Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-09662-3

Indoor objects detection and recognition for an ICT mobility assistance of visually impaired people

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

Received: 3 November 2019 / Revised: 17 July 2020 / Accepted: 18 August 2020
Published online: 22 August 2020

Abstract
Indoor object detection in real scene presents a challenging computer vision task; it is also a key component of an ICT autonomous displacement assistance of Visually Impaired People (VIP). To handle this challenge, a DCNN (Deep Convolutional Neural Networks) for indoor object detection and a new indoor dataset are proposed. The novel DCNN design is based on a pre-trained DCNN called YOLO v3. In order to train and test the proposed DCNN, a new dataset for indoor objects was created. The images of the new dataset present large variety of objects, of indoor illuminations and of indoor architectural structures potentially unsafe for a VIP independent mobility. The dataset contains about 8000 images and presents 16 indoor object categories. Experimental results prove the high performance of the proposed indoor object detection as its recognition rate (a mean average precision) is 73,19%.

Keywords Indoor object detection and recognition · Deep convolutional neural networks (DCNN) · Visually impaired people (VIP) mobility · Indoor navigation

1 Introduction

Indoor object detection and indoor scene understanding are basic tasks for many applications including autonomous robot navigation [48] and mobility assistive devices for people with visual impairments (VIP) [17].

✉ Mouna Afif
mouna.afif@outlook.fr

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

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

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

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

For independent mobility, the VIPs need to perceive relevant objects of their nearest space. As the VIPs are not able to see landmarks or such (indoor) objects, an assistive device must indicate their presence.

Indoor objects' perception in real indoor scene is a challenging task as many complex problems such as background complexity, occlusions, viewpoint changes, etc. should be taken into account. To address this problem, a fully labeled indoor object dataset was elaborated with a goal of their detection. This dataset consists of 8000 indoor images containing 16 different and the most frequent indoor landmark objects and classes.

Moreover, the robotic and human navigation assistance requires a real-time processing. A Deep Convolutional Neural Networks (DCNN) may be a solution to achieve such temporal performance.

Deep CNN combines two concepts: Deep Learning and Convolutional Neural Networks. Such combination integrates millions of values of parameters which underlay the acquired images presentation, parameters which are relevant to perform a specific task and which are taken into account during the training phase.

Furthermore, DCNN exhibits big difference from other traditional approaches for object detection. Indeed, a DCNN models use powerful *adhoc* objects' representations by providing good features extraction process in each layer of the network.

The great particularity of Deep Learning models is the hierarchical representation of features. This means that features computed in intermediate layers can be reused in different applications and tasks, while features found by the last layers are specific and a function of the targeted application and the dataset are used. The convolution part of a DCNN (layers closer to the input layers) refers to general features, while the classification part (layers closer to the outputs) refers to specific features.

Deep learning models can be divided into two principal parts: region proposal-based models (such as R-CNN [15], Fast R-CNN [16], Faster R-CNN [43] and Mask R-CNN [19]) and proposal-free methods (such as YOLO [42], YOLO 9000 [41], and YOLOv3 [40], and SSD [35]).

The efficiency and the accuracy of object detection with the deep learning models is that DCNN extracts per-pixel features through a large number of images during the training process, although deep structures CNN models ensure a good extraction of the most relevant image features.

State-of-the-art models of deep learning widely rely on large-scale dataset such as ImageNet [9], MS COCO [33], PASCAL VOC 2007 [13], and VOC 2012 [14], and all of them generally fail in the indoor object detection; indeed, listed images' datasets present complex scenes. However, despite of the variety of backgrounds, multiple indoor objects, multiple positions, different scales, etc. the considered objects are samples of classic situations and do not consider specific needs of visually impaired people (VIP).

This paper proposes a new fully labeled dataset for indoor object detection and recognition, and relevant to VIP mobility. The originality of the proposed dataset comes from the inclusion of new characteristics of a 3D scene not considered so far, and relevant to VIP mobility. Such new training data will robustify the object recognition and may be used in any (assistive) navigation system. This paper presents the first approach evaluating YOLOv3 architecture on indoor object detection. Our aim from this work is to provide the Visually impaired person with a robust indoor object detection system to help them to more explore and interact with their surrounding environments and to more integrate in the daily life. Our proposed work achieved very encouraging results in term of detection accuracy and speed which meet the VIP

mobility requirements. Also, the proposed work achieved high detection performances in challenging conditions as: extreme lighting conditions, heavy occlusion, high intra and inter-class variation.

The remainder of the paper is organized as follows: Section 2 reviews related works on indoor object detection. Section 3 presents the proposed multi-class indoor object detection and recognition dataset. In Section 4, the proposed approach for indoor objects detection based on a pre-trained DCNN is presented. Section 5 outlines the experimental evaluation of the proposed approach and discusses the obtained results. Finally, Section 6 concludes the paper and proposes further extensions of the work.

2 Indoor object recognition: related works

Many researchers and academics show their big interest in real time indoor object detection. The challenge is to detect correctly and accurately the object in an image or in a video. Generally, the indoor environments are different from the outdoor scenery. Indoor scenery is composed generally of a wide range of background elements and different interior decorations.

Since the appearance of RGB-D sensors, such as Kinect cameras, that provide not only image color but also depth information, many works based on RGB-D sensors have been used to guide the robot indoor navigation [25]. However, the detection becomes very challenging when it is used for recognition of specific objects or unknown obstacles in unfamiliar natural environments [20].

