noticeable for interior navigating aids for vision impaired individuals and for bling.

Object localization and object classification are two aspects of the issue of indoor object detection that we address in this study. Object localization involves locating objects in the input photos, while object classification involves identifying the type of the item. In this work, we offer a novel indoor object detection system that can efficiently handle the detection of particular interior items that are essential for everyday navigation. This system will accommodate the varied demands of those who are blind and those who are sighted. Throughout the course of this study, an attempt is made to gather and annotate eleven thousand interior photographs that contain twenty-four indoor landmark icons. During the training phase, we ensured that a variety of tough situations were used, which allowed us to demonstrate our strong detection method. Additionally, it is important to highlight that the suggested indoor object identification system was constructed by us on the basis of a modified version of YOLOv3, which employs MobileNet v1 as the backbone of the features extractor algorithm. Using MobileNet V1 as the network backbone for YOLOv3 rather than Darknet-53 has substantial benefits, particularly in terms of efficiency and speed. MobileNet V1, built for mobile and embedded applications, uses depthwise separable convolutions, which significantly lower the amount of parameters and computational load when compared to the classic convolutions used in Darknet-53. This makes MobileNet V1 significantly more suitable for real-time object identification on resource-constrained devices like smartphones and embedded systems, while maintaining high accuracy. The lower power consumption and faster inference times achieved by MobileNet V1 increase the feasibility of deploying YOLOv3 in practical, real-world applications requiring timely and reliable detection, such as assisting visually impaired individuals with indoor navigation and object recognition.

Existing indoor object identification systems usually face substantial hurdles, such as poor performance in changing illumination conditions and difficulty distinguishing closely spaced or overlapping objects. These constraints are frequently caused by standard machine learning approaches that struggle to generalize across varied interior contexts, yielding inconsistent results. Furthermore, many current systems lack the ability to recognize a broad range of object classes, limiting their practical utility. The suggested approach tackles these concerns by employing a Feature Pyramid Network (FPN)-style structure. This method improves multi-scale feature representation, allowing the system to effectively detect objects of varying sizes and resolutions. The FPN-like structure enables consistent performance in a variety of illumination settings and enhances the system's capacity to distinguish between closely positioned or overlapping objects.

What follows is a summary of the remaining sections of this paper: Section 2 presents related works about indoor object detection systems. Section 3 introduces the dataset that was suggested for usage. The suggested architecture for indoor object detection is described in section 4. Experiments and findings are described in Section 5, and the study is concluded in Section 6.

2. RELATED WORKS

In the realms of computer vision and artificial intelligence, indoor item detection and identification constitute a formidable axis. Indoor spaces, with their crowded décor and severe lighting conditions, make this work even more complex. Building a novel deep learning-based indoor object identification system is the primary goal of this effort. We need to make sure that our planned work can withstand a lot of different kinds of extremes, such complete darkness, different lighting, different decorations, and fluctuation both within and across classes.

A lot of research has gone into finding solutions to the challenges of indoor object detection and recognition. Using machine learning approaches to solve this challenge is especially important for classical works [11, 12]. To get the most out of the network inputs, though, algorithms of this kind need intricate pipeline architecture. In order to provide a thorough understanding of the interior geometry, some efforts depend on creating statically models [13, 14]. Many works have been suggested based on RGB-D sensors since their emergence, such as Kinect cameras, which offer depth information in addition to color information about their surroundings. Indoor robot navigation makes extensive use of these cameras [15]. An indoor objects detection for indoor robots navigation-based method was proposed in the study by

Jia et al. [16].

The area of indoor navigation has attracted a large number of academic scholars. Researchers employ SLAM method for localization approaches since it is a powerful component [17].

Researchers have tackled the difficult task of developing reliable indoor object recognition systems for both visually impaired and sighted people's navigation needs. Because of factors such as elaborate décor, varying lighting conditions, and significant occlusion, object detection problems become more problematic in interior contexts.

The challenge of indoor object detection lies not only in properly localizing items but also in categorizing them from input photos or videos. This is a particularly difficult topic for us to solve since our work is tailored to a certain group of people: those who are blind, partially blind, or seeing. To aid these people in exploring their interior environments, we need to create an application that can recognize objects extremely well.

The navigation of mobile robots relies heavily on the development of a reliable object detecting system. The authors [18] provide a multi-model place classification system that mobile robots may employ to correctly detect interior locations and determine semantic class categories.

Deep learning algorithms have been the focus of a lot of attention from computer vision experts as of late. Many proposals utilizing deep learning and DCNN architectures have been made in light of this reality. The authors [19] put forward a convolutional neural network (CNN)–based system for indoor item identification. The Fragments of Videos (Fov) dataset and the public indoor dataset were utilized to train the model. To aid the visually handicapped, Bhandari et al. presented an object identification and detection system in study by Bhandari et al. [20]. Deep learning methods were utilized in the development of this application.

