Robot Indoor Text Contents Recognition Based on Visual SLAM

Lan Hu
Guangzhou Yew Wah International Education School, Guangzhou, China.
Email: t88439172@163.com

Abstract. Although research on indoor mobile robots has evolved over the years and has made good progress in many areas, few researchers have noticed their combination in text/character recognition and navigation applications. In fact, robots navigate and move with obstacles and various visual disturbances without prior knowledge, and people and objects are possible obstacles. Indoor robots may need to recognize text/characters to help them determine their location and move on. In order to enable mobile robots to navigate and locate more accurately, we improved the visual text content recognition system based on Faster RCNN. This method uses the attention mechanism to replace the original re-sharp method in Faster RCNN, and the visual attention mechanism can be adopted for the anchor. Reduce the number and error, and reduce the interference of non-target information on the identification system. This method can reduce the problem of large system consumption time. Experiments on the indoor robot dataset show that the method has certain advantages in accuracy and cost time.

1. Introduction

There is a growing market demand for vision-based autonomous mobile service robots that typically require radar, ultrasound, and cameras as a perceptual module for the robot. Cameras are often an essential part of a robot, but more software algorithms are needed to drive these devices [1].

Real-time motion and structural estimation of a single mobile camera has potential applications in the fields of robotics, wearable computing, augmented reality and automotive. In recent years, advances in computer processing power and algorithms have made great progress in these areas, and many good systems have existed. The main method includes: the filtering method sequentially fuses the measured values of all the images by updating the probability distribution of the features and camera pose parameters; and the bundle adjustment method performs batch optimization on the selected images from the real-time stream. Both methods are used for stereo vision as well as monocular vision [2].

Many robotic applications, such as navigation and mapping, require accurate and drift-free attitude estimation of moving cameras. Previous solutions supported visual feature-based methods as well as beam adjustment or pose map optimization. Although these methods are prior art, the process of selecting the relevant key points discards most of the acquired image data. In SLAM, the classic problem is to estimate the motion of the mobile robot in real time as the mobile robot continuously observes and uses sensors that may or may not include the camera to map its unknown environment. Here the sequential filtering technology has come to the fore [3].

Simultaneous Localization and Map building (SLAM) is the challenging problem in mobile robots that has attracted the interest of more and more researchers over the past decade. The self-positioning of mobile robots is clearly a fundamental problem in autonomous navigation: mobile robots must be able to estimate their position and orientation (posture) within the map of their navigation environment.
However, in many practically relevant applications (such as exploration missions or operations in harsh environments), maps are unavailable or highly uncertain. Therefore, in this case, the robot must use the measurements provided by its sensing device to estimate the environmental map while positioning itself within the map [4].

Research in related literatures include [5]: EKF is used to estimate position and motion information, and it is recommended that the FPGA provide an additional execution engine for the matching processing module. By introducing multi-agent ideas into particle swarm optimization particle filter; a simultaneous localization and mapping method based on multi-agent particle swarm optimization particle filter; neural evolution optimization SLAM (NeoSLAM); based on image feature recognition, describes the actual use Successful experiment of mobile robot's automatic navigation and proposed map representation; a new single-eye EKF SLAM feature initialization method that utilizes the 3D measurement model in the camera frame instead of the 2D pixel coordinates in the image plane; determined in the map The estimated location of the landmarks is closely related to the actual location of these landmarks in real life; the combined control and SLAM system is described, and the attractions it has gained in successful application in the regional world environment are discussed; the use of three is called CPS- The newly developed measurement system consisting of mobile robots successfully carried out measurement experiments in unknown and large indoor/outdoor environments, including halls, buildings, urban areas and cultural heritage; an RGB-D simultaneous positioning and mapping (SLAM) a new method of the system that uses both vanishing points and door panels in the corridor environment Air China.

