

## Research Article

# Visual Semantic Navigation Based on Deep Learning for Indoor Mobile Robots

Li Wang,<sup>1</sup> Lijun Zhao ,<sup>1</sup> Guanglei Huo,<sup>2</sup> Ruifeng Li,<sup>1</sup> Zhenghua Hou,<sup>1</sup> Pan Luo,<sup>1</sup> Zhenye Sun,<sup>1</sup> Ke Wang,<sup>1</sup> and Chenguang Yang <sup>3</sup>

<sup>1</sup>State Key Laboratory of Robotics and System, Harbin Institute of Technology, Harbin 150001, China

<sup>2</sup>HNA Technology Group, Shanghai 200122, China

<sup>3</sup>Key Laboratory of Autonomous Systems and Networked Control, College of Automation Science and Engineering, South China University of Technology, Guangzhou 510640, China

Correspondence should be addressed to Lijun Zhao; zhaolj@hit.edu.cn

Received 14 July 2017; Accepted 11 February 2018; Published 22 April 2018

Academic Editor: Thierry Floquet

Copyright © 2018 Li Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to improve the environmental perception ability of mobile robots during semantic navigation, a three-layer perception framework based on transfer learning is proposed, including a place recognition model, a rotation region recognition model, and a “side” recognition model. The first model is used to recognize different regions in rooms and corridors, the second one is used to determine where the robot should be rotated, and the third one is used to decide the walking side of corridors or aisles in the room. Furthermore, the “side” recognition model can also correct the motion of robots in real time, according to which accurate arrival to the specific target is guaranteed. Moreover, semantic navigation is accomplished using only one sensor (a camera). Several experiments are conducted in a real indoor environment, demonstrating the effectiveness and robustness of the proposed perception framework.

## 1. Introduction

Enabling robots to navigate autonomously in a real world environment is a very challenging topic in the field of robotics associated closely with signal processing, machine vision, and so forth. A robot should have adaptive capacities of planning optimal paths in maps when implementing tasks [1]. Traditional navigation approaches strongly rely on metric or topological maps and constraints which are described in terms of geometry, assuming the shortest path to be the best [2–4]. However, human navigation does not depend on the “best,” but on what to be seen [5]. Semantic information can be further abstracted from images to decide where we go based on it. Normally, we can recognize rooms, corridors, doors, aisles, and so on for reference to plan the motion from one place of a room to another in a building. Moreover, we should also know the exact side within the scenario in order to keep moving on the right path. In other words, we can adjust back if we realize that we are walking in a skew direction. Therefore, mobile robots should

have the abilities mentioned above to perform human-like navigation.

Semantic navigation is regarded as a system considering semantic information to express the environment and then to implement the robot’s localization and navigation. In recent decades, a great deal of attempts have been made focusing on finding applicable solutions for robot semantic navigation. Semantic navigation approaches usually adopt topological structures [6–8], in which semantic places and objects are abstracted to different nodes. It is expected that each node is observed accurately during the motion. However, those nodes may not be observed straightforwardly via the motion offset of mobile robots. Moreover, humans’ navigation depends on their two eyes, which is the motivation behind equipping multiple sensors on mobile robots when dealing with the navigation task.

The main contribution of this paper is to propose a three-layer perception framework based on transfer learning using only visual information, including place recognition model, rotation region recognition model, and “side” recognition

model. Using this framework, semantic navigation can be achieved via only one camera and the motion offset of mobile robots can be solved. Different from traditional semantic navigation methods, the proposed algorithm uses transfer learning to train and recognize the semantic information in the environment and only uses one RGB camera to realize the whole semantic mapping and navigation. Through the recognition of input images, it can provide the robot with key semantic information for navigation, such as navigation in corridors and recognition of turning areas.

The rest of this work is organized as follows. After discussing some related work in Section 2.1, Section 2.2 discusses the details of the proposed three-layer perception framework. Section 3 shows some experimental results obtained by our approach. Finally, Section 4 concludes the paper.

## 2. Materials and Methods

*2.1. Related Work.* Semantic information has been used to infer the indoor environment information and to improve the planning efficiency [5, 9–11]. Also, it has drawn a deal of attention in the area of large-scale navigation, seeking to deal with problems in a higher dimension [12]. This type of navigation is inspired by humans, where places are not described in terms of a global map but by semantic information. Semantics in mobile robot navigation has been mainly used for place recognition, allowing mobile robots to build relationships based on places [13]. The topological structure is usually adopted for the semantic navigation, which allows robots to plan their paths at a high dimension [14, 15]. In topological structure, places are often abstracted to nodes, and visiting orders are abstracted to edges.

A variety of approaches are attempted to solve the semantic navigation problem in different perspectives; for instance, Joseph et al. [16] used a human motion mode to predict a path based on how real humans ambulated towards a goal while avoiding obstacles. Posada et al. [17] presented a semantic navigation approach which could be parsed directly from natural language (e.g., “enter or get out of the room, follow the corridor until the next door, etc.”). Zhao and Chen [18] encoded scene information, semantic context, and geometric context into a condition random field (CRF) model, which computed a simultaneous labeling of image regions into semantic classes and structural object classes. Horne et al. [19] used semantic labeling techniques to achieve path planning. In these systems, each pixel in images was classified automatically into a semantic class, and then an image was produced from the induced visual percepts that highlighted certain classes. Recently, neural networks based on learning have been widely used in robots [20]. The deep learning method has become a significant way to solve semantic navigation problems showing the powerful ability to obtain semantic information [21–23]. Zhu et al. [24] proposed a target-driven visual navigation method using a reinforcement learning model that generalizes across targets and scenes. Furuta et al. [25] proposed semantic map based navigation which consisted of generating a deep learning enabled semantic map from annotated world and object

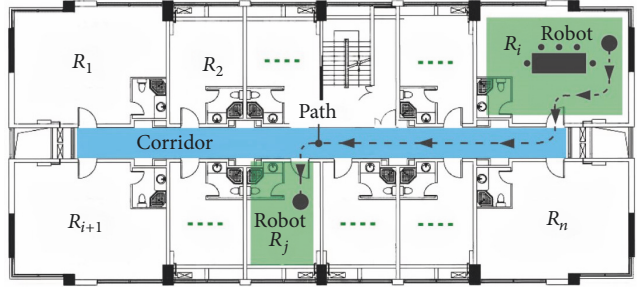


FIGURE 1: The diagram for mobile robot working in an indoor environment. There are several rooms and a corridor in the diagram. A trajectory with a dashed line shows a path for the navigation from a room to another.

based navigation using learned semantic map representation.