Frequently, depth sensors have been widely used for object detection and recognition during the simultaneous localization and mapping (SLAM). Chae et al. [6] introduced a framework for indoor object recognition for SLAM in indoor scenery.

Many others classic works rely on indoor objects detection based on machine learning techniques [30, 37]. However, this category of methods usually contributes to some complex pipeline design which make them highly depended on computational resources and of a very high computational cost; the real-time constraint is usually not met.

During the last few years, DCNN models have gained a great attention in many computer visions tasks. This approach has been used for indoor object recognition [10, 11], for indoor object segmentation [8, 44], detection tasks [46], internet of things (IoT) [31] grasping force prediction [36] and authentication systems [24, 32]. In order to enhance indoor object detection, it is necessary to build and create new reliable classification and detection systems. Kim et al. [27] trained a deep CNN model, ConvNet that will be used for autonomous indoor navigation of robots. Another work based on DCNN models, which focuses on objects' prediction knowing their poses (and named PoseNet), was introduced by Kendall et al. [26]. This model combines the strengths of DCNN models with SLAM techniques, with the respect of a hard-real-time constraint.

Chen et al. [7] presented a new visual indoor positioning system based on CNN models. To better address the problem of indoor positioning, authors proposed a localization method which consists of features extraction using DCNN and pose estimation.

Sho et al. [45] proposed an indoor positioning deep learning-based system in order to satisfy the increasing demand for these types of systems. Authors adopted DCNN to implement their orientation-free positioning system. Their hybrid system is based on a location part with wifi and fingerprint images; a CNN is used to classify locations.

After many years of research in the field of deep neural networks, DCNNs still present the best choice for many computer vision problems. DCNN are widely used for indoor scene recognition [34]. To design their scene classification system authors used ResNet with all its versions (Renet-50, ResNet-101, ResNet-152) [18], datasets ImageNet 11 K [29] and places 365 [49] for training. Bashiri et al. [3] developed a detection system dedicated to detect three objects (doors, stairs and sign) using deep learning method. Their construction of an appropriate representation of a specific environment still presents a challenging issue in the robotic field. Escalona et al. [12] present a 3D object detection system based on RGB-D cameras. They used the semantic labeling image concept which is very suitable for human-robot interaction as it facilitates the interaction of the robot with the surrounding environment.

Over the last few years, data-driven DCNN outperform the classic approaches. Many indoor object detection approaches was proposed. However, none of them is suitable for independent mobility of VIP. The next section presents a novel efficient indoor objects detection approach based on a DCNN model. In addition, a new dataset was created to train and test the proposed detection system.

3 Proposed multi-class indoor object detection and recognition dataset (IODR)

Several indoor object datasets were proposed in the literature. Bachiri et al. [4] proposed a new indoor dataset used for indoor object classification presenting 20,000 indoor images and containing 3 indoor classes (door, sign, stairs). Quattoni et al. [39] proposed a new dataset used specially to deal with indoor scene recognition problems. This dataset presents 15,620 images with 67 indoor scene categories. Xiao et al. [47] proposed an extensive database named "Scene UNderstanding" (SUN) containing 899 scene categories with over 130,519 images, it is especially used for scene understanding. Their dataset present various categories such as indoor urban and nature's categories. Indoor datasets used to solve indoor object detection problems, are constrained by the indoor dataset present that do not capture a variety of indoor objects while covering the challenging situations as luminosity invariance, occlusion and various objects positions. For this fact, we will collect and fully label an indoor object dataset that will cover various challenging situations to be used in a second time to test and train the proposed indoor object detection system.

The Multi-class Indoor Object Detection and Recognition (IODR) Dataset presents a new fully labeled indoor object dataset [1]. This fully annotated dataset (§3.1) can be highly recommended for training and testing different DCNN models (§3.2) while prototyping a new indoor navigation assistance system.

3.1 Data preparation and annotation

The biggest challenge is to provide to VIPs the relevant information on their indoor navigation environment. Some images of the proposed dataset are collected from the NAVIIS project [21]; new indoor images contain vital indoor objects for VIP mobility (identified with the VIP) with different lighting conditions and complex backgrounds. The dataset was labeled using LabelImg software [22]. The new original dataset encompasses 8000 indoor annotated images where 16 indoor object classes are considered.

Figure 1 presents an example of the proposed annotation process via labelImg tool: objects are delimited by their rectangular bounding box (with their coordinates in image defined in video streaming modes).

The proposed dataset is composed of many categories of indoor objects. It contains 8000 indoor images captured, which present different lighting conditions, to obtain a very robust (scene illumination invariant) dataset. Two resolutions are present in the dataset:1616 × 1232 and 4592 × 3448.

The collected dataset contains 16 main landmark objects that are usually present in any indoor scene especially in corridors. They are: doors, light switches, smoke detector, chair, fire extinguisher, sign, window, heating, electricity box, stairs, table, security button, trash can, elevator and notice table. All images are in the .jpg format.

The proposed dataset provides various characteristics important for the VIP mobility, it is original in term of:

- Light invariance: objects are taken under different lighting conditions (day, night, blurred).
- Geometrical change invariance: the objects are taken under different angles and poses.
- Objects provided in in the proposed dataset are vital for VIP indoor mobility.
- Occlusion: parts of the objects are hidden or overlapped by other objects.
- Highlighting the presence of dangerous situations to ensure a safe mobility for the VIP person as the downstairs.
- The proposed dataset is very suitable to develop new robust indoor object detection systems.
- High inter and intra-class variation.

