Neural Processing Letters (2020) 51:2265–2279
https://doi.org/10.1007/s11063-020-10197-9

An Evaluation of RetinaNet on Indoor Object Detection for Blind and Visually Impaired Persons Assistance Navigation

Mouna Afif[1] · Riadh Ayachi[1] · Yahia Said[1,2] · Edwige Pissaloux[3] · Mohamed Atri[4]

Published online: 23 January 2020

Abstract
Indoor object detection presents a computer vision task that deals with the detection of specific indoor classes. This task attracts a lot of attention, especially in the last few years. The strong interest related to this field can be explained by the big importance of this task for indoor assistance navigation for visually impaired people and also by the phenomenal development of the deep convolutional neural networks (Deep CNN). In this paper, an effort is made to perform a new indoor object detector using the deep convolutional neural network-based framework. The framework is built based on the deep convolutional neural network "RetinaNet". Evaluation is done by using various backbones as ResNet, DenseNet, and VGGNet in order to improve detection performances and processing time. We obtained very encouraging results coming up to 84.61% mAP as detection precision.

Keywords Indoor object recognition · Visually impaired people (VIP) · Deep convolutional neural network (DCNN) · Deep learning · Indoor object detection and recognition dataset (IODR)

1 Introduction

Detecting and recognizing indoor objects present a challenging problem in the artificial intelligence field. Especially for indoor robot navigation and for visually impaired people assistance navigation. Blindness and visual impairments pose a parsing need to develop new automatic systems to assist persons presenting visual impairments. It will be very useful to develop a new application for detecting and recognizing indoor objects to fully assist

✉ Mouna Afif
mouna.afif@outlook.fr

[1] Laboratory of Electronics and Microelectronics (EμE), Faculty of Sciences of Monastir, University of Monastir, Monastir, Tunisia

[2] Electrical Engineering Department, College of Engineering, Northern Border University, Arar, Saudi Arabia

[3] LITIS Laboratory & CNRS FR 3638, University of Rouen Normandy, Rouen, France

[4] College of Computer Science, King Khalid University, Abha, Saudi Arabia

persons with visual impairments in their daily indoor navigation. Detecting specific indoor objects poses a challenging problem since many objects present few pixels wide the input images, also some objects are occluded with other objects and others cannot be detected because of the lighting conditions. All these issues make our work especially hard and challenging. Having the ability to navigate from one place to another presents a significant part from our daily lives.

Detecting and recognizing objects in images and videos present an important task by making computers able to understand the surrounding environment. For a human being it is vital to recognize and interact with the surrounding objects. Solving the perception problem presents a challenging task especially for visually impaired persons (VIP).

Indoor object detection presents a typical process widely used in computer vision and image processing areas in order to detect a particular class. Basically, an indoor object detection algorithm requires a huge amount of data pre-annotated in order to train DCNN and to extract features related to specific classes. Indoor object detection and classification present a highly challenging task due to the high-level variation among objects of the same category and the variation between objects belonging to different object categories (intra-class and inter-class variation).

Locating and recognizing indoor objects in an input image present a very important application which makes computers able to understand the surrounding environment. To develop an application for indoor object detection, we need to make sense of the object that we want to recognize as good as the human brain because we develop this application for visually impaired people. The object detection problem is divided into the classification and regression part. The classification part provides the object class while the regression part provides object bounding box (bbox) coordinates.

The aim of the object detection task is to determine whether the input image or video presents a given class and estimating the object location by outputting the bounding box (bbox) that overlaps the object instance. Detection step is obtained by extracting important features from the input image or video. This task forms a preliminary step for several computer vision tasks as object recognition [1], image segmentation [2] and road sign detection [3]. DCNNs were first used for image classification task and by showing best performances in image classification, DCNN has been used to deal with more complex tasks as object detection from images. Many works were proposed to enhance the Convolutional neural networks (CNNs) and to increase their performances. A deformable convolution network was proposed in [4] to enhance the transformation modeling capacity of deep convolutional neural networks.

The Indoor navigation of visually impaired people presents an important computer vision issue to be considered. Exploring new unfamiliar environments, blind, partially-blind and persons with visual impairments require assistive systems for their indoor navigation. By applying the advantages of deep learning algorithms, we propose in this work to develop an indoor object detector based on one stage convolutional neural network called “RetinaNet” [5].

Our aim from this work is to design an indoor object detection system that will accurately locate and classify the indoor objects present in an input image. The developed method was trained and tested under our proposed indoor object detection and recognition dataset (IODR). The apported value of the proposed work compared to the-state-of-the-art works is that the train and test process were done under challenging images in order to increase the robustness of indoor objects detection and to ensure more security for blind and visually impaired persons during their navigation in indoor environments. Results obtained outperforms those obtained in the previous works.