Most approaches mentioned above have two main problems:

- (1) Each node in its topological structure is a specific target, which may not be observed through the motion offset of mobile robots on the edge.
- (2) More than one sensor is used, such as a camera for image collection and a laser for mobile robot mapping and motion.

The two problems have motivated our current work, aiming at achieving visual semantic navigation in a human-like way using only one camera.

*2.2. Visual Semantic Navigation Based on Deep Learning.* People achieve the perception of the environment through images seen by eyes and then guide the behavior. Therefore, we can learn from the “perception-guidance” model to control the robot. In this paper, a three-layer perception framework based on transfer learning is conducted with a common scene (composed of multiple rooms and corridors, as shown in Figure 1). This framework can only rely on image information of a single camera to perceive the surroundings and identify the region where the robot stands and the current pose, which provides decision information for semantic navigation.

*2.2.1. Three-Layer Perception Framework.* Mobile robots usually work in the environment shown in Figure 1. It can be supposed that the number of rooms is  $n$  ( $n \in N^+$ ) and the semantic task is to move the robot from a room named  $R_i$  ( $i < n$ ,  $i \in N^+$ ) to  $R_j$  ( $j < n$ ,  $j \in N^+$ ). To achieve this semantic task, the robot is required to determine the initial semantic region firstly and then plan the path to reach the target region (the dashed line for the navigation path as shown in Figure 1). As the input information of the robot is merely images acquired by a camera, the learning algorithm can be used to train the robot’s perception model of the environment to realize the semantic navigation purpose.

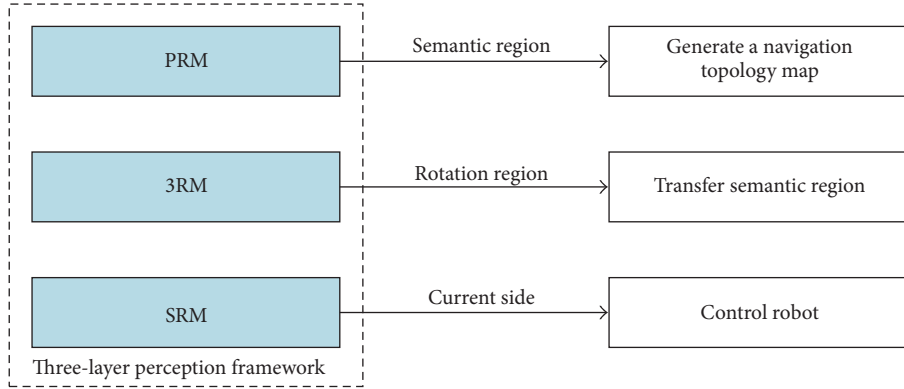


FIGURE 2: Three-layer perception framework.

Each room and each corridor is usually classified as a category, but only the region where the robot is located can be identified. Additional sensors are needed to implement automatic navigation for the robot, although we have already obtained a semantic map. It is difficult to complete the whole semantic navigation in a single neural network model because it cannot provide the robot with all the navigation information simultaneously. Therefore, we design a three-layer perception framework consisting of three perceptual models, which are place recognition model (PRM), rotation region recognition model (3RM), and “side” recognition model (SRM), as shown in Figure 2.

The PRM is used to identify the semantic region where the robot is currently located. And then it gets the navigation topological map according to the semantic task. The 3RM is to identify the key regions when transferring between regions, such as the rotation position at a door when a robot moves from a room to the corridor. The SRM is to provide the relative pose information between a robot and the environment to control the movement. The “side” means that a robot is located at a side. A robot perhaps locates in the left side, center, or right side, when it moves in a corridor or aisle.

(1) *Place Recognition Model (PRM)*. When the robot implements semantic tasks, it is necessary to determine the semantic region where it stands and the target places and then to carry out semantic navigation planning. In an environment similar to Figure 1, the semantic task may be moving from a position of one room to a target region of another room, or from a corridor to a specific region of one room. It is hard to perform navigation planning because the initial position and orientation are both uncertain. There are several main aisles for walking in a room; therefore, several regions can be divided according to these aisles, and each one is regarded as a semantic region. The recognition model of semantic regions can be trained using the method of image classification in machine learning. It needs to collect images of different positions and perspectives in each semantic region as training samples.

Set the number of semantic regions divided in the  $i$ th room as  $n_i$ , and the number of semantic regions in the corridor

as  $n_{\text{Corr}}$ ; then, the total number of semantic regions can be calculated by

$$N_{\text{Sem}} = \sum_{i=1}^n n_i + n_{\text{Corr}}. \quad (1)$$

Deep learning is widely used in image classification and has obtained excellent achievements in the ImageNet Challenge; for instance, the top-5 network model accuracy rate of Google’s Inception-V3 reaches up to 96.5% [26, 27]. In addition, transfer learning can use the complex trained neural network model to train the new classification to reduce the amount of training samples and save training time [28, 29]. Therefore, the neural network model for recognizing semantic regions is designed using transfer learning. The Inception-V3 model consists of 11 layers of inception module, which uses multiple branches to extract high-level features of different abstraction levels to enrich the expressive ability. The neural network model framework of semantic region perception based on transfer learning is shown in Figure 3. The model’s input is RGB images. The parameters of Inception-V3 model trained on the ImageNet dataset are used to calculate the network forward transmission, and 2048 nodes are obtained in the bottleneck layer. Then, the last fully connected layer FC is replaced. The number of output categories is the total number of semantic regions  $N_{\text{Sem}}$ . Then, the Softmax layer is calculated, and the output is the probability of each semantic region category. We use Inception-V3 model for image feature extraction directly and then take the extracted bottleneck feature vector to train a single-layer fully connected neural network.

