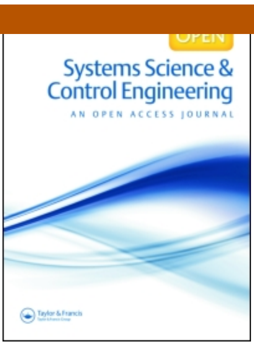

Taylor & Francis Group

Systems Science & Control Engineering

An Open Access Journal

ISSN: (Print) 2164-2583 (Online) Journal homepage: http://www.tandfonline.com/loi/tssc20

Prior knowledge-based deep learning method for indoor object recognition and application

Xintao Ding, Yonglong Luo, Qingde Li, Yongqiang Cheng, Guorong Cai, Robert Munnoch, Dongfei Xue, Qingying Yu, Xiaoyao Zheng & Bing Wang

To cite this article: Xintao Ding, Yonglong Luo, Qingde Li, Yongqiang Cheng, Guorong Cai, Robert Munnoch, Dongfei Xue, Qingying Yu, Xiaoyao Zheng & Bing Wang (2018) Prior knowledge-based deep learning method for indoor object recognition and application, Systems Science & Control Engineering, 6:1, 249-257, DOI: 10.1080/21642583.2018.1482477

To link to this article: https://doi.org/10.1080/21642583.2018.1482477

SYSTEMS SCIENCE & CONTROL ENGINEERING: AN OPEN ACCESS JOURNAL
2018, VOL. 6, NO. 1, 249–257
https://doi.org/10.1080/21642583.2018.1482477

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Prior knowledge-based deep learning method for indoor object recognition and application

Xintao Ding[a,b], Yonglong Luo[a,b], Qingde Li[c], Yongqiang Cheng[c], Guorong Cai[d], Robert Munnoch[c], Dongfei Xue[c], Qingying Yu[a,b], Xiaoyao Zheng[a,b] and Bing Wang[c]

[a]School of Computer and Information, Anhui Normal University, Wuhu, China; [b]Anhui Province Key Laboratory of Network and Information Security, Wuhu, China; [c]Department of Computer Science, Hull University, Hull, UK; [d]Computer Engineering College, Jimei University, Xiamen, China

ABSTRACT
Indoor object recognition is a key task for indoor navigation by mobile robots. Although previous work has produced impressive results in recognizing known and familiar objects, the research of indoor object recognition for robot is still insufficient. In order to improve the detection precision, our study proposed a prior knowledge-based deep learning method aimed to enable the robot to recognize indoor objects on sight. First, we integrate the public Indoor dataset and the private frames of videos (FoVs) dataset to train a convolutional neural network (CNN). Second, mean images, which are used as a type of colour knowledge, are generated for all the classes in the Indoor dataset. The distance between every mean image and the input image produces the class weight vector. Scene knowledge, which consists of frequencies of occurrence of objects in the scene, is then employed as another prior knowledge to determine the scene weight. Finally, when a detection request is launched, the two vectors together with a vector of classification probability instigated by the deep model are multiplied to produce a decision vector for classification. Experiments show that detection precision can be improved by employing the prior colour and scene knowledge. In addition, we applied the method to object recognition in a video. The results showed potential application of the method for robot vision.

ARTICLE HISTORY
Received 23 February 2018
Accepted 26 May 2018

KEYWORDS
Indoor object recognition; deep learning; prior knowledge

1. Introduction

The detection and recognition of indoor objects is an essential task in robot vision. Real-time, highly accurate indoor object recognition can greatly assist in robot navigation and manipulation (Khan, Hayat, Bennamoun, Togneri, & Sohel, 2016). In fact, many tasks associated with robot navigation depend directly on the recognition of indoor objects (Collet, Berenson, Srinivasa, & Ferguson, 2009; Ramisa, Alenyà, Moreno-Noguer, & Torras, 2014; Srinivasa et al., 2010). To enhance the robot's performance during indoor navigation, it is therefore necessary to design a reliable recognition method.

Classical studies on indoor object recognition mainly relied on machine learning techniques (Ding et al., 2016, 2017; Mei, Yang, & Yin, 2017; Nan, Xie, & Sharf, 2012; Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007; Uijlings, van de Sande, Gevers, & Smeulders, 2013). However, these methods involve a complex pipeline design and cannot learn deep features to generalize their extension. Other studies focussed on the design of statistical models to understand indoor geometry (Espinace, Kollar, Roy, & Soto, 2013; Pero et al., 2012; Wang, Gould, & Roller, 2013). However, these models lack sufficient precision (Pero et al., 2012; Wang et al., 2013).

Because RGB-D sensors, such as Kinect, provide not only colour but also depth information in scenes, RGB-D cameras are being widely used to guide indoor robot navigation (Husain, Schulz, Dellen, Torras, & Behnke, 2017). Jiang, Koch, and Zell (2016) developed a real-time recognition system for fruit and small-textured objects for a mobile robot equipped with the Kinect RGB-D sensor. Other studies also contributed the design of RGB-D descriptors for object recognition. Blum, Springenberg, Wülfing, and Riedmiller (2012) proposed a convolutional *k*-means descriptor for object recognition in RGB-D data. Chae, Park, Yu, and Song (2016) proposed a way to recognize objects for simultaneous localisation and mapping (SLAM) based on an object-level descriptor using a depth

CONTACT Xintao Ding xintaoding@163.com

Supplemental data for this article can be accessed here. https://doi.org/10.1080/21642583.2018.1482477

sensor. Bo, Ren, and Fox (2013) proposed an unsupervised feature learning method for RGB-D data, and the features were employed for object recognition using linear support vector machines. By using RGB-D data, Asif, Bennamoun, and Sohel (2017) employed convolutional neural networks (CNNs) to extract features for object recognition and grasp detection. Although depth information contained in the RGB-D data can produce more robust results, relevant techniques are usually more complex and computationally expensive. Furthermore, because depth information is generally captured by infrared lasers, RGB-D implementation involves a process of multimode optimization. For the sake of brevity, we focus on object recognition within the scope of the RGB mode.