2 Related Work

Indoor objects detection and recognition present a challenging task in computer vision area. This task is crucial and difficult as indoor environments images with complex background which makes object recognition difficult. Our object from this work is to develop an indoor object detection system robust for identifying indoor objects under challenging conditions as various lighting conditions, different image backgrounds, inter and intra-class variation. Indoor scene recognition presents a crucial problem in image and video processing area. There have been numerous works done on object detection and recognition. Various applications for locating automatically have gone so far enough for identifying objects.

A large number of works were elaborated showing the big interest on detecting objects by developing new applications touching several areas as visual question answering [6], autonomous robot navigation [7], pedestrian detection [8], face detection [9], object detection [10, 11], image captioning [12] and visual grounding [13].

Many academic researchers have been elaborated on the indoor navigation field. SLAM technique presents a popular way in order to implement localization techniques [14]. In [15], the authors proposed an unsupervised deep learning model used especially for dimension reduction based on local deep features alignment (LDFA). To ensure a full image understanding, we should not only try to classify and recognize images only but also to precise accurately objects locations present in the input image. This task is object detection [16] which contains many sub-tasks as face detection [17] and pedestrian detection [18]. Object detection task contributes to providing valuable information, it is related to many applications including human behavior analysis [19], autonomous driving systems [20] and image recognition [21].

Building an appropriate indoor environment representation for an autonomous robot presents an addressed problem for the robotic community. In [22] authors presented a multi-model place classification that allows the robot identifying places and recognize indoor environments. Yu et al. [23] proposed an end to end deep neural network used for place recognition. In this work, authors used spatial pyramid structures of input images to enhance the vector of locally aggregated descriptors (VLAD). Numerous associated works on indoor object detection used for guiding indoor robot navigation employed RGB-D cameras sensors as these types of cameras provide not only the images' color but also images' depth information. Jiang et al. [24] developed a system which is used to recognize small-texture objects using a Kinect RGB-D camera.

Recently, deep learning algorithms have gained impressive attention in the artificial intelligence area. Ding et al. [25] presented a prior knowledge-based deep learning method aiming to enable the robot to recognize indoor objects places. Ding et al. [26] proposed a pipeline based on DCNNs for indoor object detection. A 3D object detection system based on multi-channel CNN is proposed in [27] in order to train the detection system. The authors conducted training and testing process using NYU V2 [28] and SUN RGB-D dataset [29]. Li et al. [30] proposed an object detection system based on deep learning algorithms specially designed for detecting small samples. In our previous work [31, 32], an indoor object recognition system demonstrated a high recognition rate comparing to state-of-the-art methods. DCNNs used for detection tasks are usually divided into two main categories: single-stage detectors or one-stage detectors and two-stage object detectors. In two-stage object detectors as Faster-RCNN [33] and mask–RCNN [34], it uses the Region Proposal Network (RPN) to generates ROIs for the first stage then the ROIs are transferred to the second stage for object classification and bounding box (bbox) regression.

The big weakness of two-stage object detectors is that they are very slow however, they contribute generally for high detection performances of the foreground and the background candidates. On the other hand, the one-stage object detectors as SSD [35] and YOLO [36] do not present a pre-treatment for foreground candidates but they treat the object detection problem as a simple regression problem.

The proposed indoor detection system presents an outperformance compared with the state-of-the-art models. The developed detector was evaluated under various feature extractor backbones; it achieves high performance in terms of detection precision as well as processing speed. This work presents the first indoor detection system that treats dangerous situations as the downstairs and prevents the blind and visually impaired people in order to ensure for them more security for their indoor navigation.

The remainder of the rest of this paper is the following: Sect. 3 outlines the contributions ported by this work. In Sect. 4, we provide an overview of the proposed indoor detection system. Section 5 provides the results of the experiments conducted for indoor object detection and recognition. Section 6 concludes the paper.

3 Contributions

- This paper presents the first approach evaluating RetinaNet architecture in detecting and recognizing indoor objects.
- The proposed detection system achieved high accuracies for indoor object detection and classification in complex image conditions.
- This is the first approach using a one stage DCNN detector for indoor object detection.
- This method achieved high detection performances in challenging conditions, such as extreme illumination changes, occlusion and high inter-class and intra-class variation.
- Finally, the proposed indoor detection system was trained using the proposed indoor object detection and recognition dataset that was not previously studied.

4 Proposed Approach for Indoor Object Detection Based on RetinaNet

The proposed detection system is specially designed to detect indoor objects. This method is built to be operated in embedded systems. It requires much less energy consumption than a computer desktop. Based on these processing aspects from speed and accuracy, we will use the one-stage object detector to develop our indoor object detection system using RetinaNet [5].

4.1 Overview of the Proposed Method

The biggest problem faced by state-of-the-art detectors is the class imbalance issue. For this fact, RetinaNet [5] introduces a reshaping standard cross-entropy function, so it down weights the loss assignment to increase the good classification of classes. RetinaNet is a specific dense detector that can include various model backbones. For a DCNN, the backbone is a necessary part as it is used for feature extraction from the input image. RetinaNet uses various backbones as ResNet [37], DenseNet [38], or VGGNet [39]. The model backbone is an encoder that processes the input images using convolution kernels in order to extract the high relevant image features.