Suppose that the input RGB image is  $I$ , the bottleneck output is  $y_{bp}$  ( $p \in [1, N_b]$ ) (the subscript “b” means bottleneck;  $N_b$  is the number of bottleneck layer nodes) after calculating function  $f_{V3}$  of Inception-V3 module, and the output of FC layer is  $y_{cq}$  ( $q \in [1, N_{\text{Sem}}]$ ) (the subscript “c” means fully connected layer).

The bottleneck output can be calculated by

$$y_{bp} = f_{V3}(I). \quad (2)$$

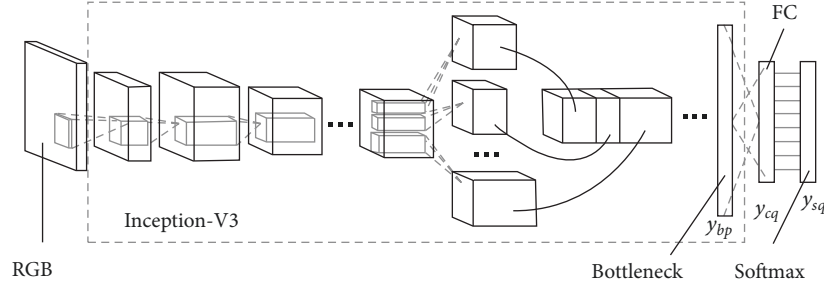


FIGURE 3: The diagram of PRM based on transfer learning.

And the ReLU is selected as an activation function in the FC layer, and then the output of the FC layer can be given by

$$y_{cq} = f_{\text{ReLU}} \left( \sum_{p=1}^{N_b} w_{pq} y_{bp} + b_q \right), \quad (3)$$

where parameters  $w$  and  $b$  are the weight and bias of the FC layer, respectively.

Model parameters of the FC layer are needed to be trained in the network, and the number of parameters can be given by

$$\text{Num} = N_b \cdot N_{\text{Sem}} + N_{\text{Sem}}, \quad (4)$$

where Num is the amount of parameters.

Finally, the Softmax function is used to obtain the probability of each output. The output  $y_{sq}$  ( $q \in [1, N_{\text{Sem}}]$ ) (the subscript “s” means Softmax layer) can be obtained by

$$y_{sq} = f_{\text{Soft max}} (y_{cq}) = \frac{e^{y_{cq}}}{\sum_{q=1}^{N_{\text{Sem}}} e^{y_{cq}}}. \quad (5)$$

For every input image, the probability value belonging to each category can be calculated by the model PRM. And the category with a maximum probability is the final recognition result.

(2) *Rotation Region Recognition Model (3RM)*. Through the place recognition model, the robot can recognize the region location where it stays, but it also needs the recognition information of region transfer to reach the target one, such as how to reach the door from a region of the room. The robot usually performs rotation during the transfer among semantic regions, and it needs to recognize where to rotate. In order to realize the recognition of the rotation region for the robot by the image information, a rotation recognition perception model is proposed based on transfer learning. The rotation position of the robot is identified by sensing room regions and door regions connected with the corridor.

Usually, the movement of the robot between room and corridor is divided into four cases: (a) from room to the left side of corridor, (b) from room to the right side of corridor, (c) from the left side of corridor to room, and (d) from the right side of corridor to room. We can obtain the best rotation region by analyzing the volume of the robot and the turning

radius, which means reaching nearly the center line of the doorway or corridor after rotating. Therefore, the navigation recognition region can be divided at the door as shown in Figure 4.

The navigation recognition regions are drawn as three blue dashed boxes in Figure 4. And  $RR_{i1}$  is the recognition region entering the room from the left side of the corridor (Path 1).  $RR_{i2}$  is the recognition region entering the room from the right side of the corridor (Path 4).  $RR_{i3}$  is the recognition region entering the left or right corridor from the room (Path 2 and Path 3). In brief,  $RR_{i1}$  and  $RR_{i2}$  are referred to as the entrance recognition regions, and  $RR_{i3}$  is called the exit recognition region. The symbol “ $i$ ” in the subscript means the  $i$ th room.

The navigation in the room among semantic regions also needs to identify the rotation recognition region, shown in Figure 5. Semantic regions are divided in the room according to the method described in the previous section. The number of semantic regions divided in the  $i$ th room is  $n_i$ . The robot should recognize the rotating positions between the two adjacent semantic regions. Therefore, it is necessary to determine the location of the recognition region and to collect the images. Recognition regions are described as dashed boxes in Figure 5, and arrows indicate the movement direction of the robot.

In order to determine the number of rotation recognition regions, it is necessary to analyze the distribution of the recognition region in the room. The center line of the region can be connected (such as the red dashed line in Figure 5), and then the required recognition regions can be obtained according to the connection. For connections like the “ $T$ ” type, three regions are required, while the “ $L$ ” type requires two regions.