Recently, deep learning has gained increasing attention in the area of computer vision. It has been employed to undertake many computer vision tasks, such as recognition (Eitel, Springenberg, Spinello, Riedmiller, & Burgard, 2015; Neverova, Wolf, Taylor, & Nebout, 2014; Schwarz, Schulz, & Behnke, 2015; Zhang et al., 2016), detection (Bianco, Celona, & Schettini, 2016; Gupta, Girshick, Arbeláez, & Malik, 2014), and image segmentation (Couprie, Farabet, Najman, & Lecun, 2013; Gupta et al., 2014). The task of indoor recognition may be divided into two main types: scene recognition and object recognition. In this study, we focus on indoor object recognition. However, the practice of extending existing deep models to indoor objects is still in its infancy, partially owing to the insufficiency of training datasets.

In order to improve detection precision, our study carries out indoor object recognition using a prior knowledge-based deep learning method, which learns deep features using annotated objects and predicts unknown objects using the features. Generally, public deep learning datasets (e.g. ImageNet (Schwarz et al., 2015), Chalearn's Looking at People dataset (Neverova et al., 2014), and Washington RGB-D Object (Eitel et al., 2015)) or private datasets (e.g. MIT campus buildings (Zhang et al., 2016)) are employed for training indoor objects. In this study, we first combine the public Indoor dataset (Quattoni & Torralba, 2009) and the private frames of videos (FoVs) dataset to train a CNN model, because the integration datasets are in favour of improving the detection precision (Ding et al., 2017). Because object colour may be helpful for object recognition, we second employ colour as a type of prior knowledge to enhance the detection precision of the resulting deep model. In addition, due to particular objects having a tendency to occur in certain scenes, we then employ scene as another type of prior knowledge to enhance the detection precision of the model.

The remainder of this paper is structured as follows. Section 2 describes our proposed method. Training and experiments are presented in Section 3 and Section 4, respectively. Section 5 focuses on an application of the method for robot vision. Finally, some concluding remarks follow in Section 6.

2. Proposed method

Figure 1 shows the architecture of the method proposed in this paper. For the implementation of indoor object recognition, we propose deep features involving colour knowledge and scene knowledge for recognition

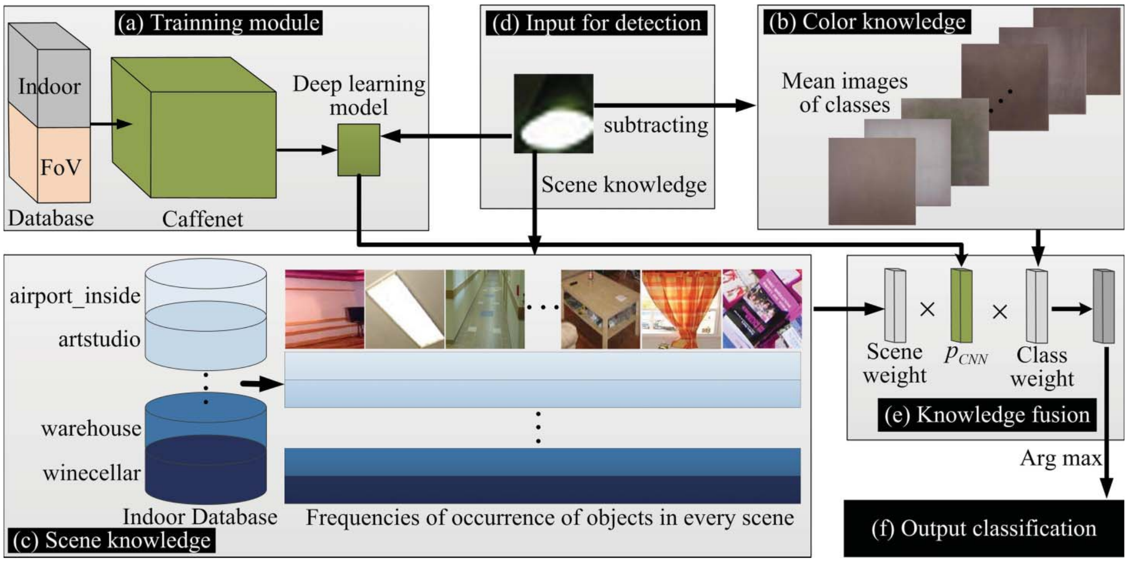

Figure 1. Outline of the proposed method. (a) CNN training module. (b) Colour knowledge. (c) Scene knowledge. (d) Input for detection. (e) Knowledge fusion. (f) Output classification.