As presented in Fig. 1 of the proposed method, the input images are from RGB color space. This indoor image is used as an input for the RetinaNet network. As an output of RetinaNet, classes, and positions coordinate of the indoor objects present in the image are determined.

4.2 Proposed Indoor Detection System Architecture Details

The proposed detection system used for indoor object detection and for visually impaired people assistance navigation is based on the RetinaNet DCNN. RetinaNet architecture proposes a new loss function that contributes to more effectiveness compared with other approaches on class imbalance. The loss function presents a dynamical scale cross-entropy loss which consists on decaying to zero and the correct class confidence decrease. Focal loss function handles perfectly class imbalance caused by the few numbers of object instances. It perfectly addresses the one-stage object detection process and increases the imbalance between foreground and background during the training step.

The focal loss is the reshaping of the cross-entropy loss. It consists of down-weights less attributed to the well-classified samples when having a high-class imbalance, a "balancing parameter is added".

To effectively address the one-stage detectors problem, this is the extreme imbalance between foreground and background class during the training process. Equation one presents the cross-entropy function.

$$CE(p, y) = \begin{cases} -log(p), & p = 1 \\ -log(1 - p), & otherwise \end{cases} \tag{1}$$

This loss function is used for binary classification where:

y ϵ {-1,1}: Ground truth class

p ϵ {0,1}: model estimated probability

Focal loss function handles perfectly the class imbalance caused by the few numbers of the object instances. The common method to address the class imbalance problem is to introduce a rectifier weight α, while α_t ϵ [0, 1] for class 1 and 1-α_t for class -1. For national convinces they define p_t. The α_t balanced Cross entropy is the following:

$$CE(P_t) = -\alpha_t \log(P_t) \tag{2}$$

$$P_t = \begin{cases} 1 & if\ y = 1 \\ 1 - P, & otherwise \end{cases}$$

During the training process, a large class imbalance will appear for the one-stage detector. α_t balances the positive/negative examples while it does not differentiate between easy

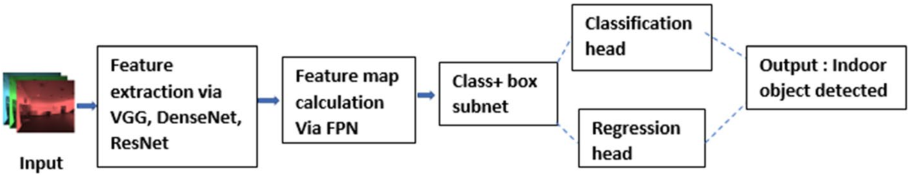

Input
Feature extraction via VGG, DenseNet, ResNet
Feature map calculation Via FPN
Class+ box subnet
Classification head
Regression head
Output : Indoor object detected

Fig. 1 Proposed pipeline for indoor object detection

or hard samples. For this fact, RetinaNet introduces the “focal loss” in order to reshape the loss function to down-weight the easy examples and focusing on training the network on the hard negative. The focal loss function is the following:

$$FL\left(P_t\right) = -\alpha_t(1 - p_t)^{\gamma}\log(P_t) \tag{3}$$

For the focal loss function, a modulating factor $(1-p_t)^{\gamma}$ is added while $\gamma >= 0$. The focal loss is applied several volumes of $\gamma \in [0, 5]$. The focal loss presents two main properties:

(1) For the miss classified examples and a small p_t value, the modulating factor is near 1 and the loss function is unaffected. For $p_t \rightarrow 1$, the modulating factor goes to 0 and the loss for the well-classified examples is down-weighted.

(2) For the focusing parameter γ adjusting smoothly the down weighting the easy example rate. When $\gamma = 0$, FL equivalent to CE. If γ is increased the effect of the modulating parameter is increased. The focal loss function yielding to improve slightly network accuracy.

For binary classification, models are initialized to have equal outputs ($y = -1$ or 1). Depending on the model initialization and on the presence of the class imbalance, the frequent class loss can dominate the total loss. To address this problem, RetinaNet architecture introduces a “prior” value of p-estimated used for the rare classes (foreground) at the start of the training step. This technique improves the training process stability for cross-entropy and focal loss when a heavy class imbalance is present.

Usually, two-stage object detectors are trained using cross-entropy with non-use of α balancing parameter. Two-stage detectors address the class imbalance problem by using either the two-stage cascade or by using mini-batch size sampling. The proposed indoor detection system is built to respect embedded system resources limits, which present less energy and memory consumption compared with a desktop computer resources requirements.

By considering all the challenging aspects of detection precision and processing speed, we used the RetinaNet-based architecture as mentioned in Fig. 2.