We can suppose that the number of “ $T$ ” type connections in the  $i$ th room is  $n_T^i$ , and the number of “ $L$ ” type connections is  $n_L^i$ ; then, the number  $n_{\text{Rot}}^i$  of regions in the  $i$ th room can be calculated by

$$n_{\text{Rot}}^i = 3n_T^i + 2n_L^i. \quad (6)$$

In addition, there are three rotation recognition regions at the door of each room; the total number  $N_{\text{Rot}}$  of regions is given by

$$N_{\text{Rot}} = \sum_{i=1}^n n_{\text{Rot}}^i + 3n. \quad (7)$$

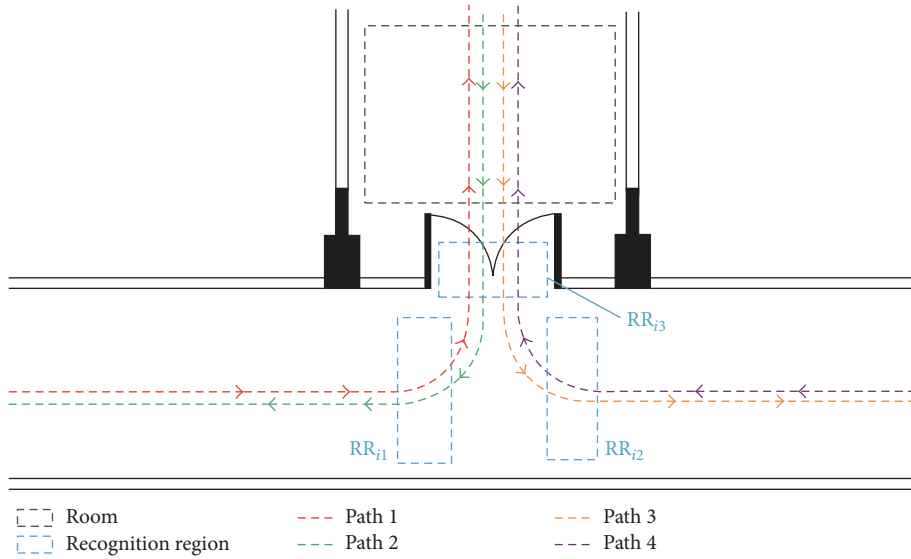


FIGURE 4: The diagram of rotation recognition region at the door. The green dashed box means room area. The three blue dashed boxes mean recognition regions. The four dashed lines show the different movement directions of the robot.

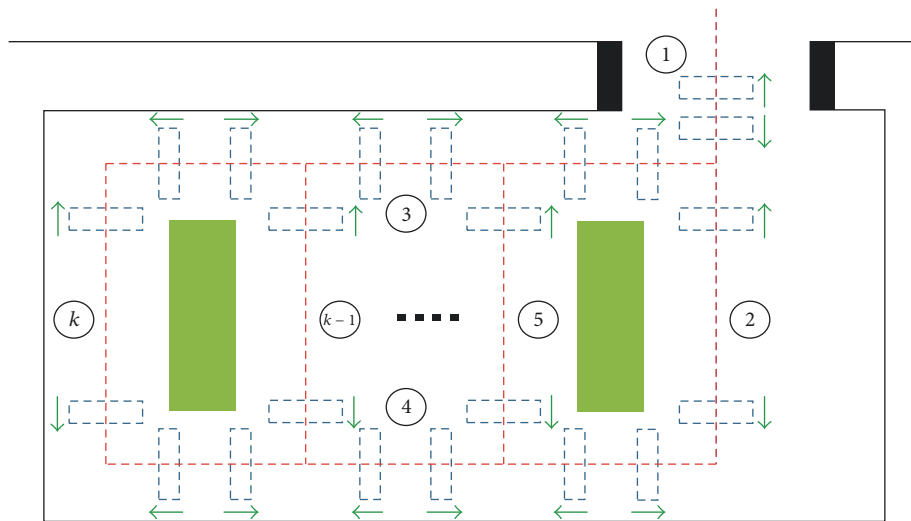


FIGURE 5: Diagram of semantic regions and rotation recognition regions divided in the room. The circled numbers indicate semantic regions. The green blocks describe objects in the room. The robot's accessible area is indicated by the red dashed line. The blue dashed boxes are recognition regions and arrows mean the movement direction of the robot.

When the robot moves between rooms and corridors, it needs to recognize the rotation recognition region firstly and then rotate to the corresponding direction. The locations of recognition regions are determined in each semantic region and images should be collected in the corresponding directions. Then, each region is trained as one category. The location and orientation of the robot are limited in a fixed range for each recognition region when collecting images. In the region  $RR_{i1}$ , the robot should be in the center line of the corridor with the direction towards the right. In the region  $RR_{i2}$ , the robot locates in the center line of the corridor with the direction towards the left. In the region  $RR_{i3}$ , the robot should be at the center of the doorway with the direction towards the corridor (perpendicular to the corridor extension

direction). Besides, we need to collect images outside the rotation recognition region for training in neural networks as nonrecognition region.

The neural network is trained by the method of transfer learning, and the output nodes are the rotation recognition regions and the nonrecognition region; the amount is  $N'_{Rot}$ , given by

$$N'_{Rot} = N_{Rot} + 1. \quad (8)$$

As the neural network structure is similar to Figure 2, its structure is not given here. The trained neural network model is used as a rotation recognition region perception model to guide the robot through different semantic regions.

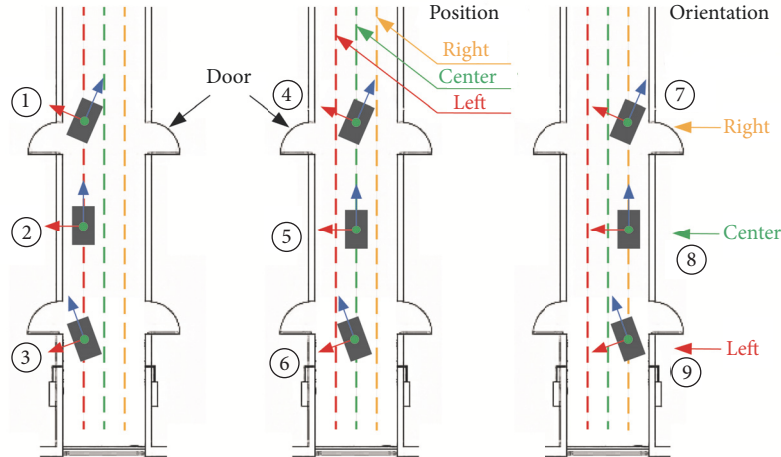


FIGURE 6: Positions and orientations of the mobile robot in the corridor. The nine states of the robot are shown.