(Figure 1). After combining the public Indoor dataset (Quattoni & Torralba, 2009) and the private FoVs dataset, we first train a CNN model (Figure 1a). Mean images, which are used for colour knowledge, are then generated for all the corresponding classes in the Indoor dataset (Figure 1b). After that, scene knowledge, which consists of frequencies of occurrence of objects in the scene, is employed as another type of prior knowledge (Figure 1c). When a detection request is launched, as shown in Figure 1d, the input image is first forwarded to the deep learning model to produce a vector of classification probability p_{CNN}. Second, the input image subtracts every mean image in Figure 1b to produce a class weight vector. Similarly, its scene knowledge is used to produce a scene weight vector. After that, the three vectors are multiplied to produce a decision vector, as shown in Figure 1e. Finally, the output classification of the input image is the index with the maximum value in the decision vector (Figure 1f).

2.1. Convolutional neural network

In order to implement recognition of indoor objects, we employ a CNN to train a deep model for classification. In detail, we use CaffeNet as our reference implementation, as shown in Figure 2. The images used for training consist of the public Indoor dataset (Quattoni & Torralba, 2009) and the private FoVs dataset. They were scaled to 256 × 256 pixels without regard for their original width and height ratio, since Caffenet requires input images of this size. Every private video was recorded from the surroundings of an object. The Indoor dataset contains 481 categories, while the number of annotations among categories varies. There are over 300 categories containing no more than 100 objects. Because a category with small object numbers cannot be used to train a deep model, and the number of samples used for training must be greater than the number of parameters, we rebuilt Caffenet by designing the size of the full connection layers, i.e. fc6 and fc7, to be 2048 (Figure 2).

2.2. Class weighting

Because object colour may be helpful for object recognition, we employ colour as a type of prior knowledge to enhance the hit rate of the resulting deep model. Let D be the Indoor dataset (Quattoni & Torralba, 2009), let D_k, $(k = 1, 2, \cdots, K)$ be the k-th class in D. Let $I_i^{(k)}$ be the i-th image in D_k, i.e. $I_i^{(k)} \in D_k \subset D$. Let the mean image of D_k be MI_{D_k}. Then, MI_{D_k} is given by equation (1):

$$MI_{D_k} = \frac{\sum_{I_i \in D_k} (I_i)_{M\times N}}{\mathrm{card}(D_k)} \tag{1}$$

where M and N are the number of rows and the number of columns of I, respectively; $\mathrm{card}(D_k)$ is the number of elements in D_k.

An input image I is compared with MI_{D_k} to produce class distance, as shown in equation (2):

$$d_k = \sum_{y=1}^{M} \sum_{x=1}^{N} |I(x,y) - MI_{D_k}(x,y)| \tag{2}$$

where $I(x,y)$ is the intensity of I at a pixel lying at the x-column and y-row. Similarly, we have $MI_{D_k}(x,y)$.

The class weight is then defined as in equation (3).

$$w_c = \frac{(d_1, d_2, \cdots, d_K)^T}{255MN} \tag{3}$$

2.3. Scene weighting

Generally, particular objects are found in certain scenes, such as a bed is usually found in a bedroom. Therefore, we also employ scene as another type of prior knowledge to help the deep model in decision-making. Let S_l, $(l = 1, 2, \cdots, L)$ be the l-th scene in D. The scene weight vector of the l-th scene f_l is defined as in equation (4).

$$f_l = (f_{l,1}, f_{l,2}, \cdots, f_{l,K})^T \tag{4}$$

where $f_{l,k}$ is the scene weight of the k-th class in the l-th scene, and it may be calculated as equation (5).

$$f_{l,k} = \frac{\mathrm{card}(I_i \in S_l \cap D_k)}{\mathrm{card}(S_l)} \tag{5}$$

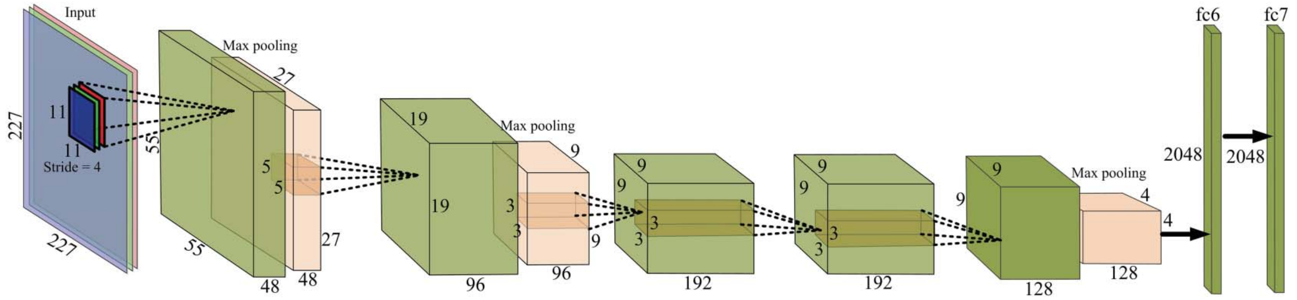

Figure 2. Illustration of the architecture of our Caffenet.

2.4. Knowledge fusion

When a detection request is launched, colour and scene knowledge are fused to help detection. The detection image is first forwarded to the deep model to produce a vector of classification probability p_{CNN}. Then, its colour and scene weights are generated from (3) and (4), respectively. The output classification of image I, which comes from the l-th scene, is the index for which the decision vector is at a maximum, as shown in equation (6).

$$c = \arg\max_{i} w_c \circ f_l \circ p_{CNN} \qquad (6)$$

where $w_c \circ f_l$ is the Hadamard product of w_c and f_l, which is defined as $(w_c \circ f_l)_i = w_c(i) f_l(i)$.