With the development of technology and increasing market demand, indoor robots are gradually becoming an important member of the robot family. In addition to the above positioning and navigation problems, the recognition of characters and words is also a hot research topic, especially the characters and words of handwriting. The difficulty lies in the identification of the constraints formed by the original handwriting. In fact, we have no control over authors, writing tools, or writing styles. In addition, varying degrees of neatness are possible, from very sloppy to extremely neat. However, these difficulties can be offset by the constraints of the input words from fixed vocabulary [6].

Text localization and recognition in real-world scenarios is an open question because it is a key component in many computer vision applications, such as searching for images through their text content, reading commercial tags in map applications, or assisting visually impaired people. Several competitions have been played in the past few years, and the winning method in the recent ICDAR 2011 competition is able to correctly locate 62% of the words. Locating text in an image can be a computationally expensive task, and the text localization method handles the problem in two different ways. The sliding window-based method limits the search to a subset of the image rectangle. This will reduce the number of subsets that check for the presence of text. The second method finds individual characters by grouping pixels into regions using connected component analysis, if pixels belonging to the same character have similar properties. The advantage of the connected component approach is that their complexity is generally independent of the attributes of the text, and they also provide segments that can be utilized in the OCR step, with the disadvantage that sensitivity to clutter and occlusion can change the connected component structure [7].

Although research on indoor mobile robots has evolved over many years and has made good progress in many areas, few researchers have noted their combination in text/character recognition and navigation applications. In fact, indoor robots are likely to need to recognize text/characters to help them determine their position and continue to move. For example, it may be necessary to identify the door’s number to prevent entering the wrong room; it may be necessary to read and understand the text on the object. To deal with these objects that have not been known; and among people with disabilities who have language barriers, it is a common practice to write orders. So, the handwritten commands of these people need to be identified by robots.

This paper proposes an improved indoor text content recognition method based on SLAM technology. The method acquires RGB and depth image information based on multiple sensors, integrates the information based on the improved Faster RCNN structure, identifies the text type content in the room, and selects the optimization objective function to maximize the effective
coverage area of the SLAM-based navigation. The goal is to achieve a balance between efficiency and robustness in real-time recognition and navigation.

The rest of the paper is as follows: In Section II, we present the methods of acquiring RGB and depth image information and building SLAM. The proposed Faster RCNN structure is given in Section III. Our experimental results and conclusions are presented finally.

2. Acquiring Image Data and Building SLAM

2.1. Acquiring RGB and Depth Image Information
Providing full-size information, including true 3-D images of depth maps and color images of any scene, remains challenging. Real-time capture of depth map information is critical in a variety of applications, such as robotics, security, automotive and interactive gaming, because it enables electronic devices to identify and track target objects without the need for complex post-processing.

Time-of-flight (TOF) technology is one of the most advantageous methods for 3D image capture because they allow all pixels to provide their own depth information in real time. There are two ways to use TOF technology. One is to measure the round-trip time of the emitted light between the light source and the target object by using a single photon avalanche diode. However, since the required TDC and memory size is larger than the entire pixel array, they have a limitation of reducing the pixel size. Very high sensitivity to pixel signal saturation limits their use even under normal lighting conditions. In contrast, another method detects the phase difference of a modulated optical signal from a light mixing device capable of electro-optical demodulation, commonly referred to as a locked pixel [8].

Recent advances in 3D depth cameras such as Microsoft Kinect sensors [9] have created many opportunities for multimedia computing. The Kinect sensor allows the computer to directly sense the depth information of the player and the environment. It also understands when users are talking, knowing who they are going to, and can explain their actions and turn them into a format that developers can use to build new experiences. Kinect sensors integrate a variety of advanced sensing hardware.

2.2. Building SLAM on Indoor Robots
Autonomous mobile robots must be able to navigate in an unknown environment. SLAM issues are related to this autonomy. Vision sensors are attractive devices for autonomous mobile robots because they are informative and rarely limited in various applications. However, many vision-based SLAM methods that use a universal pinhole camera suffer from variations in illumination and occlusion because they primarily extract corner points of the feature map. Moreover, due to the narrow field of view of pinhole cameras, they are not suitable for high speed camera motion. And affected by lighting and partial occlusion [10].