(3) “Side” Recognition Model (SRM). In order to reduce the collision when the robot moves in the corridor or rooms, it is necessary to be able to perceive the pose of the robot. As the input for the robot is images, we can learn from the process of people walking in the corridor and then design a “side” recognition model. Firstly, we analyze the operation state of the robot in the corridor. Secondly, the neural network model of recognizing robot pose is trained by transfer learning.

The mobile robot moves in a two-dimensional plane; its pose includes position and orientation. There are nine different states according to the position and orientation when moving in the corridor as shown in Figure 6. Among them, the positions are separated into corridor center, left side, and right side, and the orientations are divided into center direction, left, and right. For convenient description, the nine states of the robot are abbreviated as shown in Table 1. When the robot moves in a room, its pose state is similar to that in the corridor, which has nine states too. As the door region is relatively small, in order to avoid the collision of the robot, it needs to sense and adjust its pose when passing through the door. Thus, the method of image collection at the door of each room is similar to that in the corridor, as shown in Figure 7.

When the robot is in a different state, its camera (fixed on the robot with the forward direction) observes different images. Therefore, we can recognize the pose state through image classification. Similar to the training model above, a neural network model of robot recognition in the corridor is designed using transfer learning. The last full connection layer of Inception-V3 model is modified to output the nine states of the robot, and then the single-layer fully connected neural network is trained.

Images that robots observed in different poses need to be collected when training the model. For facilitating control and reducing the number of perception models, images at the same pose state are put together, as a category to train network model. To cover possible scenarios, data collection takes two movements in the corridor and doorway.

TABLE 1: Nine pose states of mobile robot.

Orientation (O)	Position (P)		
	Right (R)	Center (C)	Left (L)
Right (R)	PROR	PCOR	PLOR
Center (C)	PROC	PCOC	PLOC
Left (L)	PROL	PCOL	PLOL

TABLE 2: Robot poses and control strategies.

Robot poses	Control strategies
① PLOR	Turn left and move to center from left
② PLOC	Move to center from left
③ PLOL	Turn right and move to center from left
④ PCOR	Turn left
⑤ PCOC	Move forward
⑥ PCOL	Turn right
⑦ PROR	Turn left and move to center from right
⑧ PROC	Move to center from right
⑨ PROL	Turn right and move to center from right

The initial position and orientation may be in any cases above, and the movement may deviate from the center direction; it is necessary to adjust the control according to each case. The pose states with corresponding control strategies are given in Table 2. The robot’s state can be recognized through the input image, and corresponding control is conducted, which makes the robot move along the center.

2.2.2. *Semantic Navigation of the Mobile Robot.* When the mobile robot performs semantic navigation between multiple rooms and corridors, it is necessary to determine the topological relations between semantic regions according to prior information. An indoor environment usually contains several rooms and corridors; the topological relationship (shown in Figure 8) between any semantic regions can be given combined with the indoor semantic region division (shown

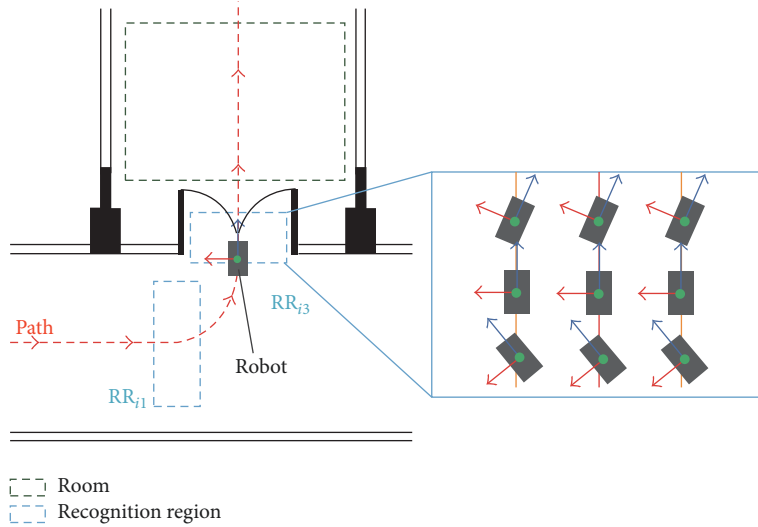


FIGURE 7: Positions and orientations of the mobile robot in the doorway.

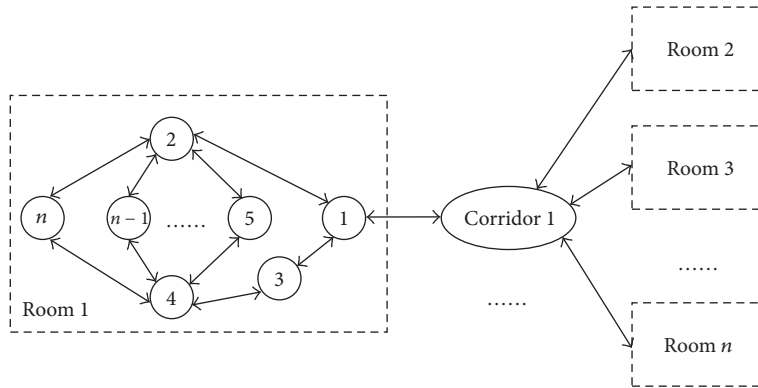


FIGURE 8: Semantic region topology diagram.

in Figure 5). The semantic topological relation diagram is a directed graph of connected nodes, the node is a set of semantic regions, and the edge is a set of rotation recognition regions. The semantic regions are connected by rotation recognition regions. The topological relations between any two semantic regions can be calculated by the directed graph, and the navigation path with the smallest number of semantic regions is selected as the optimal path.