2.5. Time analysis

For our knowledge-based method, there is no additional work involved in the training stage. During the stage of detection, the input image is subtracted by every mean colour image, which is in size of 256 × 256. Therefore there are K × 256 × 256 subtractions for using colour knowledge. In order to use its scene knowledge, a L loop is required to index the input scene knowledge. During the knowledge fusion step, two Hadamard productions tailed by the probability vector normalization are required. The operations for this step total to 6 K. In all, additional K × 256 × 256 + L + 6 K operations are required compared with the non-knowledge method.

3. Training the deep model

Our pre-training module runs on a dual-core i-3 4160 CPU with 16GB RAM equipped with the NVIDIA GeForce GTX 1080 graphics card with 8 GB memory. Caffenet is compiled under Ubuntu 16.04 with CUDA Toolkit 8.0, cuDNN 5.1 library, Anaconda3, and OpenCV 3.1. The protocol buffer version employed for Python 3.5 is 3.0.0.

We use the Indoor dataset (Quattoni & Torralba, 2009) and the FoVs for our CNN model. The experimental dataset consists of indoor objects belonging to 21 different classes. Because a category with a small number of objects cannot be used to train a deep model, an Indoor dataset class was retained if it contained more than 500 members. Using these parameters, we obtained 18 object categories, as shown in Table 1. In order to take personalized objects into account, we extended the resulting dataset using the FoVs. The extension included 17 categories. Every category extension employed multiple videos, and each video was created from the surroundings of a particular object. The extended categories DP, screen, and TM are three new categories added to the categories from the Indoor dataset. The other fourteen categories are shared by both the Indoor dataset and the FoVs dataset. FoVs were appended to the corresponding Indoor categories. In all, Indoor and FoVs datasets are employed for Caffenet in this study (Table 1). For convenience, the four categories that are not extended by FoVs are marked with asterisks.

Table 1. Categories used for our CNN model.

Class index	Semantic class	Class index	Semantic class	Class index	Semantic class
0	Wall	7	Picture	14	Bottle
1	Lamp	8	Plant	15	Table
2	Floor	9*	Chair top	16	Curtain
3	Window	10*	Painting	17*	Book
4	Ceiling	11*	Chair occluded	18	Desk partition
5	Chair	12	Door	19	Screen
6	Books	13	Pillow	20	TV monitor

The class indexes 0–17 are those categories with more than 500 members from the Indoor dataset. Class indexes 18–20 are extended personal categories using FoVs. The semantic classes marked with asterisks are the four categories not extended by FoVs.

The whole dataset was divided into three subsets used for training, validation, and testing. The division was implemented using a loop choice in the ratio 2:1:1. The resulting element counts were 36,268, 18,134, and 18,133. After 450,000 iterations, the resulting accuracy of the Indoor+FoVs model was 0.9009. The time spent on training was 27.5 h.

4. Experiments

4.1. Test experiments

In this section, we describe the test experiments implemented against the test subset. After parsing all the object images from the Indoor dataset, all images annotated with the same object were placed in a folder. The mean images were generated by a folder scan using equation (1). Figure 3 shows the mean images of classes 0–17 in the Indoor dataset.

In this study, all the annotation files of the Indoor dataset were scanned to count the occurrence of the 18 classes. The images from the Indoor dataset were placed in 67 different folders. When the object images of the Indoor dataset were parsed, we counted their scene occurrences to find the scene weight. Figure 4 shows the resulting scene weight, in which the scenes were sorted by their names in alphabetical order. The scene with tag 0 represented 'airport_inside' in the Indoor dataset.

Figures 3 and 4 present heterogeneity, which may be helpful for object recognition. The mean image (*MI*) of plant, i.e. class 8, in Figure 3 presents green. The *MI* of painting, i.e. class 10, in Figure 3 shows four borders like a frame around it. Class 1 presents high intensity in its centre. It is lamp. On the contrary, classes 5 and 11 show

Figure 3. The mean image (*MI*) of 18 classes in the Indoor dataset.

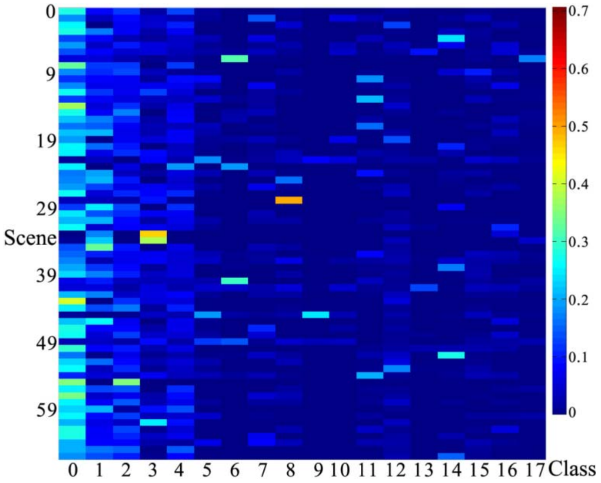

Figure 4. Heat map of scene knowledge of all classes found in the Indoor dataset.

low intensity. In Figure 4, certain classes in a certain scene present high weights. Overall, the deep model may take advantage of prior knowledge, such as colour and scene, during detection.