Figure 2 illustrates the wide intra-class variation between doors in the proposed dataset. Doors present many shapes, many poses, many colors, different textures). Annotations were done on different doors poses, their status (opened or closed), the material used is under different textures (wood, glass, iron). The biggest strength of the proposed dataset is that it provides many challenging conditions in order to perform a robust training to deal with different indoor environments belonging to various buildings and to be relevant to VIP mobility.

Fig. 1 LabelImg annotation example: .jpg image (left picture) and its annotated equivalent (bbox) (right picture)

Fig. 2 Doors intra-class variation

In contrast to the exiting indoor datasets, the proposed dataset provides many challenging conditions taken into account as: heavy occlusions, different lighting conditions, complex background, etc.… in order to increase the robustness of the indoor object detector. Also, the proposed dataset provides a high inter and intra-class variation to build an accurate detector. The proposed dataset is highly recommended for multi-objects problems as it provides various indoor object classes.

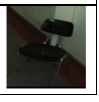
3.2 Training and testing subsets

After the selected dataset annotation, training and testing configurations must be prepared. The dataset was divided into train and test sets. For the training set, 66% of the dataset was reserved and the rest was used as testing set. The proposed dataset contains 16 indoor object classes. Table 1 presents all the indoor object classes with all classes' names and IDs to ensure a better scene understanding of the indoor images. The original images included in the dataset are selected with respect to two main issues of VIP mobility: firstly, providing the most relevant indoor objects and landmarks, and secondly, providing a good annotation to better understand the indoor scene.

4 Proposed architecture for indoor object detection

Deep learning models have proved their big performances in the computer vision area in particular for object detection tasks. Precise and fast indoor object detection and recognition, in

Table 1 Indoor object class names and IDs present in the collected indoor dataset

Class Name	Window	notice table	elevator	Door	electricity box	sign	light	trash can
Class ID	0	1	2	3	4	5	6	7
Class image								
Class Name	Stairs	security button	table	smoke detector	heating	fire extinguisher	light switch	Chair
Class ID	8	9	10	11	12	13	14	15
Class image								

images and videos, is a very important task as it supports the VIP understanding and interaction with the external world.

YOLOv3presents the best compromise between speed and accuracy for object detection [40], and makes YOLOv3 the best choice for this type of applications especially indoor navigation assistance to visually impaired persons. Indoor assistance navigation systems require (fast) real-time object detection as well as the high accuracy of the detection as the secure displacements should be targeted.

As the classic DCNN training requires a long time, the proposed system will use the transfer DCNN learning training technique [38] which uses less data. Indeed, transfer learning, a fast component in artificial intelligence and especially in deep learning field, is usually expanding in using deep CNN pretrained models.

This section provides an overview of the Darknet-53 used by YOLO v3 as a feature extractor (§ 4.1) followed by details of the proposed architecture which is used for indoor object detection (§ 4.2).

4.1 YOLO V3 backbone: Darknet-53

YOLO v3 presents a custom fully convolutional neural network named "Darknet-53" [40]. It makes use of residual blocks, of connections' skipping and of up-sampling and allows to detect fine-grained features in images. Darknet-53 originally presents 53 convolution layers trained on ImageNet [9]. Darknet-53 is mainly composed of 3 × 3 and 1 × 1 convolution layers

Table 2 Darknet-53 Architecture contents

Type	Filter size	Stride	Output size
Convolution	32 3 × 3	1	256 × 256
Convolution	64 3 × 3	2	128 × 128
1 x	32 1 × 1	1	128 × 128
Convolution	64 3 × 3	1	
Convolution			
Stride			
Convolution	128 3 × 3	2	64 × 64
2 x	64 1 × 1	1	64 × 64
Convolution	128 3 × 3	1	
Convolution			
Stride			
Convolution	256 3 × 3	2	32 × 32
8 x	128 1 × 1	1	32 × 32
Convolution	256 3 × 3	1	
Convolution			
Stride			
Convolution	512 3 × 3	2	16 × 16
8 x	256 1 × 1	1	16 × 16
Convolution	512 3 × 3	1	
Convolution			
Stride			
Convolution	1024 3 × 3	2	8 × 8
4 x	512 1 × 1	1	8 × 8
Convolution	1024 3 × 3	1	
Convolution			
Stride			

with skip-connections. Table 2 lists all processing layers of the Darknet-53, while Fig. 3 outlines the architecture of its residual blocks.

First, the top of the network uses a convolution layer with a 3×3 kernel size. It down-samples image size by using the strided convolution instead of using pooling layers. As mentioned in [2] using the strided convolution instead of pooling is more efficient in term of memory and temporal performances.

Darknet-53 deploys also a set of residual blocks where each block is composed of 3×3 and 1×1 convolution layers. Table 3 provides a comparison of performance of Darknet-53 and ResNet [18] in term of accuracy and BFLOP occupancy.

From Table 3 it can be easily found that the Darknet-53 implementation is more efficient than that of ResNet-152 [18] as it achieves 1457 BFLOP per second which makes it two times faster than ResNet-152 with the same accuracy.

4.2 YOLOv3 for object detection

YOLOv3 [40] is the 3rd version of the YOLO family [41, 42]. This version shows many temporal and accuracy improvements. YOLOv3 adopts an architecture based on two consecutive powerful Darknet-53 convolution layers whatleadsto106 convolution layers. Generally, the YOLO family models solve the detection problem as a regression problem.

Usually, objects in the images are of different size: small, medium and big (based on the object's size when compared to the image size). For indoor object detection it is important to detect all objects whatever is the object's size.