The robot's semantic navigation path can be generated automatically through the topology diagram of Figure 8, which is used to guide the robot motion. Meanwhile, the three-layer perception framework described in Section 3.1 is used to obtain corresponding perceptual information to make decisions. Assuming that the robot is currently in the door region of the room  $R_j$ , the semantic task is to reach the room  $R_j$ ; then, the robot's decision process is as follows:

- It determines the semantic region using PRM.
- It obtains the pose using the pose perception model and adjusts position and orientation to move towards the door.
- It determines whether the robot is in nonrecognition region or rotation recognition region using 3RM. The

robot keeps going straight if it is in the nonrecognition region. And the robot rotates to corridor when the exit recognition region is detected.

- It obtains the robot's pose relative to the corridor using SRM, and the robot moves along the corridor center line through the control strategy in Table 2.
- The robot rotates towards the room when the entrance recognition region of room  $R_j$  is detected using 3RM.
- The robot moves along the center line of the doorway through the control strategies in Table 2.

We achieve the robot semantic navigation from the current region to the target through the algorithm above. In the whole process, the robot only relies on images information for perception and decisions, without using odometer, laser, or other sensor information.

### 3. Results and Discussion

In order to verify the validity of the proposed visual semantic navigation algorithm, several experiments are carried out on

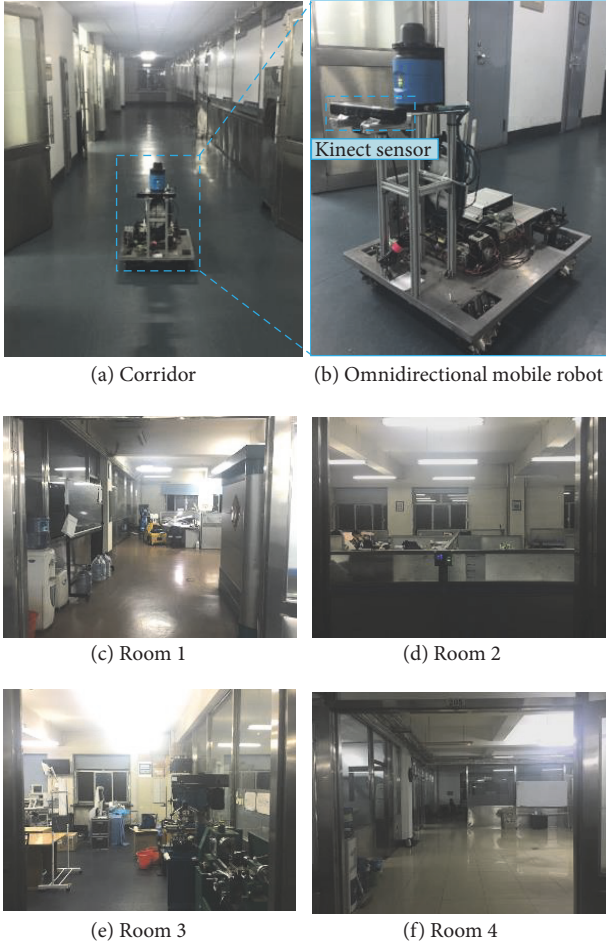


FIGURE 9: Experimental scene and omnidirectional mobile robot.

a mobile robot platform. Firstly, the experimental environment is introduced, and then the training process of the three-level environment perception model is given. Finally, the semantic navigation experiments of the robot from the corridor to room and from room to room are carried out.

**3.1. Introduction to the Experimental Environment.** An indoor environment including one corridor and four rooms is selected to verify the semantic navigation algorithm, shown in Figure 9. The experimental mobile robot is an omnidirectional platform using Mecanum wheels. It is able to implement movements in any direction. A Kinect sensor is used in the experiment only using RGB color images.

**3.2. Training of the Three-Layer Perception Framework.** RGB images are collected using a Kinect sensor to train three perception models in the environment as shown in Figure 9.

**3.2.1. Training of PRM.** The model is used to classify semantic regions in four rooms and a corridor. We control the robot to move and collect images simultaneously. The robot rotates in a circle when it moves one meter forward. The frame rate of Kinect is 30 f/s, which can capture sufficient training data. In

TABLE 3: Training data for three-layer perception framework.

Model	Output nodes	Images quantity	Test set quantity	Accuracy (%)
PRM	9	74600	7460	96.8
3RM	21	83750	8375	94.2
SRM	9	136580	13658	96.5

each room, two main aisles are selected as semantic regions. The nodes of model's output are nine when adding up a corridor region.

**3.2.2. Training of 3RM.** The model provides rotation positions for robot navigation among semantic regions. Firstly, it is vital to determine the recognition region at the doorway of each room. According to the description in Section 3.1, we collect images in entrance recognition regions, exit recognition regions, and nonrecognition region. In addition, images at recognition regions in each room should be collected. The trained network model has 21 output nodes according to (7) and (8).

**3.2.3. Training of SRM.** The model provides position and orientation perception information relative to the environment for motion control. RGB images are collected in the method described in Section 3.1 in the corridor. Respectively, nine categories of images in different positions and orientations of the corridor and rooms are collected. The position and orientation are divided into center, left side, and right side separately.

In addition, we need to collect the corridor images in two directions. The correction strategies for the nine poses are consistent whether the robot is in the corridor or rooms. Therefore, images in the same state are trained as one category and the number of output nodes is nine.