After the acquisition of prior knowledge, we implement detection using equation (6). In order to evaluate our approach, we use top-1 precision and mean average precision (mAP) as two measures of performance, where mAP is the average of all the precisions obtained for all queries. Together with mAP, the resulting classes' top-1 precisions on the test subset are shown in Table 2.

Compared with Caffenet, the prior knowledge-based method achieves a better result with a mAP of 86.3%. The increase in the detection precision for wall, books, painting, chair occluded, and book categories are remarkable. Although the resulting precisions for ceiling, door, table, and curtain categories are inferior, the small differences show the comparability of the method. Results demonstrate that detection precision can be improved by employing colour and scene knowledge.

Table 3 shows some top-3 classifications together with their probabilities. The semantic classes marked with asterisks are top-1 classification. (F|T) reveals the example, in which CaffeNet results in a false classification whereas our proposed method results in a correct classification under top-1 classification. Similarly, we have (F|F) and (T|F).

It is unavoidable for top-1 classification to incur misclassification due to samples that are difficult to categorize, such as chair, wall, picture, and ceiling in Table 3. However, misclassification may be reduced if top-3 evaluation is implemented, as shown in Table 3. For CaffeNet and our proposed method, both chair and ceiling are correctly classified under top-3 evaluation. The wall

Table 2. Top-1 precision (%) on test dataset. mAP is the abbreviation of mean average precision.

Class index	0	1	2	3	4	5	6	7	8	9*	10*
Caffenet	85.9	90.8	92.2	88.8	**90.1**	93.6	89.9	92.3	95.9	74.7	45.2
Proposed method	**90.0**	**92.3**	92.2	**91.1**	90.0	**94.4**	**93.2**	**93.8**	**96.1**	**80.7**	**51.0**
Class index	**11***	12	13	14	15	16	**17***	18	19	20	mAP
Caffenet	38.6	**92.7**	93.4	92.9	**91.5**	**91.3**	27.9	1.00	1.00	1.00	84.2
Proposed method	**50.3**	92.5	**95.3**	**94.0**	90.8	90.3	**33.3**	1.00	1.00	1.00	86.3

Table 3. Top-3 classifications together with their probabilities.

Object		True class	CaffeNet	Proposed method
(F|T)		ceiling	(window*, ceiling, books) (0.48, 0.30, 0.12)	(ceiling*, window, wall) (0.69, 0.18, 0.11)
(F|T)		chair	(bottle*, chair, plant) (0.64, 0.32, 0.01)	(chair*, lamp, plant) (0.89, 0.06, 0.02)
(F|F)		wall	(ceiling*, floor, lamp) (0.53, 0.41, 0.02)	(ceiling*, floor, wall) (0.53, 0.45, 0.02)
(F|F)		picture	(books*, chair, painting) (0.55, 0.33, 0.05)	(chair*, painting, window) (0.36, 0.26, 0.13)
(T|F)		ceiling	(ceiling*, wall, table) (0.65, 0.35, 0.00)	(wall*, ceiling, table) (0.66, 0.34, 0.00)
(T|F)		painting	(painting*, floor, lamp) (0.30, 0.24, 0.09)	(floor*, painting, picture) (0.41, 0.37, 0.10)

Table 4. Timing (s) on a dual-core i-3 4160 CPU equipped with GTX 1080 GPU. ATpS is the abbreviation for average time per sample.

Method	Time (s)	ATpS (s)
CaffeNet	1862.2	0.10
Our method	2471.8	0.13

Table 5. Comparison experiments for NYU v2 on 7 classes.

Method	Wall	Floor	Win	Ceiling	Books	Pic	Table
Couprie et al.	89.4	68.0	37.8	33.2	31.7	38.5	18.0
Hermans et al.	71.8	91.5	46.1	83.4	45.4	35.8	27.7
Caffenet	83.2	66.2	47.0	77.8	13.5	12.8	15.8
Proposed method	86.7	73.1	46.8	73.0	14.6	19.4	15.4

is correctly detected by our proposed method. Neither CaffeNet nor knowledge-based method is able to correctly detect picture in (F|F). However, painting, which occurs in top-3 classification, is a close classification of the object picture. It can be inferred that top-3 accuracy may alleviate misclassification.

In Table 4 we summarize the running time of the entire object detection on test dataset using python. Since the total number of samples in test dataset is 18,133, the test results reveal that our proposed method requires approximately 30 ms for every input.

4.2. Comparison experiments

In this section, we describe the comparison experiments implemented on the NYU v2 dataset. The dataset consists of a total of 1449 samples of different indoor scenes. We parsed seven object classes from images in the labelled dataset based on annotated labels. After resizing the parsed patches to 256 × 256 pixels, they are inputted into the proposed model for recognition. Couprie et al., 2013 applied a multiscale convolutional network to learn features combining the images and its depth information. Hermans, Floros, and Leibe (2014) proposed a fast 2D semantic segmentation approach based on a novel 2D-3D label transfer method. Table 5 shows our individual labelling top-1 precisions compared with the Couprie et al. (2013), Hermans et al. (2014), and Caffenet methods.

Although Couprie et al. (2013) and Hermans et al. (2014) result in higher accuracies, the results of Caffenet and the proposed method shown in Table 5 are transfer results from the Indoor dataset to the NYU dataset. The average accuracies of Caffenet and the proposed method are 45.2 and 47.0, respectively. Compared with Caffenet, our proposed method achieves a better result overall.