One of the most difficult and complex issues in the world of artificial intelligence is object detection and recognition. Indoor mobile robot navigation finds this job highly handy. The authors [21] suggest a method for classifying three-dimensional objects using a hybrid of convolutional and recursive neural networks. Learned features and 3D RGB-D picture classification are the system's strong suits.

There are two main types of DCNNs used for object detection tasks: one-stage and two-stage. Two-stage detectors, like Faster-RCNN [22] and Mask-RCNN [23], have two stages. In the first stage, they use region proposal networks (RPN) to extract ROIs from input images. Then, in the second stage, they classify objects and locate their positions using bounding box regression. Results from object detection exercises showed that this sort of detector performed very well and efficiently. Having said that, two-stage detectors are tedious and resource-intensive.

The one-stage object detectors, such as SSD [24], YOLO family [25-27], and RetinaNet [28], handle the detection issue as a simple regression problem and only conduct detection on one stage.

Comparing the suggested indoor object detection technology to current best practices yields promising results. In addition to the proposed detection method, we include a new indoor object dataset in this study for training and testing purposes. This dataset includes new indoor landmark items that are highly recommended for navigation by both sighted and blind people, and it also includes a variety of demanding settings. We address hazardous circumstances so that people who are blind or have other mobility impairments may navigate more safely as part of our job.

Various works have been proposed in the literature that address the problem of indoor objects detection but almost none of them can detect a set of 25 indoor objects highly valuable for blinds and visually impaired mobility.

The proposed modification of the YOLOv3 architecture by replacing the Darknet-53 backbone with MobileNet V1 represents a significant advancement in the field of object detection. Traditional state-of-the-art methods, such as those using the original YOLOv3 with Darknet-53, excel in accuracy but often fall short in terms of computational efficiency and real-time applicability, particularly on resource-constrained devices. In contrast, MobileNet V1, designed for lightweight applications, reduces the computational burden through depthwise separable convolutions while maintaining competitive accuracy. This modification results in a faster and more efficient object detection system that is well-suited for real-time applications, including those assisting visually impaired users in navigating indoor environments.

3. DATASET COLLECTION

Accurate and robust item detection and identification systems rely on data collecting and labeling. In addition, there is a labor-intensive challenge in annotating a large amount of data.

Heavy occlusions, variable lighting conditions, complicated backgrounds, multiple points of view, etc. are just a few of the hard aspects that the suggested dataset takes into consideration in order to create extremely robust detection algorithms, in contrast to the existent indoor datasets. We also want to mention that the suggested dataset has a lot of variances both between classes and within them, which is great for developing new accurate detectors. One example of inter-class variance is shown in Figure 1.

The most novel aspect of the proposed dataset is its treatment of previously unrecognized landmark indoor objects and extremely risky scenarios; these are addressed in an effort to develop safer detection systems that can assist the visually impaired and the blind with indoor navigation.

The various items and decorations that make up indoor scenery (such as doors, signs, stairs, elevators, hallways, etc.) make it very different from outside landscape.

In the accompanying picture, we can see that the suggested indoor dataset records a large amount of intra-class variance among the doors. A wide variety of door styles, sizes, colors, and textures are available. Additionally, the doors included in the proposed project come in a variety of materials, including wood, iron, and glass. Additionally, doors are photographed from various angles and with various attitudes, with some doors open and others closed. Our suggested indoor object dataset is well-suited for training and testing new indoor object detectors due to all these questions.

Figure 1. Inter-class variation

What makes the suggested indoor dataset unique is its focus on:

- Invariance with Light Intensity: this foundation offers a wider range of photos captured in varying lighting conditions. The items in this base are captured from diverse angles, thus they remain invariant to strict geometric change.
- The occlusion effect, in which some or all of an object's surface is obscured.
- The accessibility of necessary items for individuals with vision impairments and other disabilities. To make sure the visually impaired can safely go downstairs, we've highlighted potentially hazardous circumstances.
- The ability to make decisions is given to the sight challenged by seeing visuals of crossings and corral intersections.
- Problems with safety caused by the presence of impediments in the hallways.
- 11000 photos taken at various spots around the structure in August 2019. There are twenty-five different types of internal items. In order to create new indoor object identification systems, convolutional neural networks will be trained using this database.
- A With a resolution of 4032×3024 pixels, the suggested database displays pictures.

Two additional facts are considered in this database, which gives it its uniqueness: first, the spatial relations between the items in the scene, and second, the various actions that may be done to these things, such as the relationships between objects and VIPs.

