DOI: 10.54254/2753-8818/51/2024CH0185

Analysis of modeling nasal narrow space based on visual slam

Changru Li

School of Mechanical and Electric Engineering, Soochow University, Suzhou, China

2129403068@stu.suda.edu.cn

Abstract. Now, surgeries are becoming more stable, safe, efficient and low-cost. During the surgical treatment of nasal diseases, surgical robotic robotics can help operate accurately and reduce the discomfort after surgery. However, due to the internal space of the nasal cavity being relatively narrow, it is difficult for the nasal surgical robot to contain multiple vision sensors and the monocular camera could not get information about the depth of the 3D objects in the scene, so the existing surgical robots cannot accomplish the three-dimensional modeling about the internal space of nasal cavity well. In practice, doctors still have to analyze pictures from the robotic, which may decrease the efficiency of the surgery and increase the risk to patients. This article designed a SLAM algorithm framework based on a depth estimation network, it can simulate the internal structure of the nasal cavity more accurately through pictures, which come from monocular endoscopic on surgical robotic. The insights gained in this study verify that the method of image segmentation can also make the depth representation of the nasal internal space more accurate and this method may help robots realize their self-position in the narrow area of the nasal cavity, which lays the foundations for the development of fully autonomous surgical robots.

Keywords: Nasal surgery, visual SLAM, depth estimation, narrow space, feature extraction.

1. Introduction

Minimally invasive surgery is a multi-person indirect hand-eye coordination process with the help of an endoscopic imaging system. In Table 1, the control group is traditional surgery and the observation group is endoscopic surgery [1]. The data in the table were analyzed by t test, and the data contained statistical significance when p≤0.05.

Table 1. Comparing two groups of clinical indicators(Δ±s).

Group	Number of case	Time of operation (min)	The nasal ventilation time (d)	Hospital stays (d)	Preoperative bleeding (ml)	The mucosa recovery time (d)
Observation group	62	26.38±5.14	3.74±1.05	4.65±1.22	32.68±3.11	5.34±1.31
Control group	62	35.76±5.21	4.59±1.23	6.78±1.54	46.57±4.19	7.65±2.04
t		10.092	4.139	8.537	20.960	7.502
P		0.000	0.018	0.002	0.000	0.008

It can be seen from the table that compared with traditional surgery, minimally invasive surgery has demonstrated advantages in all surgical parameters. Due to its advantages of light pain, small incisions during operation, and quicker recovery after operation, it was soon applied to various surgeries. However, the nasal cavity is filled with delicate facial nerves and numerous blood vessels, requiring high positioning precision. Sparse texture features and a large amount of bodily fluids further complicate visual sensor-based image construction algorithms during surgery. These challenges prompting the application of Simulation Localization and Mapping technology in endoscopic fields.

In recent years, the research of visual SLAM has become more popular. Although computer vision technology has developed for nearly half-century, the applications in endoscopy can still be considered a new field. By searching the articles on endoscopic visual processing in the SCI Extended database, only less than 15% of the articles are about the endoscopic visual SLAM method and related 3D reconstruction research [2]. So, the research institutions in endoscopic visual processing-related research are still in their infancy.

The narrow space in the nasal cavity limits the types of sensors that can be carried by the end effector of surgical robots. Because of this, necessitates the use of a monocular endoscope for visual imaging in endoscopic surgical robot design and depth estimation become the primary problem in 3D scene reconstruction. Some classical algorithms for monocular endoscope depth estimation include Structure from Motion method and Shape From Shading method [3, 4]. These methods recover shapes from motion and shadows, but they only have low precision and cannot meet practical requirements. Nowadays, deep learning has been increasingly applied in various industries, containing different network models such as deep belief networks, Auto Encoder (AE) networks, and convolution neural networks. With the help of the powerful modelling advantages of deep learning and the strong learning capabilities of data, deep learning networks can accomplish accurate depth estimation using monocular images [5]. However, encoder networks have a significant distortion in the depth prediction during down sampling in 3D reconstruction. To address this issue, research fused shallow features, which contain more detailed information with deep features [6]. While these methods demonstrate good performance in reducing depth errors, they are limited by scene instability and may lead to incorrect estimations of relative depths within 3D scenes or exhibit poor predictive performance on certain datasets containing multiple images spanning an entire plane [7].

This article addresses the problem of edge and inaccurate depth estimation of the maximum depth area in monocular images and proposes a method for monocular depth estimation based on Pyramid Split attention Network (PS-Net). First, PS-Net is based on boundary guidance and scene aggregation networks and introduces the Pyramid Split Attention (PSA) module to process multi-scale feature spatial information and extract some areas where depth gradient changes intensely or have the maximum depth. Then, the Mish function is used as the activation function in the decoder to further improve the performance of this network. Finally, the network is trained and evaluated on the SurgT and iBims-1 datasets.