RetinaNet network is based especially on:

- A backbone builded on the top of the feature extractor (backbone) called feature pyramid network FPN used to calculate convolution feature map of the input image.
- A classification head: a subnetwork used to perform classification using the backbone’s output.

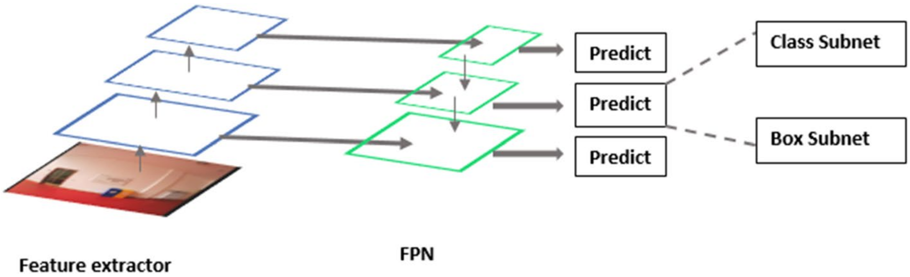


Fig. 2 RetinaNet architecture used for indoor object detection system

- A regression head: a subnetwork responsible for performing bounding ox coordinates using the backbone's output.

Feature Pyramid Networks (FPNs) are composed of a bottom-up and a top-down pathway. The bottom-up pathway usually presents a convolutional neural network used for feature extraction. FPN provides construct higher resolution layers from the semantic layers. Even the reconstructed layers are semantically strong, but after down-sampling and up-sampling objects locations are not precise. For this fact, lateral connections are added to predict object locations better.

Region proposal networks (RPN) present feature map sub-sampling that crosses a lot of pertinent image information; this enables the network to detect small objects. For this fact, the Features Pyramid Network (FPN) solves this problem. Using a backbone background is extremely necessary to extract the feature map from the entire image. RetinaNet architecture consists of an FPN backbone including subnets for class recognition and for regression pipelined with an encoder using residual networks ResNet [37] or dense convolutional networks DenseNet [38] or Visual Geometry Group VGG [39].

ResNet 152 [37] presents the deepest network of ResNet family. It presents a very powerful block named residual block which adds input to the output via shortcut connection in order to reduce the network calculation complexity.

DenseNet [38] presents shorter connections between layers in order to achieve more accurate network results. Dense connectivity requires few numbers of parameters compared with other traditional DCNNs which makes them more lightweight. DenseNet architecture presents an improved features map information through the network which makes it an easy neural network to train. DenseNet architecture provides four dense blocks with three transition layers. A dense block is composed of different convolutions and non-linear layers followed by RELU and batch normalization BN functions; each layer is connected to all other layers of the same dense block. DenseNet architecture shows an important advantage as it alleviates gradient decent problem and it reduces considerably parameters number of the network. The transition layer is composed of convolution and average pooling layers in order to decrease the feature map size and minimize time and calculation complexity. The classification head is composed of an average polling layers followed by fully connected layers (FC) used to calculate the class score and a softmax layer. The following table presents a detailed diagram of the dense block and transition layers present in DenseNet-121 architecture used in our experiments (Table 1).

In VGG Net 16 the input image is passed through a stack of convolutional layers. The convolution stride is fixed to 1 pixel. VGG Net 16 presents also max pooling layers which contribute to minimize the feature map dimension. Fully connected layers are used to calculate classes score.

The proposed indoor detection system is based on a one stage neural network composed of a backbone network and two sub-networks. The first network is responsible for performing convolutional object classification on the backbones output and the second sub-network is responsible for extracting bounding box regression and the object class name. The proposed detection application adopts FPN as a backbone. Briefly, FPN increase the standard convolution of the network and construct a rich multi-scale information from the input image each level of the FPN architecture is used for detecting objects at different scales. Using FPN highly improves multi-scales predictions as it detects object in different scales (small, medium and large). Detecting objects over multiple scales is a challenging problem in particular for small objects. This feature pyramid network combines low-resolutions

Table 1 DenseNet-121 diagram: composition details of dense block and transition layers

Layers	DenseNet-121
Convolution	7 × 7 conv, stride 2
Pooling	3 × 3 max pool, stride 2
Dense block (1)	$\begin{bmatrix} 1\times1 & conv \\ 3\times3 & conv \end{bmatrix} \times 6$
Transition layer (1)	1 × 1, conv 2 × 2, average pooling, stride 2
Dense block (2)	$\begin{bmatrix} 1\times1 & conv \\ 3\times3 & conv \end{bmatrix} \times 12$
Transition layer (2)	1 × 1, conv 2 × 2, average pooling, stride 2
Dense block (3)	$\begin{bmatrix} 1\times1 & conv \\ 3\times3 & conv \end{bmatrix} \times 24$
Transition layer (3)	1 × 1, conv 2 × 2, average pooling, stride 2
Dense block (4)	$\begin{bmatrix} 1\times1 & conv \\ 3\times3 & conv \end{bmatrix} \times 16$
Classification head	7 × 7 average pooling 1000 D FC, softmax

semantically strong features with semantically weak features with high resolutions via button-up pathway and top-down pathway and lateral connections.