The neural network training is carried out using the transfer learning algorithm with Leadtek Quadro K4200 graphics card. 10% of the sample data is randomly selected as validation set and 10% is selected as test set. The training data and results of the three models are shown in Table 3. The accuracy of test results in the test set is quite high, indicating that the trained neural network models have quiet good recognition effects on semantic regions, recognition regions, and robot poses.

**3.3. Visual Semantic Navigation Experiments.** In order to verify the validity of the proposed three-layer perception framework for the robot semantic navigation task, experiments of the robot from corridor to room and room to room are carried out.

**3.3.1. Semantic Navigation Experiments from Corridor to Room.** In order to verify the effectiveness and robustness of the model, semantic navigation experiments of all nine initial poses are carried out. The experimental results are shown in Table 4, which shows the initial position images of the robot in corridor, the corridor images observed by



TABLE 4: Semantic navigation experiment from corridor to room.



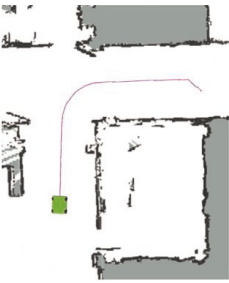






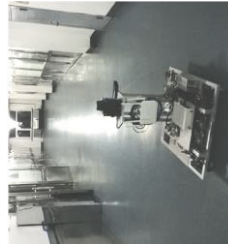




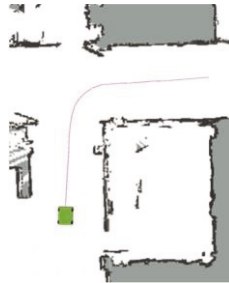

Seq.	Robot states	Initial poses of the robot	Images observed by the robot's camera	Movement trajectories
(1)	PLOR			
(2)	PLOC			
(3)	PLOL			
(4)	PCOR			
(5)	PCOC			

TABLE 4: Continued.

Seq.	Robot states	Initial poses of the robot	Images observed by the robot's camera	Movement trajectories
(6)	PCOL			
(7)	PROR			
(8)	PROC			
(9)	PROL			

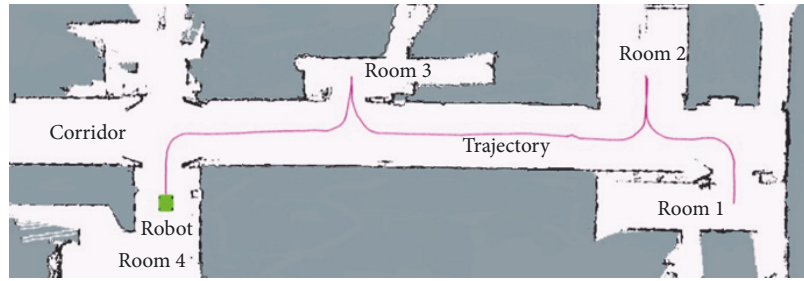


FIGURE 10: Semantic navigation experiment from room to room.

camera, and the trajectories of semantic navigation. It is obvious that the corridor images in different positions are different, which is beneficial to classify. In order to describe the process of semantic navigation, the environment map is established using a two-dimensional laser, and the trajectory of the robot is displayed on the map. From the trajectory diagrams, it can be observed that when there is a deviation between the current pose state and the corridor center, pose adjustment is implemented autonomously. In the doorway, rotation operation is implemented when the robot recognizes the entrance recognition region, and multiple pose adjustments are conducted according to the observed state. The robot can move along the center of the doorway in this method.

The validity and stability of the SRM are verified by the experiments above. The robot can correct the position and orientation by recognizing the current state in different initial poses and correct its own pose in real time to ensure stable movement.

*3.3.2. Semantic Navigation Experiments from Room to Room.* Semantic navigation experiments from room to room are carried out to verify the validity of the proposed framework. A series of semantic tasks which are “Room 1 → Room 2 → Room 3 → Room 4” are conducted. The robot is initially located at the semantic region of the doorway in Room 1, but the initial information is not given. The semantic tasks are to arrive at regions in the other three rooms, respectively.

The movement trajectory is shown in Figure 10. Firstly, the robot recognizes its semantic region by PRM and generates a semantic navigation topology. Then, the robot rotates to the left into the corridor when the exit recognition region is detected. The robot goes straight until recognizing the entrance recognition region of Room 2 and rotates to the right to enter the room. The same process is implemented to arrive at Rooms 3 and 4. Through the trajectory, it can be seen that the robot can adjust itself to realize the semantic tasks and keep moving along the center line until the target region is recognized.

Experiments show that the three-level perception framework can perform well in the semantic task only using a camera. It provides vital information to guide the navigation. In addition, it can correct the attitude of movement continuously which yields higher reliability and stability.

## 4. Conclusion

In this paper, a novel visual semantic navigation approach is presented using a three-layer perception framework based on transfer learning. The model comprises place recognition model, rotation region recognition model, and “side” recognition model, which are used to determine the semantic region and recognize the position of the rotating region and the pose information relative to the environment. Only a single camera sensor is employed in our system. Additionally, the “side” recognition model is able to correct the robot’s pose automatically and improve the operational reliability. Semantic navigation experiments are carried out in corridors and rooms, and the results verify the applicability and robustness of our method. We would like to explore the adaptability of changing environment and semantic planning algorithm considering dynamic pedestrians in the future work.

## Disclosure

Li Wang and Guanglei Huo are joint first authors.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (61473103, 61673136, and 61473120), the Natural Science Foundation of Heilongjiang Province, China (F2015010), Self-Planned Task (nos. SKLRS201715A, SKLRS201609B, and SKLRS-2017-KF-13) of the State Key Laboratory of Robotics and System (HIT), the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (no. 51521003), and Science and Technology Planning Project of Guangzhou (201607010006).

## References

- [1] A. Pandey, S. Kumar, K. K. Pandey, and D. R. Parhi, “Mobile robot navigation in unknown static environments using ANFIS controller,” *Perspectives in Science*, vol. 8, pp. 421–423, 2016.