2. Overall design scheme

The entire research and experimental plan for surgical robots requires strict adherence to relevant ethical and safety regulations. Operations involving the human body necessitate prior approval from the government and full informed consent from the patients. Therefore, in designing the experiments, this research chooses a combined approach of simulation and data verification. The safety and feasibility of the experiment were tested by using simulation software such as Solid Works and Matlab, and the real-world operations used animal experiments to conduct.

Firstly, this article used the publicly available SurgT datasets to collect and analyze the anatomical structures of the nasal cavity in the human body and then create a proprietary dataset specific to this area. This dataset includes the nasal cavity internal structures, easily reflective areas, flexible areas and weak feature areas. These features were utilized to optimize the design of the structure in the surgical robot arm end, resulting in a more suitable and reasonably flexible endoscopic surgical robot arm, which will be used in the human body and surgical operations. Subsequently, Solid Works will be used to model the arm and test the rationality and feasibility of its design. After that, 3D simulation models will be employed to test the flexibility of the robotic arm and the efficiency of image capture.

In the implementation of the SLAM system framework, the RGB image is first input into the network for simple processing in advance. Then, a newly designed network is trained to learn features, which is based on the structure of boundary guidance and scene aggregation networks. This network introduces the Pyramid Split Attention (PSA) module and effectively enhances network performance by employing a new Mish activation function in the decoder [8]. Following this, the depth information of images is predicted by a stripe refinement module through learning features. Finally, the visualization output of predicted depth maps is generated.

Last, this article collects and statistically analyses the data and results in the experiments to adjust the scheme of experiments, and compares them with data obtained from other existing SLAM modelling software. In the quantitative analysis of result comparison, accuracy rate, recall rate, and a comprehensive parameter are used as parameters to assess the performance and accuracy of the depth estimation framework.

3. Algorithm network structure

Convolution neural networks have been widely used in image recognition, because of their better accuracy and higher robustness. In this section, a brief summary of depth estimation networks will be given.

The supervised depth estimation network is learned and trained by taking input RGB images with their real-world depth parameters, resulting in the optimal network model which can predict the depth of the monocular endoscope picture more accurately. This model can be used to obtain the depth values of new images when they are inputted into it. PS-Net is built upon boundary guidance and scene aggregation networks based on the traditional encoder-decoder structure [7].

This article introduces the Pyramid Split Attention module on this traditional structure and replaces the ReLU function in the decoder with the Mish activation function [8]. The network adopts ResNet50 (Residual Network) as the basic network and uses dilated convolution to replace the original 3*3 convolution in the 4th and 5th down sampling stages in order to obtain a larger receptive field about the picture and reduce computational cost. Furthermore, information from shallow layers is progressively put into deeper layers of the Bottom-Up-Top-Down Fusion (BUBF) modules through the networks, facilitating extraction of depth variations during down sampling and details present in high-resolution areas. Then the output of BUBF modules will be input to the SR (Super-Resolution) modules for further processing.

The depth predicted module takes the output of the encoder as input and utilizes the dilated convolution and scene encoder of the Pyramid Split Attention module to capture long-distance pixel and multi-scale region correlations, and then they are put to the decoder. The decoder mainly works through four steps. The first two steps involve compressing channels and maintaining resolution by using the Large-Kernel Refinement Blocks (l-RB). In this article, the l-RB employs the Mish activation function to work. The last two steps employ a combination of l-RB and up sampling, similar to the upward projection method which is described in Laina and team’s paper [9].

The PSA module uses a 114*152*64 image as an input, which is obtained through down sampling of the datasets. In order to achieve better results in deep prediction, it is necessary to input shallow images to represent more details in the feature map. The output is combined with BUBF and sent into the SR module. In the SR module, the output in decode is fused with that in BUBF and PSA modules, then generates the final required depth picture. This research indicates that depth pictures obtained using this method can exhibit fewer depth errors.

4. PSA module

In the process of depth estimation, the first step is collecting spatial feature points. This is achieved by utilizing visual sensors to capture the three-dimensional coordinates of these pictures and employing image processing algorithms to detect and extract these coordinates. This article integrated a pyramid slip attention module into the traditional extraction stage. This module enhances inter-scale and cross-channel information correlation through the utilization of multi-scale convolution kernels and global average pooling operations, because of this action the accuracy of predictions for edge and farthest distance areas has been improved [10].

The execution plan of this module is approximately three steps. Firstly, the input image is split it into S groups based on channels by using the Split and Concat module (SPC). Different-sized convolution kernels are then used to obtain feature maps with multi-scale information at the channel level. Subsequently, the output of the split and fusion module is put into a weight module to derive weights for different channels, thereby obtaining attention weights for each level's feature map, and then normalising the attention weight values for each group. Finally, through these operations, multi-scale information is integrated with cross-channel attention, and then they are put into every block which is split from feature groups to improve pixel-level attention. In order to produce improved pixel-level attention, multi-scale spatial information and cross-channel attention are ultimately incorporated into blocks of each split feature group using these techniques.