The vSLAM object recognition system relies on vision-based object recognition, which focuses on identifying planar texture objects. In addition to using a camera instead of a laser scanner as its primary sensor, vSLAM works in a similar way to other SLAM software packages under study. The robot running vSLAM sends images from the onboard network camera to the object recognition system of the vSLAM. The object recognition system first extracts features from each image frame it processes, and if there are enough interesting features in the frame, the object recognition system stores these features as "objects." Each "object" records the current encoder data of the robot. As the robot moves through the environment, it records the location of obstacles encountered. This information is combined with the object/encoder data to create an occupied grid map that provides a bird's eye view of the robot that has arrived and provides the vSLAM with the necessary information for path planning [11]. The SLAM example which is shown in Fig.1 was collected indoor.
Figure 1. The SLAM example collected indoor.

3. Improved Faster RCNN for Contents Recognition

3.1. Basic RCNN Structure
For many practical applications of indoor robots, it is very important to find and recognize text content in natural images. This technique includes two subtasks: text detection and recognition. The big changes in the text model and the highly chaotic background constitute the main challenge [13].

The Region-Based Convolutional Network Method (RCNN) achieves excellent object detection accuracy by classifying object proposals using deep ConvNet. However, R-CNN has obvious drawbacks: its training is a multi-stage approach; training is expensive both in space and time; R-CNN is slow because it proposes ConvNet forward delivery for each object, and Do not share calculations. Space Pyramid Pooling Networks (SPPnets) accelerate R-CNN by sharing computations. The SPPnet method calculates a convolutional feature map for the entire input image and then classifies each object offer using the feature vectors extracted from the shared feature map. SPPnet accelerates R-CNN by 10 to 100 times during testing [14].

3.2. Improved Faster RCNN
The fast R-CNN network, which is shown in Fig.2, takes the entire image and a set of object proposals as input. The network first processes the entire image using several convolutions and a maximum pooling layer to produce a transformed feature map. Then, for each object proposal, the region of interest collection layer extracts a fixed length feature vector from the feature map. Each feature vector is fed into a series of fully connected layers that are ultimately branched into two sibling output layers: one that produces a softmax probability estimate for the K class object class plus one versatile "background" class and another output four real layer evaluates the number of each K object class. Each set of 4 values encodes the fine bounding box position of one of the K classes [14].

Figure 2. Fast R-CNN architecture [14].

However, it is difficult to directly apply the object detection system using Faster RCNN to text detection and recognition in a robot scene because the latter requires higher positioning accuracy. In fact, in generic object detection, each object has a well-defined closed boundary, and this well-defined boundary may not exist in the text because the text line or word consists of many separate characters or strokes. For object detection, typical correct detection is loosely defined. But full-reading text is a fine-grained recognition task that needs to cover the correct detection of a complete line of text lines
or words. Therefore, text detection often requires more accurate positioning, resulting in different evaluation criteria [13].

This paper proposes an improved Faster RCNN method for visual text detection and recognition of indoor robots. When performing visual text detection, the Faster RCNN uses the RPN to generate the detection frame without using the sliding window and the SS method. This method accelerates the generation of the detection frame. The RPN network classifies anchors through softmax and requires accurate proposal. Proposal combines the two for post processing. In contrast, the proposed method which is shown in Fig.3 uses the attention resharp method in the Resharp phase, specifically using the attention principle of the human eye, and using the learning ability of the neural network to train the resharp method that conforms to the human eye perception, that is, the size of the anchors. It is obtained through the attention method instead of being assigned before. The advantage is to further focus on the area to be detected, remove irrelevant pixel information, and further improve the confidence of the detection.