The bottom-up pathway is the feed-forward computation of the DCNN backbone. For each stage, one feature pyramid layer is associated. Each last output of each stage will be used as a feature map to feed the top-down pathway using the lateral connections. To reduce the number of the anchor boxes detected, a non-maximum-suppression (NMS) is applied to select the anchor with the height confidence score. In this stage, every remaining anchor will be used as a bounding box prediction.

5 Experiments and Results

We trained and tested the DCNN model on our collected dataset [40] which present complex conditions, complex image backgrounds with multiple illumination conditions. The proposed dataset contains 8000 indoor images with 16 indoor landmark objects. The proposed dataset contains two image resolutions: 1616 × 1232 and 4592 × 3448. For the proposed experiments, we selected 5300 images for the network training and 2700 to test the DCNN. Parameters used during the training process are the following: 50 epochs, each epoch contains 10.000 iterations, learning rate initialized at 0. 0001. All the experiments were performed on a Desktop computer equipped with Intel Xeon E5-2683 V4 processor, NVIDIA Quadro M 4000 graphic card with 8 GB of graphic memory. The network implementation was done by using the Keras framework. For more specifications, we used python 3.6 setups, TensorFlow-GPU 1.13, NVIDIA CUDA toolkit 10.0 and deep neural network library CUDNN version 7.0.

The proposed dataset used in these experiments is provided in our previous work [40]. In addition, we trained the proposed indoor detection system using different backbones

(ResNet, DenseNet, VGGNet). The dataset was divided into two subsets: training and testing subsets respectively. The overall performance of the proposed indoor object detection system was measured using mean average precision (mAP). Training a DCNN requires a huge amount of data. For this fact, data augmentation was used to increase the training data number. Data augmentation was used by performing multiple processing in images as scaling; rotations; translation; and flipping. Figure 3 presents some examples of the processed images.

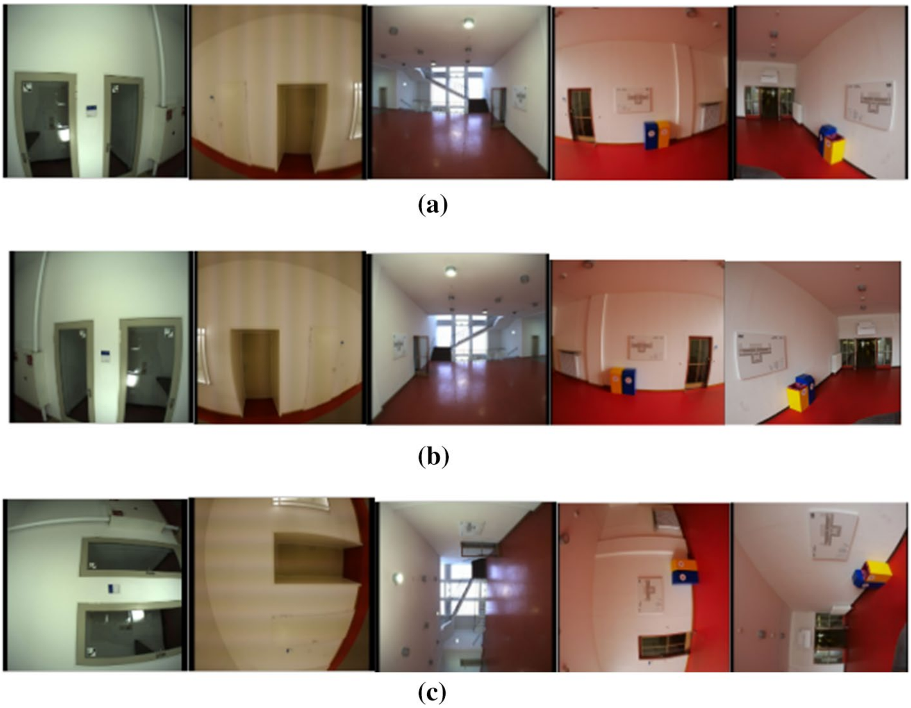
For the training process, RetinaNet architecture uses the stochastic gradient descent [41] SGD as an optimizer of the loss function. SGD updates the network parameters at each training step. But SGD performs updates followed by high oscillations of the objective function. These high oscillations cause many troubles for the loss function and lead to do not reach the local minimum. SGD updates are measured following Eq. (4).

$$w = w - \eta * \nabla w * J\left(w; x^i; y^i\right) \tag{4}$$

where w is model's parameters (weights + bias), $\nabla w * J(w)$ is the objective function, η is the learning rate.

The following Tables 2, 3 and 4 present the results of our implementation of the proposed indoor object detection system with different backbones for feature extraction and by using SGD as a network optimizer.