Gathering data in a variety of lighting circumstances and against a complicated image background, this collection of acquired photographs includes several characteristics and strengths, guaranteeing higher resilience when recognizing these items using a detection system.

The significant level of inter-class variety provided by our picture library is one of its strengths. For example, standard doors and elevator doors are examples of extremely similar classes. Figure 2 below illustrates the intra-class variance with an example.

When visually impaired people are walking within buildings, there are 25 crucial classes that must be identified. These classes include fire extinguishers, doors, stairs (both up and down), signs, windows, and more. All of the item types stored in our database are summarized in Table 1.

Figure 2. Intra-class variation example

Table 1. Classes names

Light Switch	
fire extinguisher	trash bin
door	printer
person	heater
elevator	microwave
showcase	Plant
exhibition table	window
exit	disabled exit
chair	water dispenser
table	drink dispenser
elevator	wc
stairs	security button
confidence zone	podotactile tape

The suggested indoor dataset is made up of a variety of 11,000 photos taken within buildings, each comprising 25 landmark items that are vital for visually impaired and blind people to navigate their way around indoors on a daily basis.

This dataset is ideal for building effective indoor object recognition systems, since it contains fully-labeled data that can be used to train and evaluate deep learning models. Both visually impaired and sighted people are expected to benefit greatly from this resource.

In order to facilitate their movement within enclosed spaces such as medical offices, classrooms, libraries, and hospitals, among others.

Using the acquired information, a brand-new indoor object recognition system will be tested to see how well it helps both visually impaired and sighted people navigate. With its multi-level object rightness, the suggested dataset is a step in the correct direction toward new applications that can aid a large group of people who are blind, partially blind, or sighted.

4. PROPOSED APPROACH FOR INDOOR OBJECT DETECTION

Accessing new locations presents a number of challenges for the visually impaired as they go about their everyday lives. One of the main areas of ongoing study in computer vision and AI is the detection of interior objects from input videos and photos. New deep learning approaches successfully handle the demanding problem of indoor object identification and recognition, which is one of the most important areas of computer vision.

For the sight impaired, navigating inside spaces, particularly unfamiliar ones, may be a daunting endeavor, especially without assistance. In order to tackle this problem, several different approaches have been suggested.

Indoor navigation is a delicate and difficult topic for the visually handicapped. In this paper, we provide a novel method for detecting objects within buildings using robust deep learning models to guarantee them a safer and more autonomous navigating experience. Many things and barriers might pose a threat to the safety of a visually impaired person when they navigate interior spaces. A visually impaired person may face several perilous scenarios, including the inadequacy of a stairway, when exploring unfamiliar interior spaces. We suggest building a model of a visual navigational assistance for the visually handicapped using deep learning techniques, namely convolutional neural networks (CNN), to address these risky situations in this course.

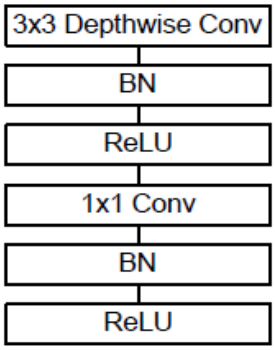

Figure 3. Depthwise separable convolution block

In the field of artificial intelligence, deep learning techniques have become game-changers. picture and video processing, object detection, natural language comprehension, speech recognition, and picture classification are just a few of the computer visions jobs that heavily utilize it.

Our suggested architecture is an indoor object detection system built on top of a tweaked YOLOv3 [27]. We see that Darknet serves as the foundation for feature extraction in the original YOLOv3 versions [27]. An end-to-end DCNN forms the basis of the suggested indoor object detection system. Feature extraction and detection are the two primary components of the proposed study. We employed the MobileNet v1 architecture [29] as a foundation for feature extraction and to guarantee a mobile implementation of the suggested method. Due to its smaller size in comparison to other DCNN models, the MobileNet [29] architecture is well-suited for use in mobile applications. Depthwise separable convolution is a new feature in MobileNet architecture. Every color channel undergoes a singular convolution. Obtaining a small, lightweight model is greatly facilitated by using these convolution types. An iterative process of 3×3 depthwise convolution, batch normalizing, and rectified linear unit (RELU) is comprised of the depthwise separable convolution. This is followed by 1×1 pointwise convolution, batch normalization, and RELU. A depth-wise separable convolution block is shown in Figure 3.

The number of parameters in the model is drastically reduced using depthwise separable convolution. As opposed to the 1x1 convolution used by pointwise convolution, each input channel in depthwise convolution is subjected to a separate filter. Combining the values of the input channels is the fundamental distinction between normal convolution and depthwise convolution. As an example, a one-channel output feature map is the consequence of applying a depthwise convolution across the features map to an input picture with three channels. Following a depthwise convolution in MobileNet architecture, a custom activation layer called RELU 6 is used to mimic an activation value of 6. The following equation is used to calculate the activation of RELU 6.

$$y=\min(\max(0,x),6)$$

If x is an input, then y is the result.