- [2] H. Omrane, M. S. Masmoudi, and M. Masmoudi, "Fuzzy logic based control for autonomous mobile robot navigation," *Computational Intelligence and Neuroscience*, vol. 2016, Article ID 9548482, 10 pages, 2016.
- [3] L. Palmieri, A. Rudenko, and O. A. Kai, "A fast random walk approach to find diverse paths for robot navigation," *IEEE Robotics Automation Letters*, vol. 2, no. 1, pp. 269–276, 2017.
- [4] R. Kümmerle, M. Ruhnke, B. Steder, C. Stachniss, and W. Burgard, "Autonomous robot navigation in highly populated pedestrian zones," *Journal of Field Robotics*, vol. 32, no. 4, pp. 565–589, 2015.
- [5] C. Galindo, J.-A. Fernández-Madrigal, J. González, and A. Saffiotti, "Robot task planning using semantic maps," *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 955–966, 2008.
- [6] D. W. Ko, C. Yi, and I. H. Suh, "Semantic mapping and navigation: A Bayesian approach," in *Proceedings of the 2013 26th IEEE/RSJ International Conference on Intelligent Robots and Systems: New Horizon, IROS 2013*, pp. 2630–2636, Japan, November 2013.
- [7] A. Borkowski, B. Siemiatkowska, and J. Szklarski, "Towards semantic navigation in mobile robotics," *Graph Transformations & Model-driven Engineering-essays*, pp. 719–748, 2010.
- [8] K. Uhl, A. Roennau, and R. Dillmann, *From structure to actions: semantic navigation planning in office environments*, *IROS Workshop on Perception and Navigation for Autonomous Vehicles in Human Environment*, From structure to actions, semantic navigation planning in office environments, 2011.
- [9] T. S. Veiga, P. Miraldo, R. Ventura, and P. U. Lima, "Efficient object search for mobile robots in dynamic environments: Semantic map as an input for the decision maker," in *Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2016*, pp. 2745–2750, Republic of Korea, October 2016.
- [10] D. Pangercic, B. Pitzer, M. Tenorth, and M. Beetz, "Semantic Object Maps for robotic housework - Representation, acquisition and use," in *Proceedings of the 25th IEEE/RSJ International Conference on Robotics and Intelligent Systems, IROS 2012*, pp. 4644–4651, Portugal, October 2012.
- [11] C. Landsiedel, R. De Nijs, K. Kuhnlenz, D. Wollherr, and M. Buss, "Route description interpretation on automatically labeled robot maps," in *Proceedings of the 2013 IEEE International Conference on Robotics and Automation, ICRA 2013*, pp. 2251–2256, Germany, May 2013.
- [12] G. Huo, L. Zhao, K. Wang, and R. Li, "Semantic region estimation of assistant robot for the elderly long-term operation in indoor environment," *China Communications*, vol. 13, no. 5, Article ID 7489969, pp. 1–15, 2016.
- [13] S. Lowry, N. Sunderhauf, P. Newman et al., "Visual Place Recognition: A Survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.
- [14] R. Drouilly, P. Rives, and B. Morisset, "Semantic representation for navigation in large-scale environments," in *Proceedings of the 2015 IEEE International Conference on Robotics and Automation, ICRA 2015*, pp. 1106–1111, USA, May 2015.
- [15] R. Drouilly, P. Rives, and B. Morisset, "Fast hybrid relocation in large scale metric-topologic-semantic map," in *Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2014*, pp. 1839–1845, USA, September 2014.
- [16] S. L. Joseph, C. Yi, J. Xiao, Y. Tian, and F. Yan, "Visual semantic parameterization - To enhance blind user perception for indoor navigation," in *Proceedings of the 2013 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2013*, USA, July 2013.
- [17] L. F. Posada, F. Hoffmann, and T. Bertram, "Visual semantic robot navigation in indoor environments," *International Symposium on Robotics*, pp. 1–7, 2014.
- [18] Z. Zhao and X. Chen, "Semantic mapping for object category and structural class," in *Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2014*, pp. 724–729, USA, September 2014.
- [19] L. Horne, J. M. Alvarez, C. McCarthy, and N. Barnes, "Semantic labelling to aid navigation in prosthetic vision," in *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2015*, pp. 3379–3382, Italy, August 2015.
- [20] Y. Jiang, C. Yang, J. Na, G. Li, Y. Li, and J. Zhong, "A brief review of neural networks based learning and control and their applications for robots," *Complexity*, vol. 2017, no. 4, Article ID 1895897, pp. 1–14, 2017.
- [21] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "SemanticFusion: Dense 3D semantic mapping with convolutional neural networks," in *Proceedings of the 2017 IEEE International Conference on Robotics and Automation, ICRA 2017*, pp. 4628–4635, Singapore, June 2017.
- [22] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 881–893, 2017.
- [23] Y. Xiang and D. Fox, "DA-RNN: Semantic Mapping with Data Associated Recurrent Neural Networks," in *Proceedings of the Robotics: Science and Systems 2017*.
- [24] Y. Zhu, R. Mottaghi, E. Kolve et al., "Target-driven visual navigation in indoor scenes using deep reinforcement learning," in *Proceedings of the 2017 IEEE International Conference on Robotics and Automation, ICRA 2017*, pp. 3357–3364, Singapore, June 2017.
- [25] Y. Furuta, K. Wada, M. Murooka et al., "Transformable semantic map based navigation using autonomous deep learning object segmentation," in *Proceedings of the 16th IEEE-RAS International Conference on Humanoid Robots, Humanoids 2016*, pp. 614–620, Mexico, November 2016.
- [26] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*, pp. 1–9, Boston, Mass, USA, June 2015.
- [27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 2818–2826, July 2016.
- [28] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
- [29] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, article no. 9, 2016.

