A Deep Learning Method for Target Pose Estimation of Planetary Surface

HUANG Lu, MAO Xiaoyan, XING Yan
Beijing Institute of Control Engineering, Beijing 100094, China
e-mail: 1216644919@qq.com

Abstract: The key technical challenge of 6D target attitude estimation from RGB-D image is to make full use of 2D and 3D image point cloud data. Previous traditional target pose estimation methods either extract information from RGB images and depth, respectively, or use expensive post-processing steps that limit their performance in highly cluttered scenarios and real-time applications. Therefore, in this paper, a method is proposed to estimate the 6D attitude of a group of known objects from RGB-D images. The network used is a heterogeneous architecture, which processes two data from different dimensions separately, and uses a new dense fusion network to extract pixel-level dense feature embedding, and estimate the target position and attitude from it. In addition, an end-to-end iterative attitude optimization method is integrated to further improve attitude estimation, while real-time visualization of point clouds and rendering of bounding boxes are realized. Experimental results show that this method is effective on open source data sets.

1. Introduction

The 6D pose estimation of the rigid body refers to the 6D pose estimation of the object in the camera coordinate system, that is, 3D position and 3D attitude. At this time, the coordinate system of the original object itself can be regarded as the world coordinate system, that is, the RT transformation between the world system of the original object and the camera system can be obtained. The significance of 6D pose estimation of rigid body is to obtain the precise pose of the object and support the fine operation of the object. If you want to interact with objects in the environment, 6D pose estimation of objects is a necessary technology, and will continue to become a research hot spot.

According to the input data used, the 6D pose estimation of rigid body can be divided into 2D image-based method and 3D point-cloud based method. The early 6D pose estimation method based on 2D images deals with objects with rich textures. By extracting salient feature points, the descriptor with strong characterization is constructed to obtain matching point pairs, and the PNP method is used to recover the 6D pose of the object. For weakly textured or non-textured objects, the template-based method can be used to retrieve the corresponding 6D pose of the most similar template image, or the optimal pose can be learned by voting method based on machine learning.

With the emergence of cheap Depth sensors represented by Kinect in 2011, 2.5D Depth images can be obtained while acquiring RGB images, which can assist the method based on 2D images. In order not to be affected by the texture, it can also be operated only in 3D space. At this point, the problem becomes a part-to-whole registration problem from the acquired single-view point cloud to the existing complete object point cloud. If the geometrical details of the object are rich, the salient 3D feature points can be extracted, and the characterization descriptor can be constructed to obtain 3D matching points, and the initial pose can be obtained by least squares. Random sampling point
consensus algorithm (RANSAC) can also be used to obtain a large number of candidate 6D poses and select the pose with the smallest error. Since 2012, deep learning has taken the lead in the field of 2D vision and introduced deep learning into 6D pose estimation of objects. It has shown excellent performance no matter based on pure RGB images, RGB and Depth images, or only based on 3D point cloud.

In the process of planetary surface exploration, it is necessary to first understand the terrain. For example, rocks will become obstacles during the driving of the rover. How to accurately and effectively obtain the 6D pose of the target through image analysis in advance, use the robotic arm to carry out grasping experiments, and avoid obstacles in advance is an important link in the deep space exploration mission. It is difficult to obtain accurate data of the position and attitude of small targets, because in the deep air, the remote control has great uncertainty factors, and the two-dimensional image changes greatly under the influence of illumination. In a relatively complex situation (such as occlusion and low illumination), the attitude estimation of known objects can only be carried out by combining color images and depth images. However, these two kinds of images come from different Spaces and dimensions. For data from such different sources, feature fusion is a challenge faced by current technologies.

Extraterrestrial spacecraft missions, the author of this paper, the advantages of high precision requirements, target shape is irregular, geared to the needs of poor soil, metal minerals, irregular rock, motion detectors, such as specific targets, analysis of target image and multiple points of view of point cloud data correlation characteristics, design based on multi-source information of end-to-end independent positioning neural network architecture, put forward the characteristics of dense pixel level fusion detection target space positioning method, break through the perception of the image characteristics and depth information fusion target location technology, multi-source perception to detect targets under the dynamic scene space positioning. Therefore, in this paper, the feature vectors of RGB images and point clouds are fused at the level of each pixel. An iterative optimization method is used for end-to-end training, and the pose estimation of the target is achieved quickly.