Using a 1×1 kernel size for pointwise convolution is the second new feature introduced by the MobileNet design. The goal of this convolutional layer is to generate new features by merging the output feature map. We get the depthwise separable convolution block by merging the depthwise and pointwise convolutions. Figure 4 shows the distinctions between the three convolutional layers utilized by MobileNet.

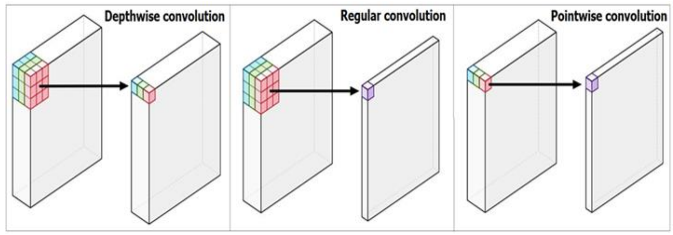

Figure 4. Different types of convolutions used in MobileNet architecture

When compared to normal convolution, the depthwise separable convolution block is both quicker and uses less computing resources while still performing the same purpose. The initial hidden layer of a MobileNet architecture, which is immediately after the input picture, uses a conventional convolution. At the end, it includes an average pooling layer and a series of depthwise convolution locks.

Using an end-to-end DCNN architecture, we introduce a novel framework for indoor item detection and recognition. As a DCNN, we employed an adaptation of the YOLOv3 architecture [27].

To address object detection challenges, YOLOv3 introduces a one-stage deep convolutional neural network. There are two primary components to the YOLOv3 architecture:

1) A component for feature extraction, which is employed to extract crucial characteristics from input photos.

2) Detection component: the section that uses bounding boxes to forecast class categories and their positions in the input pictures.

One advantage of YOLOv3 over competing models is that it can forecast both the position and class of an item at the same time. It creates bounding boxes with dimensional clusters as the "anchors" so it can make predictions about what items are in the input photos. The hole identification problem is thus treated as a regression problem by YOLOv3. The input picture is partitioned into cells that are S×S grids. In YOLOv3, each cell is limited to testing a specific number of anchors and making one object prediction. The YOLOv3 algorithm can forecast the bounding box of each grid cell. In which the objectness (confidence) of each bounding box is linked to it. Whether or whether the bounding box contains an item is determined by this word. With its YOLOv3 design, you may additionally anticipate C class probabilities, or one for every possible class. Each category's probability can take on a value between zero and one. In addition, while making a forecast, the total of all the probabilities for the C classes is 1. A 1×1 convolution layer (an independent logistic classifier) is used to predict the localization of each item, including its bounding box and class probability. The YOLOv3 architecture produces an output shape that looks like this: 1×1(B×(4+1+C)), where 1×1 is the convolution layer, B is the number of bounding boxes that can be detected by each grid cell, '4' is the coordinates of the bounding boxes (tx, ty, tw, th), 1 is the object score for each grid cell, and C is the number of classes. Three-dimensional item sizes (small, medium, and large) are anticipated by the YOLOv3 design. Our proposed method made use of a total of 25 class numbers and 2 boxes each scale, for a total of 6 boxes. To that end, we suggest an output shape of 1×1(2×(4+1+25)). The objectness score is assigned by the YOLOv3 architecture for every bounding box that is extracted. The objectness score is a measure of the likelihood that an object is included inside the grid. Objectness score in

YOLOv3 architecture is computed not with a softmax layer but with an independent logistic classifier. Decrease the computational complexity of the detection procedure significantly by utilizing the independent logistic classifier. Indoor object detection's full workflow is shown in Figure 5.

When compared to previous models of neural networks, YOLOv3 demonstrates superior capability in identifying objects of many sizes. Figure 6 shows that YOLOv3 uses an FPN structural adaptation to guarantee detection on small, medium, and large sizes. A bottom-moving FPN-like structure down-samples the picture by 2. It improves accuracy as well. In order to upsample the picture, a structure similar to an FPN also uses the top-down movement. Because of this change, the localization precision is improved.

The following stages are used to achieve the detection at three scales:

- First step: (large object detection) Making an object presence prediction using the feature extraction backbone's final feature map.

- The second step, "medium object detection," entails combining two feature maps that display things of a similar size and then using a convolution to forecast the existence of objects of a medium size.