YOLOv3 [40] shows a better ability in detecting multi-scales objects. Indeed, the YOLOv3 adopted Features Pyramid Network (FPN-like) structures to detect objects with different scales. FPN algorithm encompasses two data movements: a **bottom-up** and a**top-down**(cf. Fig. 4).In bottom-up movement (image down sampling by 2) the semantic information (object characteristics) increases but the precision of their localization decreases; in top-down movement image up-sampling allows to increase the accuracy of the localization (using the information provided by the additional lateral connections and generated in bottom-up movement).

To perform the feature detection, the input image is subdivided into a grid of detection cells.

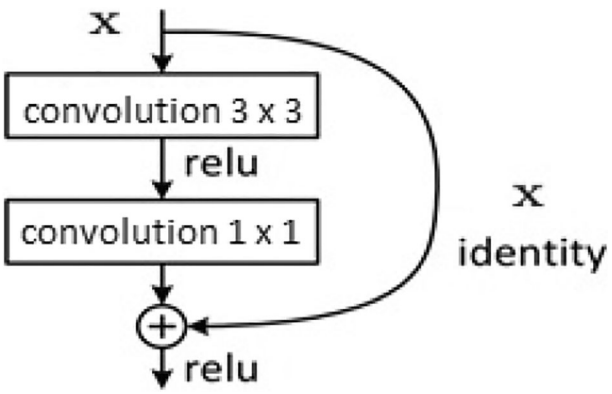


Fig. 3 Residual block architecture

Table 3 Performance comparison of Darknet-53 &ResNet backbones [40]

Backbone	Top-1Accuracy (%)	Top-51Accuracy (%)	Bn Ops	BFLOP/s	FPS
ResNet-101[18]	77.1	93.7	19.7	1039	53
ResNet-152[18]	77.6	93.8	29.4	1090	37
Darknet-53 [40]	77.2	93.8	18.7	1457	78

The multi-scale detection algorithm (top-down data movement) implements the following three steps (cf. Fig. 5):

- Step 1 (big object detection): Prediction (localization) of the features of big objects using the last feature map (of the top layer)
- Step 2 (medium object detection): merging the two corresponding feature maps of the same size (one generated in bottom-up movement with the one which is up-sampled by 2 and generated in the top-down data movement); Convolution to the merged feature map and prediction of the feature localizations for the objects of medium size.
- Step 3 (small object detection): up-sampling by 2 of the features maps of the convolution layer in step 2; concatenation of the feature maps of two ad hoc layers: one generated in bottom-up with the up-sampled map generated in top-down movement; convolution of the resulting feature map and prediction of small size objects' locations.

The prediction of an object localization (bounding box, localization) is performed by a convolution layer with the grid of a shape of 1×1 (B x $(4 + 1 + C)$) where 1×1 is the convolution layer, B is the number of rectangular bboxes that can be detected,"4" refers to the bbox attributes (tx,ty,tw,th), "1" is the object confidence for each grid cell and C presents the number of classes. In the proposed approach, 3 boxes for each grid cell and 16 indoor classes are used. Therefore, an output shape of 1×1 $(3 \times 5 + 16)$. Where 1×1 is the convolution layer, 3 is the small, medium and big object sizes, 5 refers to the four bbox attributes plus 1 as object confidence score and 16 is the number of object class.

For each extracted bbox YOLO v3attributes the objectness scores. The *objectness* score quantifies how likely an image window encompasses an object. The objectness score may be

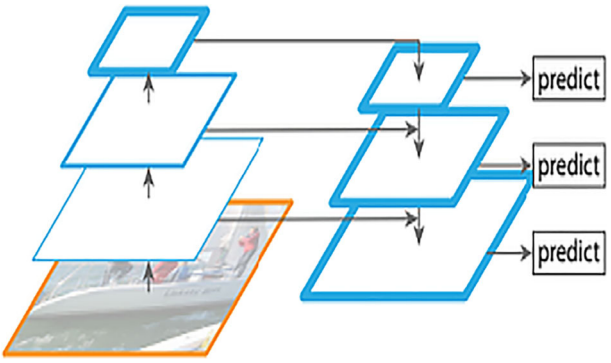


Fig. 4 FPN- like Structure used in YOLOv3 architecture [40]