Mean average precision (mAP) was selected as a metric to evaluate the performance of the proposed indoor object detection system. mean average precision presents the precision average of all the object classes present in the proposed IODR dataset. While



(a)

(b)

(c)

Fig. 3 Processed images examples: **a** input images, **b** flipped images, **c** rotated images

Table 2 Implementation results obtained by using VGG 16 as a feature extractor and SGD as a network optimizer

Class name	Window	Notice table	Elevator	Door	Electricity box	Sign	Light	Trash can
AP (%)	59.83	73.29	77.58	96.94	68.64	70.56	67.91	80.83
Class name	Stairs	Security button	Table	Smoke detector	Heating	Fire extinguisher	Light switch	Chair
AP (%)	69.63	60.47	81.17	37.74	76.63	74.19	44.97	77.73

Table 3 Implementation results obtained by using DenseNet 121 as a features extractor and SGD as a network optimizer

Class name	Window	Notice table	Elevator	Door	Electricity box	Sign	Light	Trash can
AP (%)	63.81	88.38	89.55	95.54	94.28	73.19	69.38	87.83
Class name	Stairs	Security button	Table	Smoke detector	Heating	Fire extinguisher	Light switch	Chair
AP (%)	85.93	64.73	99.55	45.62	96.12	79.82	44.78	99.67

the average precision (AP) provides the detection precision for one specific class. The detailed results obtained during our implementations with various features extractors are presented in Tables 2, 3 and 4.

As provided in these previous tables, the proposed object detection system achieves a very encouraging detection rate. Almost perfect recognition was obtained for chair, table, trash can, electricity box, door, elevator, notice table, and heating categories; good performances were obtained in the detection of many other indoor classes such as: stair, sign and fire extinguisher.

We note that by using Resnet 152 as a backbone, we achieved 83.15% mean average precision (mAP). And by using DenseNet 121 we achieved 79.88% mAP and 69.88% when using VGG Net 16.

Table 4 Implementation results obtained by using Resnet 152 as a features extractor and SGD as a network optimizer

Class name	Window	Notice table	Elevator	Door	Electricity box	Sign	Light	Trash can
AP (%)	67.95	94.38	93.58	97.67	95.95	76.82	73.88	92.83
Class name	Stairs	Security button	Table	Smoke detector	Heating	Fire extinguisher	Light switch	Chair
AP (%)	87.97	68.73	99.67	50.77	97.81	83.19	51.18	98.12

Table 5 Implementation results obtained by using VGG 16 as a features extractor and ADAM as a network optimizer

Class name	Window	Notice table	Elevator	Door	Electricity box	Sign	Light	Trash can
AP (%)	61.80	75.39	76.58	98.33	70.74	71.66	69.11	81.93
Class name	Stairs	Security button	Table	Smoke detector	Heating	Fire extinguisher	Light switch	Chair
AP (%)	71.43	63.22	82.99	38.74	77.83	77.15	45.97	79.53

At a first time in the proposed experiments, we trained the proposed detection system with SGD. To solve the problem caused by the SGD optimizer. We propose to modify the optimizer by ADAM [42]. ADAM optimizer updates quickly the parameters. By keeping an exponential decay of the past gradient decay $\mathrm{m_t}$ (the first momentum of the gradient). Moreover, ADAM optimizer stores also an exponential decay of the average of the past squared gradient $\mathrm{v_t}$ (the second momentum of the gradient). The biggest strength of the ADAM optimizer is that it computes an adaptive learning rate for each parameter during the training process using the following equations.

$$m_t = \beta_1 * m_{t-1} + \left(1 - \beta_1\right) * g_t \tag{5}$$

$$V_t = \beta_2 * V_{t-1} + \left(1 - \beta_2\right) * g_t^2 \tag{6}$$

where β_1 *and* β_2 are close to 1.

Adam performs a biases correction of the first and the second momentum. The biases corrected first ($\hat{m}_t$) and second ($\hat{v}_t$) can be estimated as the following equations:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{7}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{8}$$

Then, Adam updates the network parameters using the corrected first and second moment as Eq. (9).

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\hat{v}_t + \varepsilon}} * \hat{m}_t \tag{9}$$

The testing step was performed on the same desktop computer with the same configuration of the training step. To measure the performance of the proposed detection system on the indoor object detection and recognition dataset, the ground truth box positions including indoor objects were manually annotated in the input images, then the ground truth boxes are compared with the overlapping regions between ground truth and the detected boxes. Tables 5, 6 and 7 provide the detection accuracies obtained per object class for the three backbones using ADAM as a neural network optimizer.