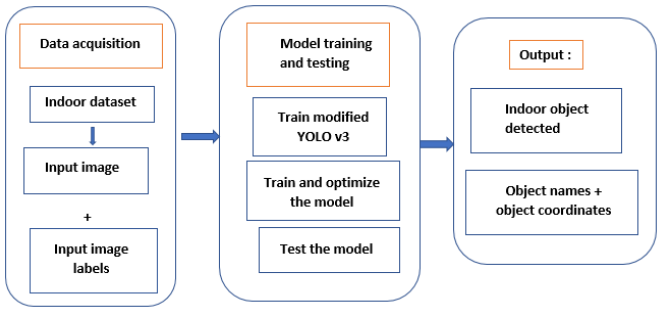

Figure 5. Overall pipeline of the indoor object detection method

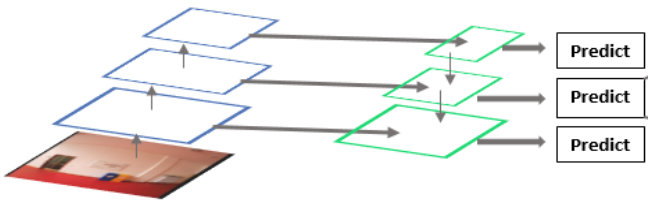

Figure 6. FPN-like structure used in YOLOv3 architecture

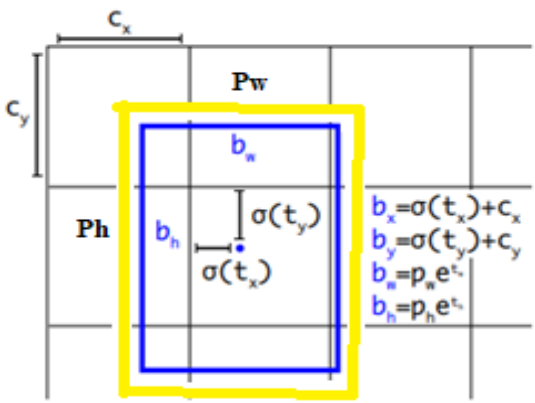

Figure 7. Bounding box technique used in YOLOv3

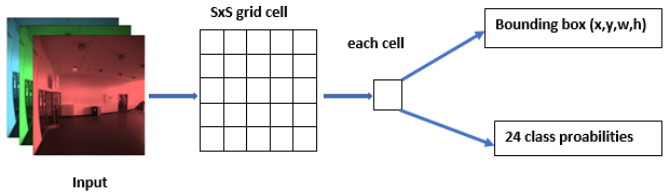

Figure 8. Output shape of the proposed indoor object detection system

- Third Step: Detecting Small Objects.

Taking the feature map from the convolution layer in step 2 and upsampling it by 2; then, merging the two feature maps, one acquired via a bottom-up approach and the other from a top-down one. Predicting the whereabouts of tiny items by using a convolution on the generated feature map.

The coordinates (tx, ty, tw, th) of each boundary box are given by itself. The center's coordinates, tx and ty, make up the bounding box (bbox). The width offset from the bbox center is denoted by tw, while the height offset is th. Given that cx and cy represent the coordinates of a grid cell in the top-left corner of the feature map, the projected parameters at the end are bx, by, bw, and bh. The following equations are used to generate these parameters:

$$b_x=\sigma(t_x)+c_x$$
$$b_y=\sigma(t_y)+c_y$$
$$b_w=p_w e^{t}_{w}$$
$$b_h=p_h e^{t}_{h}$$

where,

- σ is the sigmoid function $\sigma(x)=1/(1+e^{-x})$.
- the top-left corner of the bbox has an anchor coordinates p_w and p_h.

Figure 7 provides how the bounding box coordinates can be calculated according to the YOLOv3 architecture.

Take the following input image: ((52×52)+(26×26)+(13×13))×2=6838 bboxes. This is an example of how the suggested detection method predicts boxes at three scales. To lessen it, the objectness ratings are used to filter the bboxes using a predefined threshold. As a second point, non-maximum statistics are delayed (NMS) when compared to the ground truth. Figure 8 present the output shape of the proposed system.

We consider these factors in order to design a strong and efficient indoor object detecting system:

(1) Different kinds of illumination.
(2) There is a lot of variance both inside and between classes.
(3) Stay away from dangerous circumstances.
(4) Look for new signs indoors.
(5) Be careful with occlusion, forms, and textures.
(6) An alternative vantage point of the interior sign.

Both visually impaired and sighted people rely on indoor items as markers while navigating and finding their way around within buildings. Consequently, a crucial part of navigation assistance is accurate and fast interior item identification.