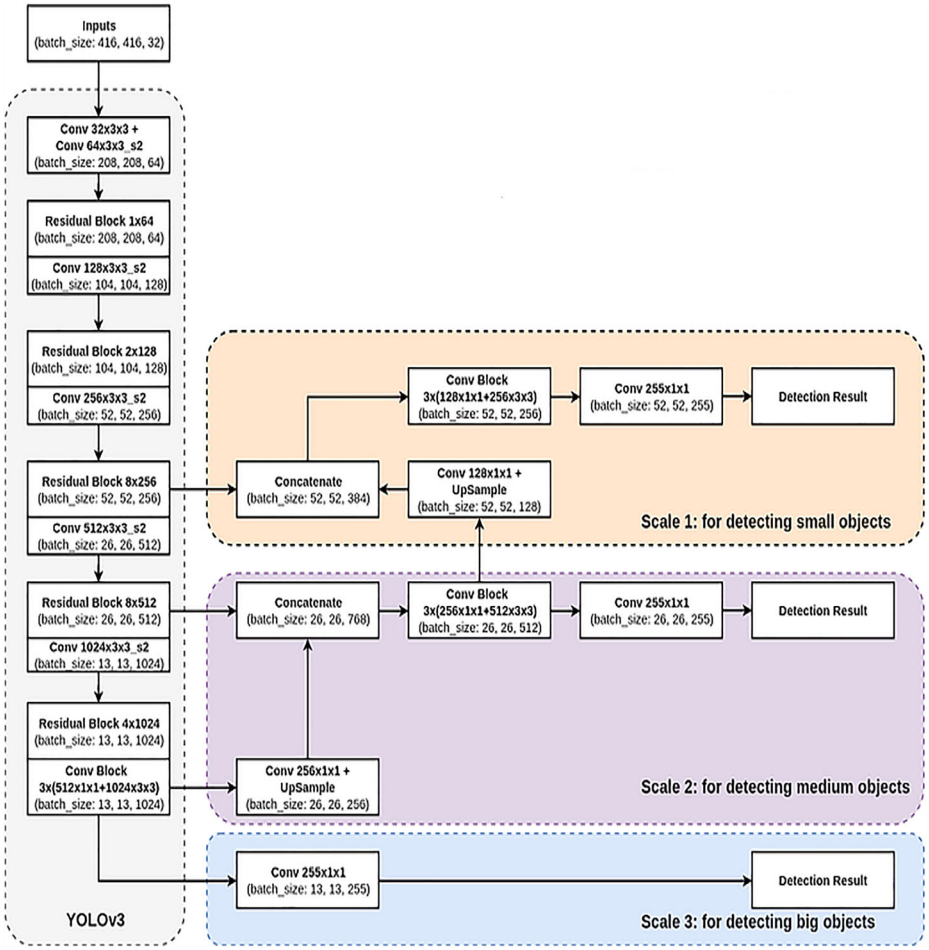


Fig. 5 YOLOv3 model simplified Architecture [23]

calculated using the independent logistic classifier; the usage of such classifier reduces the computation complexity of the processing.

A bbox is characterized by 4 coordinates (t_x,t_y,t_w,t_h)(cf. Fig. 6) which should be predicted: (t_x,t_y) are the image coordinates of the center of the bbox; t_w (resp. t_h) is width (resp. height) offset from the bbox center. Assuming (c_x,c_y) are the top left corner coordinates of a grid cell in a features map, the final predicted bbox parameters are b_x,b_y,b_h,b_w which are obtained by using the following equations:

$$b_x = \sigma(t_x) + c_x$$
$$b_y = \sigma(t_y) + y$$
$$b_w = p_w e^t_w$$
$$b_h = p_h e^t_h$$

Where:

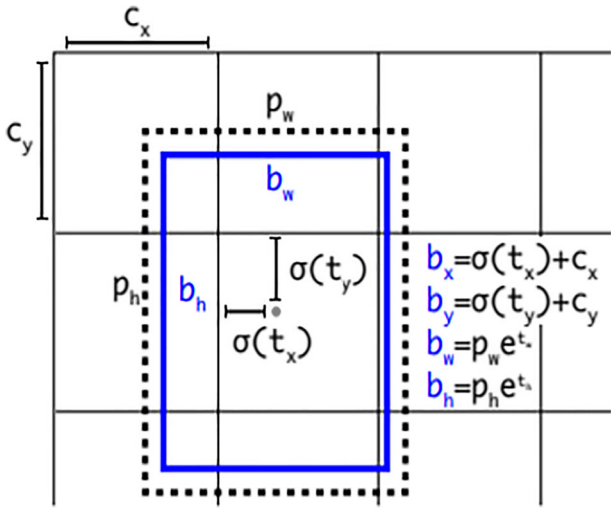


Fig. 6 Object detection approach based on the bounding box technique used in YOLOv3 [40]

- p_w, p_h are the anchor coordinates of the bbox' top-left corner (in a cell; 5 bbox can be predicted at each cell of the output feature map);
- σ is the sigmoid function $\sigma(x) = 1/(1 + e^{-x})$.

Since YOLO v 3 makes predictions at 3 different scales, and for each scale we have 3 or 5anchors it results in the use of 9 different anchor sizes.

For example, for an input image of 416 × 416 YOLOv3 predicts in 3 scales $((52 \times 52) + (26 \times 26) + (13 \times 13)) \times 3 = 10{,}647$ bboxes. This number is large. To reduce it:

- first bboxes are filtered considering the objectness scores (with a specific threshold);
- Secondly, the non-maximum (compared to the ground truth) are delayed (NMS).

5 Experiments and results

Indoor environment assistance navigation requires real-time object detection as well as height detection. Good accuracy and better speed (comparing to other DCNN models) makes YOLOv3 the best choice for real-time object detection for mobility assistive device design. This paper takes a step further to address the indoor object detection using DCNN. This is not only classifying objects but also providing the object localization in the current indoor scene. All these are experimentally tested.

The experiments on "indoor object detection and recognition" were implemented with the proposed indoor dataset Images in this dataset are taken in real interior environments. The indoor dataset collected consist of 16 indoor landmark object classes highly present in any indoor environment. The average precision of every indoor object present in the dataset is the quality criterion of the proposed approach.

This section presents the training experiments (§5.1) and test experiments with the proposed annotated dataset (§5.2).

5.1 Training experiments

Training a convolutional neural network requires a huge amount of data. For this purpose, we used the proposed indoor object detection dataset to feed the DCNN.

The training step consists in finding a set of rules to best classify objects. This process performs all tasks to train the indoor object classifier. During the training process, the pretrained model is evaluated on multiple indoor images with multiple points of view, different lighting conditions and complex backgrounds.