2. Introduction of method
The algorithm was programmed by using PyCharm and VS in Python and C++. OpenCV, developed by the Visual Interaction Group of Intel Microprocessor Research Laboratory, has built-in frame extraction functions and standard image processing algorithms for various forms of image and video source files.

3. Theory and Methods
The whole network framework structure is mainly divided into two stages: 1. The target that needs attitude estimation is separated by semantic segmentation, and the position of the target in the depth map is determined by using the mask of semantic segmentation, and then the depth information of the target area in the depth map is transformed into point cloud information. 2. Perform attitude estimation according to the input 2D image and 3D point cloud\(^1\). The implementation method is as follows:

1. Semantic segmentation

Semantic segmentation network is a CODEC structure which takes image as input and generates semantic segmentation graph of N+1 channel. Each channel is a binary mask in which the active pixel describes an object in each of the N possibly known classes.

2. Dense feature extraction
First, the segmented depth pixels are converted into three-dimensional point clouds using known camera internal functions, and then a point-net-like structure is used to extract geometric features. A geometric embedded network is proposed to generate dense point features by mapping each P-split point to an \(d_{\text{geo}}\) -dimensional feature space. According to the known camera internal parameters, it is converted to point cloud data. Then it is processed through a network similar to the point net\(^2\). Implement a variation of the point-net architecture that uses average pooling as a symmetric reduction
The goal of color embedded network is to extract the features of each pixel so as to form a dense correspondence between 3D point features and image features. The image embedding network is a CNN-based CODEC structure. Each pixel of an RGB image is mapped to another space through the CNN network. It maps an image of size $H \times W \times 3$ to an embedding space of $H \times W \times d_{rgb}$. Each embedded pixel is an n-dimensional vector that represents the appearance information of the input image at the corresponding position. The input of the network is $3^{HW}$, the output is $rgb^{HWd}$, and each pixel of the input image corresponds to $d_{rgb}$.

(3) Pixel-level intensive fusion

This paper describes a new pixel-level dense fusion network, whose principle is mainly based on one-to-one mapping from 2D images to 3D point clouds. Firstly, the known camera parameters are projected onto the image plane, and the geometric features of each point are correlated with the corresponding image feature pixels, so as to establish a good relationship between the three-dimensional point cloud and the corresponding position in the two-dimensional image. Then the obtained features are linked together and provided to other network processing, and symmetric reduction function is used to generate a global feature of fixed size. The global dense fusion feature is used to enrich the feature of each dense pixel and enrich the global semantic information. Each pixel feature is input into the final network to predict the 6D pose of the object.

(4) 6D target pose estimation

The loss function $\text{Loss}$ defined is the point cloud data of the current frame obtained by the target model according to camera parameters (rotation matrix $R$ and deflection matrix $T$) saved when photographing pictures from other perspectives. The loss function is defined as follows:

$$L_i^p = \frac{1}{M} \sum_j \| (Rx_j + t) - (\hat{R}_i x_j + \hat{t}_i) \|$$

(1)

Each point cloud selected by the target model and all predicted attitude parameters need to be transformed respectively. Finally, the combined loss value is minimized.

The above objects well define the learning method for well-textured, asymmetrical objects. But for symmetrical, non-textured objects, such a loss definition will cause fuzzy learning of the network. For symmetrical objects, use the following method:

$$L_i^p = \frac{1}{M} \sum_{j \in [0, M]} \min_k \| (Rx_j + t) - (\hat{R}_i x_j + \hat{t}_i) \|$$

(2)

In the process of network learning, the losses are optimized to minimize the total losses:

$$L = \frac{1}{N} \sum_i L_i^p$$

The confidence of each pixel is also predicted when the attitude is estimated. Therefore, when calculating loss in each point cloud, add a weight as follows:

$$L = \frac{1}{N} \sum_i (L_i^p c_i - \omega \log(c_i))$$

(3)