5. EXPERIMENTS AND RESULTS

A combination of high detection accuracies and real-time processing is necessary to construct a reliable indoor object

detection system. We present a novel indoor object identification system that is based on deep learning *. We're going above and beyond to create an indoor object detection system that both visually impaired and sighted people can use to better explore their immediate environs and take part in everyday activities. Researchers ran tests on the new indoor object dataset, both for training and for testing purposes. We highlight that the suggested dataset is ideal for interior navigation aid, as it comprises 11,000 photos taken indoors and include 25 indoor landmark items. Also, previous datasets didn't take the additional interior landmark items that the suggested dataset offers into account. There are a number of challenging situations presented by the suggested dataset, including occlusion, considerable intra-and inter-class variance, varying lighting conditions, many perspectives, and so on.

We split the suggested dataset in half, using half for training and half for testing the network, so that we could carry out the experiments as planned. In addition to using ADAM [30] to tune the network, we trained with 7300 photos and tested with 3700 images.

We used a batch normalization size of 16 and trained the DCNN models for 30 epochs, each of which included 476 iterations. The suggested dataset contains photos with a resolution of 4032 by 3024 pixels. The suggested indoor object identification program uses a 224×224 picture resizing during training. The suggested trials are executed on an HP desktop computer that has a 12-gigabyte graphics processing unit, an Intel Xeon E5-2683 v4 central processing unit, and a Tesla K40C graphics card. The Keras framework, which provides a high-level interface for programming neural networks, was used to implement the whole work. Mean average precision (mAP) was the metric used to evaluate the suggested indoor object identification system. The code for the system was developed using the following tools: python 3.6 setup, TensorFlow [31] 1.13, NVIDIA CUDA toolkit 10.0, and the deep learning library CUDNN 7.0.

This study tackles the issue of blind and disabled people navigating interior spaces securely, avoiding hazards and impediments on the way to their goals. We put the suggested detection method through its paces using our multi-object dataset. We suggested this dataset to address extremely difficult and risky circumstances that can be avoided by both visually impaired and sighted people when navigating. Our findings lead us to conclude that the suggested interior object detection system significantly improves the safety of indoor navigation for visually impaired and those without sight.

Table 2 indicated that promising results were obtained for every category of suggested interior goods. Notably, we address novel indoor classes that have not been investigated using state-of-the-art methods, but which are highly regarded for interior navigation support for the visually impaired and the blind. Our detection precisions were good across the board for indoor class categories. The overall mean average accuracy (mAP) for the test dataset was 89.78%, which is worth noting.

In order to study the efficiency of the proposed system results and performances, various evaluation metrics have been adopted in this study. The following equations details the different evaluation metrics used.

$$Accuracy = \mathrm{TP} + \mathrm{TN} / \mathrm{FP} + \mathrm{FN} + \mathrm{TP} + \mathrm{TN} \quad (1)$$

$$\mathrm{Precision} = \mathrm{TP} / \mathrm{FP} + \mathrm{TP} \quad (2)$$

$$Recall = \mathrm{TP} / \mathrm{FN} + \mathrm{TP} \quad (3)$$

Both visually impaired and sighted people will find the proposed study particularly useful in expanding their exploration of their interior environments. Both visually impaired and sighted people can use our proposed work to identify items within buildings at a rate of up to 59 frames per second (FPS).

Table 2. Per-class detection obtained results

Class	Window	Door	Person	Light Switch	Showcase	Exhibition Table	Exit	Stairs	Chair
AP (%)	87.52	89.96	91.56	84.67	88.63	89.31	88.96	90.16	93.65
Class	table	Confidence zone	Security button	Podotactile tape	Trash bin	Water dispenser	Heater	sofa	
AP (%)	86.32	89.63	87.59	86.54	87.68	93.65	94.36	89.21	
Class	microwave	plant	Fire extinguisher	Printer	Disabled exit	wc	Elevator	Drink dispenser	
AP (%)	88.34	89.96	91.38	92.36	89.36	89.65	89.69	94.38	

Table 3. Per-class detection accuracies compared to our previous work [7]

Class Name	Window	Elevator	Door	Trash Bin	Stairs	Security Button	Table	Heater	Chair	Light Switch
Previous work [7]	59.83	77.58	96.94	80.83	69.63	60.47	81.17	76.63	77.73	44.97
Proposed	87.52	77.58	89.96	87.68	90.16	87.59	86.32	94.36	93.65	84.67

Table 4. Evaluation metrics results comparison between the baseline version and the proposed improved version of YOLOv3

Method	Accuracy	Precision	Recall
Baseline	85.23	84.45	82.96
Proposed	89.78	90.32	87.97

Table 5. Per-class detection results comparison with results obtained in the previous study [19]

Indoor Object Name	Method [19] (Indoor AP %)	Ours (Proposed Dataset AP %)
door	42.9	89.96
chair	72.6	93.65
table	46.2	86.32

The findings from our prior work [7] according to Table 3 and the results from the previous study [19] are compared with the per-class detection accuracies in Tables 3 and 4. Mean average accuracy (mAP) was the statistic we settled on for evaluation. As a measure of detection accuracy, average precision (AP) shows how well each class scored. Table 4 provides a comparison between the obtained results of the baseline architecture of YOLOv3 and the proposed improved version.