The proposed system runs on a HP workstation equipped with Intel Xeon E5–2683 v4 processor and Nvidia Quadro M4000 GPU with 8 GB of integrated memory.

Several steps were performed when training the DCNN model:

- First, network initialization with weights pretrained on COCO dataset [33];
- Second, the fine-tuning of the pretrained model on proposed collected dataset.

During the training step, the binary cross-entropy loss for class prediction was applied. 3 anchors are tested at each scale which gives a tensor of N x N * [3*(4 + 1 + 16)], where 4 is the bounding boxes offset, 1 is the objectness prediction and 16 is the number of classes and N x N is the grid dimension.

The proposed dataset was split into two subsets: one for training and the other for testing. At the beginning of the training step, images were resized to the resolution of input images (608 × 608).

For the training process, YOLO v3 uses the Stochastic Gradient Decent (SGD) [5] with momentum as an optimizer for the loss function. SGD updates parameters at each training step. But SGD performs updates with high variance which causes high oscillations of the objective function. These high fluctuations enable the loss functions to reach the local minimum. For this fact in YOLO v3 architecture, they use SGD with momentum. Momentum method enhances the SGD by reducing oscillations and speeding up the convergence process. SGD is performed as Eq. (1)

$$w = w - \eta * \nabla w * J\left(w; x^i; y^i\right) \tag{1}$$

Where w is models' parameters (weights + bias)
$\nabla w * J(w)$ is the objective function
η is the learning rate

Momentum adds an γ fraction of the updated vector of the previous step to the current updated vector. The momentum updates can be performed as Eq. (2).

$$V_t = \gamma V_{t-1} + \eta \nabla w * J(w) \tag{2}$$

$$w = w - V_t$$

As a result of adding the momentum method to the SGD, the model gains faster convergence with fewer oscillations. But, because of the accumulated speed, the momentum optimizer can miss the global or the minimum local.

In the proposed experiments we used SGD with momentum. To solve the problem caused by momentum optimizer, we propose to change it by the ADAM optimizer [28]. ADAM optimizer behaves like momentum in the parameters updating speed by keeping an exponentially decayed average of the past gradient m_t (the first momentum of the gradient). In addition, ADAM optimizer stores an exponentially decayed average of the past squared gradient V_T (the second momentum of the gradient). Moreover, it computes an adaptive learning rate for each parameter. It also updates the learning parameters at each training step.

$$m_t = \beta_1 * m_{t-1} + (1-\beta_1) * g_t \tag{3}$$

$$V_t = \beta_2 * V_{t-1} + (1-\beta_2) * g_t^2 \tag{4}$$

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 ($\widehat{m}_t$) and second ($\widehat{v}_t$) can be estimated as the following equations:

$$\widehat{m}_t = \frac{m_t}{1-\beta_1^t} \tag{5}$$

$$\widehat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{6}$$

Then, Adam updates the network parameters using the corrected first and second moment as Eq. (7).

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\widehat{v}_t + \epsilon}} * \widehat{m}_t \tag{7}$$

By using the ADAM optimizer to train the proposed detection system, we gained around 2% in the mean average precision (mAP).

We trained the proposed detection system by using two optimizers: the momentum and the ADAM. Table 4 reports the results in mAP obtained by using the two methods.

5.2 Test experiments

This section, describes the performed experiments and the obtained results for indoor object detection. The mean average precision (mAP) was selected to evaluate performances of the proposed detection system. The mAP is the precision average of all class queries present in the

Table 4 Comparison of mAP when using momentum and ADAM optimizers

Optimizer name	mAP (%)
Momentum	71.35
ADAM	73.19

Table 5 Average Precision (AP) results of different indoor objects classe

Class name	Window	Notice table	elevator	Door	Electricity box	Sign	light	Trash can
AP (%)	53.61	79.38	85.04	85.55	91.28	63.79	64.03	81.88
Class name	stairs	Security button	table	Smoke detector	Heating	Fire extinguish-er	Light switch	Chair
AP (%)	76.88	56.29	99.25	34.6	91.12	70.94	36.78	99.18

collected dataset. The average precision (AP)presents the value of the detection accuracy of a specific indoor object. The obtained detection performance on the considered test subset is summarized in Table 5.

The proposed object detection system achieves a mean precision of 73.19% (mAP). Almost perfect recognition was obtained for chair and table categories; good performances were obtained in the detection of many other indoor classes such as electricity box, heating, elevator, door, trash can.

The proposed detection system struggles in two indoor classes (smoke detector and light switch). For the rest of the indoor object classes our detection system achieves good performances.

To obtain information about the model's performances we have to calculate true positive (TP), false positive (FP) and false negative (FN) to calculate precision, recall and F1-score (Table 6).

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1{-}score = 2{*}\frac{Precision{*}Recall}{Precision + Recall} \tag{10}$$

As far as the VIP mobility is considered, the system should provide ahead data on upcoming object. It is relevant to assume that:

Table 6 Evaluation metrics used in the proposed detection system

Precision (mAP)	73.19%
True Positive	9995
False Positive	3823
False Negative	3558
Average IOU	55.64%
F1-score	0.71%
Recall	0.72%

Table 7 Comparison of results obtained by our method and those obtained in [10]

Indoor object name	Method in [10] (indoor AP %)	Method in [10] (indoor + FoV AP %)	Ours (proposed dataset AP %)
door	42.9	91.4	85.55
chair	72.6	94.0	99.18
table	46.2	91.6	99.25

- The optimal distance between an indoor object and a VIP sufficient to warn him/her in advance is about 5 m.