**4. Experiments and Results**

In order to evaluate the performance of the above methods, we obtained some indoor visual data sets and evaluated and tested them on our indoor robots. The dataset contains RGB and depth images recorded by a single camera and depth camera, plus accelerometer data. Our data set includes 10 target categories, approximately 1000 training data images and 100 test data images. Our initial processing section includes 13 conv layers + 13 relu layers + 4 pooling layers. In order to test the performance of the proposed method and the Faster RCNN method, the mean accuracy (mAP) and cost time of the detection are mainly evaluated.

![The mean accuracy on indoor robot dataset](image)

**Figure 4.** The mean accuracy on two methods on indoor robot.
First, we detect the ten types of visual text content data. Figure 4 shows the accuracy results of the detection. It can be observed that the proposed method has relative advantages in most of the visual text content data categories, and has no advantage in a few categories. Further, we divided the data set into five groups for training and testing, and Figure 5 shows the results of these tests. According to the chart, we can see that our method has a consistent reduction in the consumption time compared to the fast RCNN. Since the operation of the indoor robot is generally based on embedded chip calculation, this calculation time reduction is very necessary. On the surface of these experimental results, the improved method proposed by us has better visual text content detection and recognition characteristics than the traditional method, and has certain practicability in robot navigation and content recognition.

5. Conclusions
With the rapid development of robotics, the demand for indoor robots is increasing. Robots can navigate and move with obstacles and various visual disturbances without prior knowledge. People and objects are possible obstacles. In order to enable mobile robots to navigate and locate more accurately, we improved the visual text content recognition system based on Faster RCNN. This method uses the attention mechanism to replace the original resharp method in Faster RCNN, and the visual attention mechanism can be adopted for the anchor. Reduce the number and error, and reduce the interference of non-target information on the identification system. This method can reduce the problem of large system consumption time. Experiments on the indoor robot dataset show that the method has certain advancement in accuracy and consumption time. With this method, the accuracy of the visual text content recognition of the indoor robot can be further improved, and the indoor navigation target can be completed more accurately and robustly.

6. References
[1] Florczyk S. Robot vision: Video-based indoor exploration with autonomous and mobile robots[M]. John Wiley & Sons, 2006.
[2] Strasdat H, Montiel J M M, Davison A J. Visual SLAM: why filter?[J]. Image and Vision Computing, 2012, 30(2): 65-77.
[3] Kerl C, Sturm J, Cremers D. Dense visual SLAM for RGB-D cameras[C]. Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on. IEEE, 2013: 2100-2106.
[4] Garulli A, Giannitrapani A, Rossi A, et al. Mobile robot SLAM for line-based environment representation[C]. IEEE Conference on Decision and Control. IEEE; 1998, 2005, 44(2): 2041.
[5] An Z, Hao L, Liu Y, et al. Development of Mobile Robot SLAM Based on ROS[J]. International Journal of Innovative Science and Modern Engineering, 2016, 5: 47.
[6] Hull J J. A database for handwritten text recognition research[J]. IEEE Transactions on pattern analysis and machine intelligence, 1994, 16(5): 550-554.
[7] Neumann L, Matas J. Real-time scene text localization and recognition[C]. Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012: 3538-3545.

[8] Kim S J, Kim J D K, Kang B, et al. A CMOS image sensor based on unified pixel architecture with time-division multiplexing scheme for color and depth image acquisition[J]. IEEE Journal of Solid State Circuits, 2012, 47(11): 2834.

[9] Zhang Z. Microsoft kinect sensor and its effect[J]. IEEE multimedia, 2012, 19(2): 4-10.

[10] Kim S, Oh S Y. SLAM in indoor environments using omni-directional vertical and horizontal line features[J]. Journal of Intelligent and Robotic Systems, 2008, 51(1): 31-43.

[11] Ouellette R, Hirasawa K. A comparison of SLAM implementations for indoor mobile robots[C]. Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on. IEEE, 2007: 1479-1484.

[12] https://vision.in.tum.de/research/vslam/lsdslam.

[13] Tian Z, Huang W, He T, et al. Detecting text in natural image with connectionist text proposal network[C]/European conference on computer vision. Springer, Cham, 2016: 56-72.

[14] Girshick R. Fast r-cnn[C]. Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.