As presented in Table 4, the proposed improved version demonstrates better evaluation metrics performances compared to the baseline version of YOLOv3 network. This result empowers the use of MobileNet v1 as a network backbone instead of using darknet 53.

Importantly, the proposed indoor object identification system was trained on tough pictures and novel classes of things that are important for both blind and sighted people's indoor navigation, but which were not included in the state-of-the-art datasets. Despite this, we achieved quite encouraging detection results. In comparison to other studies, our suggested approach achieves better identification accuracies across the board for most class items. Table 5 presents a comparison between the obtained results for 3 main indoor classes and the state-of-the-art works.

6. CONCLUSION

The proposed system utilizes deep learning techniques to identify objects inside a building with multiple labels. While navigating indoor spaces, the suggested indoor object detection system alerts both visually impaired and sighted users to a variety of potential hazards. An updated indoor object dataset with 11,000 photos of 24 indoor landmark items was generated for the purpose of training and testing the suggested work. A robust and accurate detector must be built to handle the several demanding situations presented by the images in the dataset. Visually impaired and visually impaired people alike will find a wealth of useful information in the specified dataset. According to the outcomes, our suggested approach for detecting objects within buildings is quite efficient and accurate. In order to enhance the quality of life for those who are blind, visually impaired, or partly impaired, the proposed study employs deep learning techniques to help them avoid hazards and get a full understanding of items and their surroundings. The performance of the object detection system may be influenced by the bias present in the dataset used for training. Addressing dataset bias and ensuring diverse representation of indoor environments and objects is essential for improving the robustness and generalization capabilities of the model. As future works, it was suggested to develop mechanisms for the system to adapt and learn from user feedback over time can improve its performance and usability. However, a critical analysis of the results reveals some limitations. The dataset, while extensive with 11,000 photos of 24 indoor landmark items, may harbor biases that could impact the model's performance across diverse environments. This bias could lead to reduced robustness and generalization capabilities, especially in varying indoor settings. Additionally, the static nature of the current model does not account for dynamic changes in the environment or user-specific variations. To address these limitations, future work should focus on expanding the dataset to include a broader range of objects and environments, thus minimizing bias.

ACKNOWLEDGMENT

This research was funded by the Deanship of Scientific Research at Northern Border University, Arar, KSA through the project number “NBU-FFR-2024-3030-05”.

REFERENCES

[1] World Health Organization. Vision Impairment and Blindness. https://www.who.int/news-room/fact-sheets/blindness-and-visual-impairment.

[2] Zhao, Z.Q., Zheng, P., Xu, S.T., Wu, X. (2019). Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems, 30(11): 3212-3232. https://doi.org/10.1109/tnnls.2018.2876865

[3] Kumar, A., Kaur, A., Kumar, M. (2018). Face detection techniques: A review. Artificial Intelligence Review, 52(2): 927-948. https://doi.org/10.1007/s10462-018-9650-2

[4] Chen, Y., Guo, S., Zhang, B., Du, K.L. (2013). A pedestrian detection and tracking system based on video processing technology. In 2013 Fourth Global Congress on Intelligent Systems, Hong Kong, China, pp. 69-73. https://doi.org/10.1109/GCIS.2013.17

[5] Wu, J., Li, Y., Wang, L., Wang, K., Li, R., Zhou, T. (2019). Skeleton based temporal action detection with yolo. In Journal of Physics: Conference Series, IOP Publishing, 1237(2): 022087. https://doi.org/10.1088/1742-6596/1237/2/022087

[6] Ayachi, R., Afif, M., Said, Y., Atri, M. (2019). Traffic signs detection for real-world application of an advanced driving assisting system using deep learning. Neural Processing Letters, 51(1): 837-851. https://doi.org/10.1007/s11063-019-10115-8

[7] Afif, M., Ayachi, R., Said, Y., Pissaloux, E., Atri, M. (2020). An evaluation of retinanet on indoor object detection for blind and visually impaired persons assistance navigation. Neural Processing Letters, 51: 2265-2279. https://doi.org/10.1007/s11063-020-10197-9

[8] Aftf, M., Ayachi, R., Said, Y., Pissaloux, E., Atri, M. (2019). Indoor object c1assification for autonomous navigation assistance based on deep CNN model. In 2019 IEEE International Symposium on Measurements & Networking (M&N), Catania, Italy, pp. 1-4. https://doi.org/10.1109/IWMN.2019.8805042