- The speed of the VIP is 1.4 m/s (speed of a normal person).

Consequently, the VIP will need 3.57 s to reach the indoor object. The temporal performance of the system should achieve a processing speed of 83 millisecond/frame, or 2 FPS. The proposed indoor object detection system achieves a processing speed of 12 FPS, therefore its match the needs of a VIP mobility.

As mentioned in Table 7, our proposed indoor object detection system achieves better results than the results obtained in [10] when using indoor dataset for the three classes. Also, our work outperforms [10] work when using (indoor+FoV) dataset. We achieved higher detection accuracies for chair and table classes. We note that we obtained better results despite we trained and tested our proposed system on challenging conditions including high inter and intra-class variation.

Figure 7 presents a detection example using images from the fully labeled indoor object detection and recognition dataset. The figure shows that all indoor objects present in the input image are detected in the considered image. Moreover, it can be observed that the door was detected despite of the fact that it was opened and it was taken with a challenging angle. We note that each of the two trash cans, very close one to one another, was detected by the proposed system. We note also that despite the small size of the smoke detector and the light switch in the input image, they were successfully detected by the proposed system.

Fig. 7 A Detection example of the proposed system

It is possible to conclude that the proposed indoor object detection system achieves high recognition rate of objects of different sizes and respect the real-time constraints required by the VIP mobility speed.

6 Conclusion

This paper presented a new indoor detection system designed for indoor assistance navigation for visually impaired people.

The proposed indoor dataset provides a data that can be used by researchers in computer vision field to develop new deep convolutional neural networks (DCNN), that can be included in many indoor robotic navigation systems, natural mobility of humanoid robotics, and in any system, which assists human being physical or virtual navigation.

The proposed indoor object detection and recognition (IODR) dataset present 8000 containing 16 landmark objects categories. Indoor image provided in this dataset presenting various challenging situations to make training and testing steps of the deep CNN robust for any complex situation given during inference process.

The proposed dataset provides images that are highly relevant for VIP mobility. The evaluation of the proposed system with the proposed new fully annotated lead to the detection precision of 73,19% mAP. This encouraging accuracy may be increased by adding more data during the DCNN model training.

A future work targets the system mAP improvement and integration of the proposed indoor detection system in embedded devices such as intelligent cane.

References

1. Afif M, Ayachi R, Said Y, Pissaloux E, Atri M (2019) A novel dataset for intelligent indoor object detection systems. Artificial Intelligence Advances, April 2019, vol.1 N°1 pp.52–58 (open-access)
2. Ayachi R, Afif M, Said Y et al (2018) Strided convolution instead of max pooling for memory efficiency of convolutional neural networks. Int. Conf. on the Sciences of Electronics, Technologies of Information and Telecommunications, Springer, Cham, pp 234–243
3. Bashiri, F. S., LaRose, E., Badger, J. C., D'Souza, R. M., Yu, Z., & Peissig, P. (November, 2018) Object detection to assist visually impaired people: a deep neural network adventure. Int.Symp. on Visual Computing, pp. 500–510, Springer, Cham
4. Bashiri FS, Larose E, Peissig P et al (2018) MCIndoor20000: A fully-labeled image dataset to advance indoor objects detection. Data in brief 17:71–75
5. Bottou L (2010) Large-scale machine learning with stochastic gradient descent, COMPSTAT'2010,Physica-Verlag HD pp. 177–186.
6. Chae, Hee-Won, Park, Chansoo, Yu, Hyejun, et al. (2016) Object recognition for SLAM in floor environments using a Depth Sensor. 13th Int.Conf. on Ubiquitous Robots and Ambient Intelligence (URAI), Xian, August, 19–22, 2016, pp. 405–410.
7. Chen Y, Chen R, Liu M, Xiao A, Wu D, Zhao S (2018) Indoor visual positioning aided by CNN-based image retrieval: training-free, 3D modeling-free. Sensors 18(8):2692
8. Couprie C, Farabet C, Najman L, Lecun Y (April 2013) Indoor semantic segmentation using depth information. In International Conference on Learning Representations (ICLR)
9. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database, IEEE CVPR (Computer Vision and Pattern Recognition), Florida, June 20–25, 2009, pp. 248–255.
10. 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, p. 1–6.