(5) The iterative optimization

The previously predicted attitude is taken as the estimation of the standard frame of the target object, and the input point cloud is transformed into the standard frame of the estimation. Then, the converted point cloud is fed back into the network and a residual attitude is predicted based on the previously estimated attitude. A dedicated pose residual estimator network is trained to refine the initial pose estimation for a given main network. In each iteration, the image features embedded from the main network are reused and the geometric features calculated from the newly transformed point cloud are closely fused. The attitude residual estimator uses the global feature of the fusion pixel feature set as input. After K iterations, the final attitude estimation is obtained as the series of each iteration estimation.
A residual network is constructed to refine the attitude of the initial estimation of the backbone network. After K iterations, the final attitude is as follows:

$$\hat{p} = [R_K | t_K] \cdot [R_{K-1} | t_K] \cdots [R_0 | t_0]$$ (4)

4. Experimental simulation results and analysis

The accuracy of pose estimation algorithm was tested with Linem od data set, and the target pose point cloud and pose data were obtained, and the bounding box was drawn.

The training time was 3 days, 02h, 53m, and the 104 epoch was trained. The results were as follows:

Train time 02h 53m 23s Epoch 104 Batch 9312 Frame 37248 Avg_dis:0.001808212633477524

The test results using the test set are as follows:

- Object 1 success rate: 0.9313632030505243
- Object 2 success rate: 0.9437439379243453
- Object 4 success rate: 0.9696078431944349
- Object 5 success rate: 0.9448818897637795
- Object 6 success rate: 0.9670658682634731
- Object 8 success rate: 0.8721506442021804
- Object 9 success rate: 0.9342723004694836
- Object 10 success rate: 0.9981185324553151
- Object 11 success rate: 0.9980694980694981
- Object 12 success rate: 0.9295908658420552
- Object 13 success rate: 0.97859540347931
- Object 14 success rate: 0.9702495201535508
- Object 15 success rate: 0.9586935638808838
- ALL success rate: 0.953595941509997

The visualization effect of real-time target point cloud and prediction point cloud using VISDOM library is as follows:

![Fig. 1 Real-time 3D display of target point cloud and predict point cloud](image)

Use Open3D to save the predict point cloud as a PCD file and draw the bounding box. The effect is as follows:
Fig. 2 Visualization of target object prediction point cloud and 3D bounding box

Use the PCL library of VS to calculate the pose information, as shown below:

Therefore, the position and pose of the surface target can be captured and sampled by the robotic arm of the rover.

5.Conclusion
Aiming at target recognition and pose estimation of planetary surface, this paper introduces a neural network deep learning and artificial intelligence method of target 6D pose estimation combining two-dimensional and three-dimensional images. Based on the original method[5], it improves the point cloud and related information to visualize in real time and display the bounding frame and its coordinates. Experimental results show that this method can identify the target’s 6D pose and display its enveloping frame[6], which provides a basis for target pose estimation and subsequent robotic arm capture in deep space exploration missions. The accuracy and effectiveness of the method are verified on the open source data set, which will lay a foundation for the rover to identify the surface obstacles in the future deep space exploration missions.

Acknowledgments
We would like to thank the anonymous reviewers for their helpful remarks. We thank the associate
editor and the reviewers for their useful feedback that improved this paper. This work is supported by Autonomous Intelligent Precision Sensing and Control of Unmanned System for Extraterrestrial Exploration (SQ2018AAA010231).

**Reference**

[1] Li Weiqing. Multi-pose face recognition based on manifold learning and tensor [D]. Huaqiao University.

[2] Qiu H. Active network technology and its application in network management [D]. University of Electronic Science and Technology of China

[3] Zhang Qian. Mechanism analysis of the effect of water on asphalt mixture [D]. 2007.

[4] Zhu Wenjuan. Research on Hand Posture Estimation Based on Feature Fusion Network [D]. 2019.

[5] Chen Wang, Danfei Xu, Yuke Zhu, and Roberto Marté Pissive Friction, "Dense Fusion: 6D Object Pose Estimation by Iterative Dose Fusion," In Computer Vision and Pattern Recognition (CVPR), 2019

[6] Wang Lu, Zhang Daoguang, Cai Zixing. Dynamic scene recognition combining prominent features and moving object filtering [J]. Journal of Journal6, 2010(26):152-154.