[9] Talla-Chumpitaz, R., Castillo-Cara, M., Orozco-Barbosa, L., García-Castro, R. (2023). A novel deep learning approach using blurring image techniques for Bluetooth-based indoor localisation. Information Fusion, 91: 173-186. https://doi.org/10.1016/j.inffus.2022.10.011

[10] Liu, Y., Jiang, D., Xu, C., Sun, Y., Jiang, G., Tao, B., Tong, X., Xu, M., Li, G., Yun, J. (2022). Deep learning based 3D target detection for indoor scenes. Applied Intelligence, 53(9): 10218-10231. https://doi.org/10.1007/s10489-022-03888-4

[11] Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W. (2013). Selective search for object recognition. International Journal of Computer Vision, 104: 154-171. https://doi.org/10.1007/s11263-013-0620-5

[12] Gill, J.S., Brar, A.S. (2019). Support vector based indoor scene classification technique using different features. In

2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, pp. 685-689. https://doi.org/10.1109/iceca.2019.8822153

[13] Wang, H., Gould, S.J., Roller, D. (2013). Discriminative learning with latent variables for cluttered indoor scene understanding. Communications of the ACM, 56(4): 92-99. https://doi.org/10.1145/2436256.2436276

[14] Del Pero, L., Bowdish, J., Fried, D., Kermgard, B., Hartley, E., Barnard, K. (2012). Bayesian geometric modeling of indoor scenes. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, pp. 2719-2726. https://doi.org/10.1109/cvpr.2012.6247994.

[15] Husain, F., Schulz, H., Dellen, B., Torras, C., Behnke, S. (2017). Combining semantic and geometric features for object class segmentation of indoor scenes. IEEE Robotics and Automation Letters, 2(1): 49-55. https://doi.org/10.1109/lra.2016.2532927

[16] Jia, Y., Ramalingam, B., Mohan, R.E., Yang, Z., Zeng, Z., Veerajagadheswar, P. (2023). Deep-Learning-based context-aware multi-level information fusion systems for indoor mobile robots safe navigation. Sensors, 23(4): 2337. https://doi.org/10.3390/s23042337

[17] Newcombe, R.A., Lovegrove, S.J., Davison, A.J. (2011). DTAM: Dense tracking and mapping in real-time. In 2011 International Conference on Computer Vision, Barcelona, Spain, pp. 2320-2327. https://doi.org/10.1109/ICCV.2011.6126513

[18] Liu, M., Chen, M., Wu, Z., Zhong, B., Deng, W. (2024). Implementation of intelligent indoor service robot based on ROS and deep learning. Machines, 12(4): 256. https://doi.org/10.3390/machines12040256

[19] Ding, X., Luo, Y., Yu, Q., Li, Q., Cheng, Y., Munnoch, R., Xue, D., Cai, G. (2017). Indoor object recognition using pre-trained convolutional neural network. In 2017 23rd International Conference on Automation and Computing (ICAC), Huddersfield, UK, pp. 1-6. https://doi.org/10.23919/IConAC.2017.8081986

[20] Bhandari, A., Prasad, P.W.C., Alsadoon, A., Maag, A. (2021). Object detection and recognition: Using deep learning to assist the visually impaired. Disability and Rehabilitation. Assistive Technology, 16(3): 280-288. https://doi.org/10.1080/17483107.2019.1673834

[21] Socher, R., Huval, B., Bath, B., Manning, C.D., Ng, A. (2012). Convolutional-recursive deep learning for 3d object classification. Advances in Neural Information Processing Systems, 25.

[22] 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

[23] He, K., Gkioxari, G., Dollar, P., Girshick, R. (2020). Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2): 386-397. https://doi.org/10.1109/tpami.2018.2844175

[24] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C. (2016). SSD: Single shot multibox detector. In Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, Proceedings, Part I. Springer International Publishing, 14: 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[25] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA, pp. 779-788. https://doi.org/10.1109/CVPR.2016.91

[26] Redmon, J., Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Honolulu, HI, USA, 6517-6525. https://doi.org/10.1109/CVPR.2017.690

[27] Redmon, J., Farhadi, A. (2018). YOLOv3: An incremental improvement. https://doi.org/10.48550/arXiv.1804.02767

[28] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P. (2020). Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2): 318-327. https://doi.org/10.1109/TPAMI.2018.2858826

[29] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. https://doi.org/10.48550/arXiv.1704.04861

[30] Kingma, D.P., Ba, J. (2014). Adam: A method for stochastic optimization. https://doi.org/10.48550/arXiv.1412.6980

[31] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X. (2016). {TensorFlow}: A system for {Large-Scale} machine learning. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), pp. 265-283. https://doi.org/10.48550/arXiv.1605.08695