11. Eitel A, Springenberg JT, Spinello L, Riedmiller M, Burgard W (2015) Multimodal deep learning for robust RGB-D object recognition. IEEE/RSJ IROS, Hambourg, 28 September-02 October, 2015, pp. 681–687.
12. Escalona F, Rodríguez Á, Gomez-Donoso F, Martinez-Gomez J, & Cazorla M (july 2017) 3D object detection with deep learning. Journal of Physical Agents vol. 8, no. 1
13. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The Pascal visual object classes (voc) challenge. Int J of Computer Vision 88(2):303–338
14. Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2015). The pascal visual object classes challenge: a retrospective Int J Computer Vision, 111(1), 98–136.
15. Girchick R, Donahue J, Darrel T, et al. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE Conf. on computer vision and pattern recognition(CVPR), Ohio, June, 24–27, 2014, pp. 580–587.
16. Girshick R (2015) FastR-CNN, IEEE Int.Conf. on Computer Vision (ICCV), 11–18 December, 2015, Chili, pp. 1440–1448.
17. Guerrero LA, Vasquez F, Ochoa SF (2012) An indoor navigation system for the visually impaired. Sensors 12(6):8236–8258
18. He Kaiming, Zhang, Xiangyu, Ren, Shaoqing, et al. (2016) Deep residual learning for image recognition, IEEE CVPR, Nevada, 26 June-1 July, 2016, pp. 770–778.
19. He K, Gkioxari G, Dollár P, Girshick R (2017) MaskR-CNN, IEEE Int Conf on Computer Vision (ICCV), 22–29 October, 2017,Venise, pp. 2980–2988.
20. Henry P, Krainin M, Herbst E, Ren X, Fox D (2012) RGB-D mapping: using kinect-style depth cameras for dense 3D modeling of indoor environments. Int Journal of Robotics Research 31(5):647–663
21. http://www.navvis.lmt.ei.tum.de/dataset/ accessed: 21-07-2018
22. https://github.com/tzutalin/labelImg accessed: 23-08-2018
23. https://www.cyberailab.com/home/a-closer-look-at-yolov3; accessed: 26-08-2018
24. Hu H, Li Y, Zhu Z, et al. (2018) CNNAuth: continuous authentication via two-stream convolutional neural networks. In : 2018 IEEE international conference on networking, architecture and storage (NAS). IEEE, 2018. p. 1–9.
25. 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
26. Kendall A, Grimes M, Cipolia R (2015) Posenet: A convolutional network for real-time 6-dof Camera Relocalization,*IEEE ICCV*, December, 7–13, 2015. Washington, pp. 2938–2946.
27. Kim DK, Chen T (2015) Deep neural network for real-time autonomous indoor navigation. arXiv preprint arXiv:1511.04668
28. Kingma DP, Jimmy BA. (2014) Adam: A Method for Stochastic Optimization, arXiv preprint arXiv: 1412.6980
29. Krizhevsky A, Sutskever I, and Hinton GE (2012) Imagenet classifcation with deep convolutional neural networks,26th Annual Conf. on Neural Information Processing Systems (NIPS '12), Nevada, December, 3–6, 2012, pp. 1097–1105.
30. LeCun Y, Huang FJ, Bottou L (2004) Learning methods for generic object recognition with invariance to pose and lighting. IEEE CVPR, Washington 27 june-2 July, 2004 2:97–104
31. Li G, Zhang L, Sun Y et al (2019) Towards the sEMG hand: internet of things sensors and haptic feedback application. Multimed Tools Appl 78(21):29765–29782
32. Li Y, Hu H, Zhu Z et al (May 2020) SCANet: sensor-based continuous authentication with two-stream convolutional neural networks. ACM Transactions on Sensor Networks (TOSN) 16(3) article no. 29:1–26. https://doi.org/10.1145/3397179
33. Lin T-Y, Michael ME, Belongie S, et al. (2014) Microsoft coco: common objects in context, European Conf. on Computer Vision (ECCV), Springer, Cham pp. 740–755.
34. Liu S, Tian G (2019) An indoor scene classification method for service robot based on CNN feature. J of Robotics
35. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, and Berg AC (2016) SSD: Single shot multibox detector, European Conf. on Computer Vision, 8–16 October, Amsterdam, pp. 21–37.
36. Ma R, Zhang, L, Li G, et al. (2020) Grasping force prediction based on sEMG signals. Alexandria Engineering Journal
37. Nan LL, Xie K, Sharf A (2012) A search-classify approach for cluttered indoor scene understanding. ACM Trans. on Graphics 31(6):Article no. 137
38. Pan SJ (2009) Et Yang, Qiang. A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
39. Quattoni A, Torralba A (2009) Recognizing indoor scenes. IEEE CVPR, Miami, June 20–25. 2009. p. 413–420.
40. Redmon J, Farhadi A (1804). Yolov3: An Incremental Improvement," CoRR, vol. abs/1804.02767, 2018.

41. Redmon J, Farhadi, A (2017) YOLO9000: better, faster, stronger, IEEE CVPR, 21–26 July 2017. Hawaii, pp. 7263–7271.
42. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection, IEEE CVPR (Conf. on Computer Vision and Pattern Recognition), 26 June-1 July 2016, Nevada, (pp. 779–788).
43. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE PAMI 39(6):1137–1149
44. Reza, M. A., &Kosecka, J. (2014) Object recognition and segmentation in indoor scenes from RGB-D images, Robotics Science and Systems (RSS) Conference-5th workshop on RGB-D: Advanced Reasoning with Depth Cameras, Berkeley, 12 July, 2014.
45. Shao W, Luo H, Zhao F, Ma Y, Zhao Z, Crivello A (2018) Indoor positioning based on fingerprint-image and deep learning. IEEE Access 6:74699–74712
46. Verschae, Rodrigo, Ruiz-del-solar, Javier. Object detection: current and future directions, Frontiers in Robotics and AI, 2015, vol. 2, Article no 29.
47. Xiao J, Hays J, Ehinger KA, et al. (2010) Sun database: large-scale scene recognition from Abbey to Zoo, *IEEE* San Fransisco, June 13–18, 2010, pp. 3485–3492.
48. Yeboah Y, Yanguang C, Wu W, He S (2018) Autonomous indoor robot navigation via siamese deep convolutional neural network. ACMInt. Conf. on Artificial Intelligence and Pattern Recognition, China, August 18–20, 2018, pp. 113–119
49. Zhou B, Lapedriza A, Xiao J et al. (2014) Learning deep features for scene recognition using places database. Int Conf on Neural Information Processing Systems, Quebec, December 08–13, 2014, pp. 487–495,