Table 6 Implementation results obtained by using DenseNet 121 as a backbone and ADAM as a network optimizer

Class name	Window	Notice table	Elevator	Door	Electricity box	Sign	Light	Trash can
AP (%)	65.37	91.21	90.55	97.11	96.39	74.99	73.38	90.13
Class name	Stairs	Security button	Table	Smoke detector	Heating	Fire extinguisher	Light switch	Chair
AP (%)	87.54	66.79	98.55	49.42	97.92	80.72	48.45	98.97

Table 7 Implementation results using Resnet 152 as a backbone and ADAM as a network optimizer

Class name	Window	Notice table	Elevator	Door	Electricity box	Sign	Light	Trash can
AP (%)	69.35	96.88	95.43	98.97	97.42	80.33	75.37	94.39
Class name	Stairs	Security button	Table	Smoke detector	Heating	Fire extinguisher	Light switch	Chair
AP (%)	90.65	67.73	98.93	51.87	98.13	85.34	54.11	98.92

By using the ADAM optimizer in the proposed detection system to train the network, we gained around 2% in the total mean average precision (mAP) for the different feature extractors used.

In the following table, we present a comparison between the results of the implementations obtained using SGD optimizer and those obtained with the ADAM optimizer for all the backbones used.

As presented in Table 8 when using the ADAM optimizer, we succeeded to improve the indoor object detection accuracy more than when using the SGD optimizer. We note that by using ADAM as a network optimizer we gained around 2% in total accuracy for the three features extractors used in our experiments.

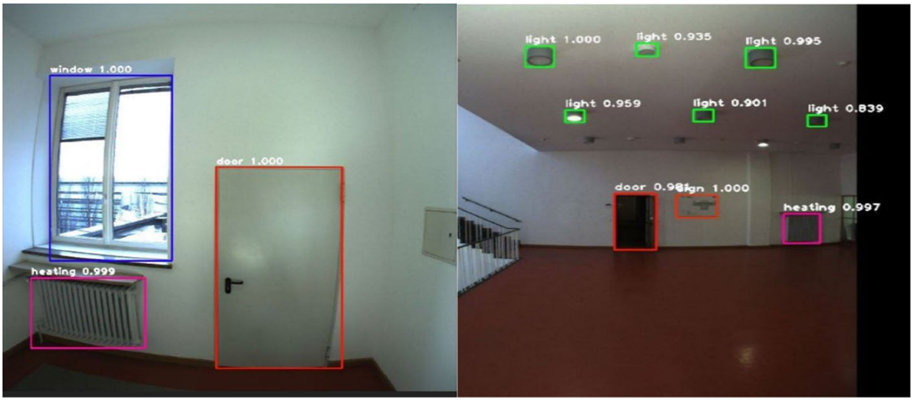


Fig. 4 A detection example of the proposed system

Table 8 Implementation results when using SGD vs results obtained when using ADAM optimizer

ADAM (mAP%) (%)	SGD (mAP%) (%)
VGG Net 16: 71.4	VGG Net 16: 69.88
DenseNet 121: 81.71	DenseNet 121: 79.88
ResNet 152: 84.61	ResNet 152: 83.15

Table 9 Comparison of results obtained by our method and those obtained in [26]

Indoor object name	Method (AP%) [26]	Ours (AP%)
Chair	94.0	98.92
Door	91.4	98.97
Table	91.6	98.93

As presented in Table 9 our proposed indoor object detection system achieves better results than the results obtained in [26] for the three common treated classes. We gained around 5%,7% and 7% for chair, door and table classes respectively.

Figure 4 provides a detection example of the proposed method. This figure shows that all the objects present in the two images are detected with high average precision.

We can conclude that the proposed indoor object detection and recognition system achieves a high detection rate for different indoor objects with different sizes.

6 Conclusion

In this paper, we propose a new indoor object detection system based on the one-stage object detector RetinaNet. The proposed detection system shows a big efficiency in detecting indoor objects in challenging conditions. It achieves high detection precisions. Through all the experiments conducted using RetinaNet architecture with different backbones as ResNet, DenseNet, and VGGNet, we improve the effectiveness of the proposed indoor detection system by modifying the network optimizer. The present work provides more interest by using deep learning applications to towards improving daily life for partially-impaired, blind and visual impaired persons by giving a comprehensive idea about the surrounding indoor environment.

References

1. Hu H, Gu J, Zhang Z et al (2018) Relation networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3588–3597
2. Liu F, Lin G, Shen C (2015) CRF learning with CNN features for image segmentation. Pattern Recognit 48(10):2983–2992
3. Ayachi R, Afif M, Said Y et al (2019) Traffic signs detection for real-world application of an advanced driving assisting system using deep learning. Neural Process Lett. https://doi.org/10.1007/s11063-019-10115-8
4. Dai J, Qi H, Xiong Y et al (2017) Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 764–773
5. Lin T-Y, Goyal P, Girshick R et al (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988