5. Application

In this section, we present an application of our proposed method for robot vision. A video was created using a camera to test indoor object recognition. The scene is a typical indoor office environment. A video of the room was captured over a span of 31 s. The video is then parsed into 940 frame images. The resolution of the FoVs is 1280 × 720.

Figure 5 illustrates the overall structure of the application presented in this paper. To implement indoor object recognition, we parse the input video into frame images (Figure 5a). Then, the regions of interest (RoIs) are extracted using a selective search (Uijlings et al., 2013) (Figure 5b). These RoIs are then resized to 256 × 256 pixels and classified into candidates using the proposed method (Figure 5c). Those candidates that are in the same category and show an overlap greater than 0.5 between the nearest frames are fused into one classification (Figure 5d). Finally, the frames annotated with bounding boxes are concatenated into video as the output (Figure 5e). The parameter *k*, which controls the size of segments in the initial segmentation, is set to 200 in this study. The number of RoIs extracted from FoVs ranges from 45 to 217. The detection was implemented offline. The detection result was resized into a 256 × 256 image and forwarded to our proposed model for classification.

Although our model (Figure 5c) may predict a classification for every RoI, misclassification is unavoidable. In order to reduce misclassification, detection fusion is employed in our design. We fused candidates that were derived from the deep model between nearest frames when they were classified in the same category and their overlap was greater than 0.5. Figure 5d shows object fusion in two frames. The top-left object shown by a black line is indexed with category 12. However, the top-left object shown by a red line is classified into category 15. As the two candidates are classified in different categories, they are misclassified. On the contrary, both the second candidates shown in the two frames are classified into category 1, and their overlap area is greater than 0.5. They are annotated in a fused state and the annotation box is the minimum coverage of the two candidates. After the decision vector is normalized to a unit vector, the frame of the box is coloured red, yellow, green, or blue to show the probability of the prediction if its probability is in the interval (0.9,1], (0.75,0.9], (0.5,0.75], or (0,0.5], respectively. Table 6 shows these intervals counting over all the prediction probabilities, which totals to 18,865. Table 6 shows that most of predictions hit their classifications with a probability more than 0.75. The low prediction probability that less than 0.5 is rare.

Table 6. Interval counting over all the prediction probabilities.

Interval	(0,0.5]	(0.5,0.75]	(0.75,0.9]	(0.9,1]
counting	9	5312	5765	7779
hit rate (%)	0.5	28.2	30.6	41.2

After all of the FoVs are assigned the aforementioned probability, we merge the frames into an annotated video at a frame rate of 6 fps (Supplement 1). For convenience, some detection results are shown in Figure 6. Figure 6a is the first output frame, in which the recognized object does not fuse with other objects. The detection results

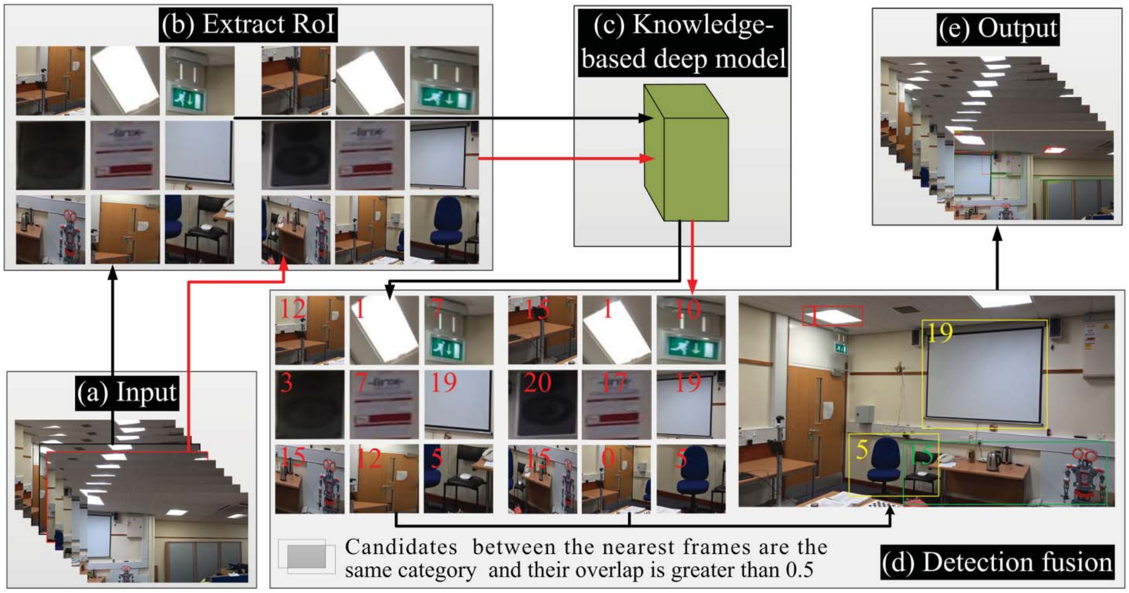
(b) Extract RoI
(c) Knowledge-based deep model
(e) Output
(a) Input
12
1
7
15
1
10
3
7
19
20
17
19
15
12
5
15
0
5
1
19
5
5
Candidates between the nearest frames are the same category and their overlap is greater than 0.5
(d) Detection fusion

Figure 5. Pipeline of application. (a) Input video. (b) Extraction of proposed regions of interest (RoIs). (c) Prior knowledge-based deep model. (d) Detection fusion. (e) Output video.

