Hindawi
Complexity
Volume 2018, Article ID 1627185, 12 pages
https://doi.org/10.1155/2018/1627185

Research Article

Visual Semantic Navigation Based on Deep Learning for Indoor Mobile Robots

Li Wang,[1] Lijun Zhao ,[1] Guanglei Huo,[2] Ruifeng Li,[1] Zhenghua Hou,[1] Pan Luo,[1] Zhenye Sun,[1] Ke Wang,[1] and Chenguang Yang [3]

[1]*State Key Laboratory of Robotics and System, Harbin Institute of Technology, Harbin 150001, China*
[2]*HNA Technology Group, Shanghai 200122, China*
[3]*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

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.

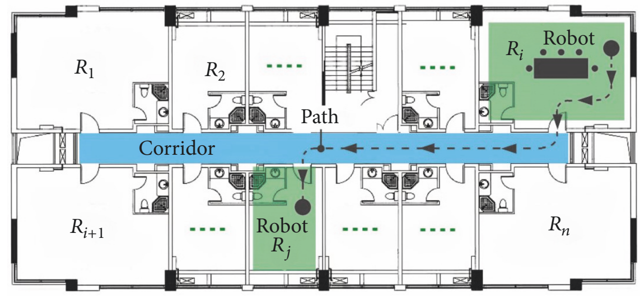

Figure 1: The diagram for mobile robot working in an indoor environment. There are several rooms and a corridor in the diagram. A trajectory with a dashed line shows a path for the navigation from a room to another.

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 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.