6. Yu Z, Yu J, Xiang C et al (2018) Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans Neural Netw Learn Syst 29(12):5947–5959
7. Cosio FA, Castaneda MAP (2004) Autonomous robot navigation using adaptive potential fields. Math Comput Model 40(9–10):1141–1156
8. Dollar P, Wojek C, Schiele B et al (2009) Pedestrian detection: a benchmark
9. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154
10. Dai J, Li Y, He K et al (2016) R-fcn: object detection via region-based fully convolutional networks. In: Advances in neural information processing systems, pp 379–387
11. Ghiasi G, Lin T-Y, Le QV (2019) NAS-FPN: learning scalable feature pyramid architecture for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7036–7045
12. Yu J, Li J, Yu Z et al (2019) Multimodal transformer with multi-view visual representation for image captioning. IEEE Trans Circuits Syst Video Technol. https://doi.org/10.1109/tcsvt.2019.2947482
13. Yu Z, Yu J, Xiang C et al (2018) Rethinking diversified and discriminative proposal generation for visual grounding. In: Proceedings of the 27th international joint conference on artificial intelligence. AAAI Press, pp 1114–1120
14. Newcombe RA, Lovegrove SJ, Davison AJ (2011) DTAM: dense tracking and mapping in real-time. In: 2011 international conference on computer vision. https://doi.org/10.1109/iccv.2011.6126513
15. Zhang J, Yu J, Tao D et al (2018) Local deep-feature alignment for unsupervised dimension reduction. IEEE Trans Image Process 27(5):2420–2432
16. Felzenszwalb PF, Girshick RB, Mcallester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell 32(9):1627
17. Sung KK, Poggio T (2002) Example-based learning for view-based human face detection. IEEE Trans Pattern Anal Mach Intell 20(1):39–51
18. Wojek C, Dollar P, Schiele B, Perona P (2012) Pedestrian detection: an evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell 34(4):743
19. Cao Z, Simon T, Wei S-E, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR
20. Chen X, Ma H, Wan J, Li B, Xia T (2017) Multi-view 3d object detection network for autonomous driving. In: CVPR
21. Yu J, Tan M, Zhang H et al (2019) Hierarchical deep click feature prediction for fine-grained image recognition. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/tpami.2019.2932058
22. Pronobis A, Martinez Mozos O, Caputo B et al (2010) Multi-modal semantic place classification. Int J Robot Res 29(2–3):298–320
23. Yu J, Zhu C, Zhang J et al (2019) Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2019.2908982
24. Jiang L, Koch A, Zell A (2016) Object recognition and tracking for indoor robots using an RGB-D sensor. In: Intelligent autonomous systems 13. Advances in intelligent systems and computing, vol 302. Springer, Cham, pp 859–871
25. Ding X, Luo Y, Li Q et al (2018) Prior knowledge-based deep learning method for indoor object recognition and application. Syst Sci Control Eng 6(1):249–257
26. Ding X, Luo Y, Yu Q et al (2017) Indoor object recognition using pre-trained convolutional neural network. In: 2017 23rd international conference on automation and computing (ICAC). IEEE, pp 1–6
27. Wang L, Li R, Shi H et al (2019) Multi-channel convolutional neural network based 3D object detection for indoor robot environmental perception. Sensors 19(4):893
28. Silberman N, Hoiem D, Kohli P, Fergus R (2012) Indoor segmentation and support inference from RGBD images. In: Proceedings of the 12th European conference on computer vision (ECCV 2012), Florence, Italy, 7–13 Oct 2012, pp 1–14
29. Song S, Lichtenberg SP, Xiao J (2015) SUN RGB-D: a RGB-D scene understanding benchmark suite. In: Proceedings of the 2015 IEEE conference on computer vision and pattern recognition (CVPR 2015), Boston, MA, USA, 7–12 June 2015; pp 567–576
30. Li C, Zhang Y, Qu Y (2018) Object detection based on deep learning of small samples. In: 2018 tenth international conference on advanced computational intelligence (ICACI). IEEE, pp 449–454
31. Afif M, Ayachi R, Said Y, Pissaloux E, Atri M (2018) Indoor image recognition and classification via deep convolutional neural network. In: International conference on the sciences of electronics, technologies of information and telecommunications. Springer, Cham, pp 364–371

32. 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). IEEE, pp 1–4
33. Ren S, He K, Girshick R et al (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99
34. He K, Gkioxari G, Dollar P et al (2017) Mask R-CNN. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969
35. Liu W, Anguelov D, Erhan D et al (2016) SSD: single shot multibox detector. In: European conference on computer vision. Springer, Cham, pp 21–37
36. Redmon J, Farhadi A () YOLOv3: an incremental improvement. arXiv arXiv:1804.02767
37. He K, Zhang X, Ren S et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
38. Huang G, Liu Z, Van Der Maaten L et al (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
39. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
40. Afif M, Ayachi R, Said Y, Pissaloux E, Atri M (2019) A novel dataset for intelligent indoor object detection systems. Artif Intell Adv 1(1):52–58
41. Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: COMPSTAT'2010. Physica-Verlag HD, pp 177–186
42. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980