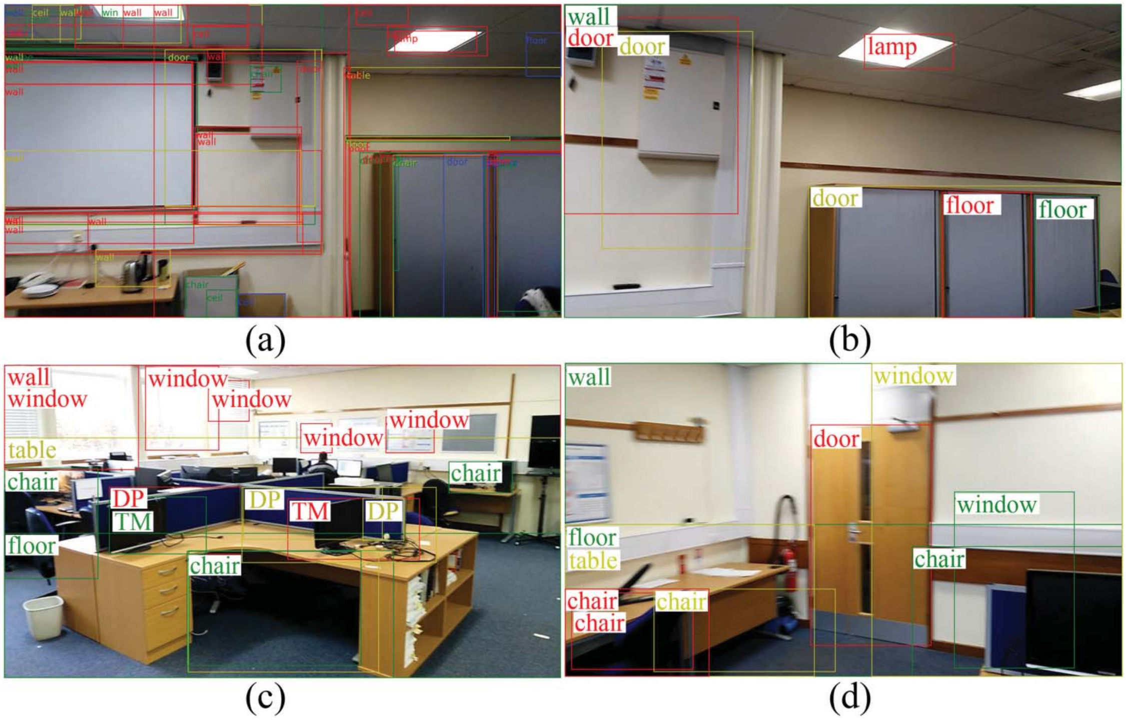

Figure 6. Some detection results. (a) The first output frame from our proposed model in which recognized objects do not fuse with others. (b), (c), and (d) are the detection results of the 70th, 494th, and 808th frames, respectively.

of the 70th, 494th, and 808th frames are shown in Figure 6b–d, respectively.

Our method may be applied to indoor object detection. Figure 6 shows that detection fusion is necessary in our pipeline. Although there are a lot of detections in Figure 6a before fusion, many of them are misclassified. Furthermore, although the door of the cupboard is misclassified to 'floor' (Figure 6b), the misclassification may be corrected when the cupboard goes through the video and comes to the centre of the scope. The video (Supplement 1) shows the door of the cupboard is detected frequently and correctly from the 40th frame to the 108th frame. The real windows, floor, table, DP, TM, chair, and door are almost all correctly detected in Figure 6c and d. In the detection video, the objects labelled 'window', 'desk partition', 'TV monitor', and 'chair' can be frequently and correctly detected most of the time when they are in the scope of the video.

The experimental results suggest that our proposed method may be applied to indoor object detection, and the use of prior knowledge is helpful in enhancing a robot vision for indoor object recognition.

6. Conclusion

In this paper, we proposed a prior knowledge-based deep model for indoor object recognition. In our design, prior knowledge of colour and scene was utilized to help the deep model to make a decision when a detection request is launched. Both the test and comparison experiments demonstrate that the knowledge-based method enhances the hit rate for object detection. Based on our proposed model, we implemented an application for an indoor video. The application experiments show that this method may be applied to robot vision. The main contribution of this work is three-fold. (a) This work contributes to indoor object recognition for robot vision. (b) Our study proposed a knowledge-based method. (c) The proposed method is in favour of improving detection precision. A potential future research project may concentrate on accelerating detection speed so that real-time detection may be implemented on a mobile robot.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Natural Science Foundation of Anhui Province under Grant [numbers 1808085MF171, 1708085MF145]; the National Natural Science Foundation of China under Grant [numbers 61672039, 61602009, and 61772034]; the University Natural Science Research Project of Anhui Province under Grant [number KJ2017A327].

References

Asif, U., Bennamoun, M., & Sohel, F. A. (2017). RGB-D object recognition and grasp detection using hierarchical cascaded forests. *IEEE Transactions on Robotics*, *33*(3), 547–564.

Bianco, S., Celona, L., & Schettini, R. (2016). Robust smile detection using convolutional neural networks. *Journal of Electronic Imaging*, *25*(6), 063002.

Blum, M., Springenberg, J. T., Wülfing, J., & Riedmiller, M. (2012). *A learned feature descriptor for object recognition in RGB-D data*. IEEE International Conference on Robotics and Automation, Minnesota, USA, 14-18 May 2012, pp. 1298–1303.

Bo, L., Ren, X., & Fox, D. (2013). Unsupervised feature learning for RGB-D based object recognition. In *Experimental robotics* (Vol. 88, pp. 387–402). Heidelberg: Springer.

Chae, H. W., Park, C., Yu, H., & Song, J. B. (2016). *Object recognition for SLAM in floor environments using a depth sensor*. International Conference on Ubiquitous Robots and Ambient Intelligence, Xian, China, 19-22 August 2016, pp. 405–410.

Collet, A., Berenson, D., Srinivasa, S. S., & Ferguson, D. (2009). *Object recognition and full pose registration from a single image for robotic manipulation*. IEEE International Conference on Robotics and Automation, Kobe, Japan, 12-17 May 2009, pp. 48–55.

Couprie, C., Farabet, C., Najman, L., & Lecun, Y. (2013). Indoor semantic segmentation using depth information. arXiv:1301.3572.

Ding, W., Gu, J., Tang, S., Shang, Z., Duodu, E. A., & Zheng, C. (2016). Development of a calibrating algorithm for delta robot's visual positioning based on artificial neural network. *Optik - International Journal for Light and Electron Optics*, *127*(20), 9095–9104.

Ding, X., Luo, Y., Yu, Q., Li, Q., Cheng, Y., Munnoch, R., . . . Cai, G. (2017). *Indoor object recognition using pre-trained convolutional neural network*. Automation and Computing (ICAC), 2017 23rd International Conference on, Huddersfield, UK, 7-8 September 2017, pp. 1–6.

Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M., & Burgard, W. (2015). *Multimodal deep learning for robust RGB-D object recognition*. IEEE International Conference on Intelligent Robots and Systems, Hamburg, Germany, 28 September-2 October 2015, pp. 681–687.

Espinace, P., Kollar, T., Roy, N., & Soto, A. (2013). Indoor scene recognition by a mobile robot through adaptive object detection. *Robotics and Autonomous Systems*, *61*(9), 932–947.

Gupta, S., Girshick, R., Arbeláez, P., & Malik, J. (2014). Learning rich features from RGB-D images for object detection and segmentation. In *Computer vision - ECCV 2014* (Vol. 8695, pp. 345–360). Cham: Springer.

Hermans, A., Floros, G., & Leibe, B. (2014). *Dense 3D semantic mapping of indoor scenes from RGB-D images*. IEEE International Conference on Robotics and Automation, Xian, China, 19-22 August 2014, pp. 2631–2638.

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.

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, pp. 859–871). Cham: Springer.

Khan, S. H., Hayat, M., Bennamoun, M., Togneri, R., & Sohel, F. A. (2016). A discriminative representation of convolutional features for indoor scene recognition. *IEEE Transactions on Image Processing*, *25*(7), 3372–3383.

Mei, S., Yang, H., & Yin, Z. (2017). Discriminative feature representation for image classification via multimodal multitask deep neural networks. *Journal of Electronic Imaging*, *26*(1), 013023.

Nan, L., Xie, K., & Sharf, A. (2012). A search-classify approach for cluttered indoor scene understanding. *ACM Transactions on Graphics*, *31*(6), article no. 137.

Neverova, N., Wolf, C., Taylor, G. W., & Nebout, F. (2014). Multiscale deep learning for gesture detection and localization. In *Computer vision - ECCV 2014 workshops* (Vol. 8925, pp. 474–490). Cham: Springer.

Pero, L. D., Bowdish, J., Fried, D., Kermgard, B., Hartley, E., & Barnard, K. (2012). *Bayesian geometric modeling of indoor scenes*. IEEE Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, 16-21 June 2012, pp. 2719–2726.

Quattoni, A., & Torralba, A. (2009). *Recognizing indoor scenes*. IEEE Conference on Computer Vision and Pattern Recognition, FL, USA, 20-25 June 2009, pp. 413–420.

Ramisa, A., Alenyà, G., Moreno-Noguer, F., & Torras, C. (2014). Learning RGB-D descriptors of garment parts for informed robot grasping. *Engineering Applications of Artificial Intelligence*, *35*, 246–258.

Schwarz, M., Schulz, H., & Behnke, S. (2015). *RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features*. IEEE International Conference on Robotics and Automation, WA, USA, 26-30 May 2015, pp. 1329–1335.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 411–426.

Srinivasa, S. S., Ferguson, D., Helfrich, C. J., Berenson, D., Collet, A., Diankov, R., . . . Weghe, M. V. (2010). HERB: A home exploring robotic butler. *Autonomous Robots*, *28*(1), 5–20.

Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., & Smeulders, A. W. M. (2013). Selective search for object recognition. *International Journal of Computer Vision*, *104*(2), 154–171.

Wang, H., Gould, S., & Roller, D. (2013). Discriminative learning with latent variables for cluttered indoor scene understanding. *Communications of the ACM*, *56*(4), 92–99.

Zhang, F., Duarte, F., Ma, R., Milioris, D., Lin, H., & Ratti, C. (2016). Indoor space recognition using deep convolutional neural network: a case study at MIT campus. arXiv:1610.02414.