In order to achieve different spatial resolutions and depths, the input pictures are split into S groups based on the different channel levels, with S=4. Each groups will be processed by convolutions of varying scales, the scales of these convolutions can be expressed by formula: $\mathrm{K_i}$ = 2×(i+1)+1 (i=0,1,2....,S-1), Where the i is the number of the split group, in order to obtain a larger receptive field. To reduce the increased computational complexity resulting from larger convolution kernel sizes, this research applies the feature grouping convolutions in each group, with the number of groups as $\mathrm{G} = 2^{\frac{\mathrm{K_i}-1}{2}}$. Finally, the feature maps of different scales are fused together in the channel dimension using the concatenation function as shown in equation (1). $\mathrm{F_n}$ is the corresponding output after each set of convolutions

$$\mathrm{F} = \mathrm{Cat}([\mathrm{F_0}, \mathrm{F_1}, \ldots.., \mathrm{F_{s-2}}, \mathrm{F_{s-1}}]) \tag{1}$$

Subsequently, the output of SPC module is utilized as an input for Squeeze-and-Excitation (SE) weights. This process encodes global information by performing a global average pooling operation on the multi-scale feature maps obtained from the splitting and fusion modules, as described in Equation (2). W is the number of columns of pixels and H is the rows of pixels, respectively. Following this, adaptive channel relationships are calibrated through two fully connected layers and activation functions, as outlined in Equation (3), to acquire channel attention weight information. Finally, the normalized attention values for different channels are integrated with the input of the SE module through weighted fusion.

$$\mathrm{g_c} = \frac{1}{\mathrm{H} \times \mathrm{W}} \times \sum_{\mathrm{i}=1}^{\mathrm{H}} \sum_{\mathrm{j}=1}^{\mathrm{W}} \mathrm{x_c}(\mathrm{i}, \mathrm{j}) \tag{2}$$

$$\mathrm{w_c} = \sigma\left(\mathrm{W_1}\delta\left(\mathrm{W_0}\left(\mathrm{g_c}\right)\right)\right) \tag{3}$$

This article integrates multi-scale spatial information into each split feature group by partitioning the image, thereby achieving the effect of grouping feature point extraction. It establishes a SLAM model framework based on a depth estimation network to accurately assess the depth position of the human nasal cavity during surgery, enhancing the adaptability and accuracy of the SLAM system. The model was tested and validated on a standard dataset, and getting a good result.

5. The operation of experimental

This research operated experiments on an NVIDIA 3060 GPU using the PyTorch framework. This article evaluated the degree of accuracy improvement of the method by operating repetitive experiments and averaging the results on the iBims-1 and NYUD v2 datasets. Training and testing of a new framework were performed on the NYUD v2 datasets, while evaluation took place on the iBims-1 datasets.

The iBims-1 datasets were specifically designed to test the method of monocular depth estimation. These datasets have high-quality pictures and depth maps with good precise because they were captured by digital cameras which only have one lens to reflex. Compared with other datasets, the quality of monocular camera pictures in the iBims-1 datasets was similar to monocular endoscope images. It is important to note that due to variations in the experimental environment, some data in this article may be different from those presented in the original paper.

Based on previous research, this article selected 50K RGB-D images from the NYUD v2 datasets for training and 654 pairs for testing [11]. In this experiment, 20 epochs were set with batch size = 8 and using the Adam optimizer with parameter was (β_1, β_2) = (0.9,0.999) and the weight was declining with 10^4. The initial learning rate was set to 0.0001 and reduced by 10% each 5 epochs.

To train the model, all images and labels were down-sampled from their original resolution of 640×480 to 320×240 using bilinear interpolation and then cropping to a size of 304×228 from the middle [8]. To align the output, the labels which were cropped further down-sampled to the resolution of 152×114. In addition, the output was up-sampled to a resolution of 304×228 pixels in order to evaluate the model.

The model proposed in this article demonstrates superior estimation capabilities for deep prediction in complex backgrounds. As illustrated in the vitro experiment, a scene featuring a bookshelf and a table was chosen, with more complex object placement and clear depth relationships. The original Bs-Net method is revealed with some issues while conducting depth estimation at greater depths and object boundaries. Compare with that, the depth map predicted by the network proposed in this article distinctly presents the scene layout and relative depth. The edges of the entire image are relatively neat and clear.

In the quantitative analysis of result comparison, accuracy rate (4), recall rate (5), and comprehensive parameter (6) are used as the three parameters for analysis. There are three parameters in these calculation formulas as follows: t_p the number of correctly predicted boundary pixels. f_p the number of incorrect boundary predictions that were mistakenly predicted as correct. f_n the number of correct boundary predictions that were mistakenly predicted as incorrect.

When counting the pixels of the number of boundaries, this article sets the threshold value which is t (t ϵ\{0.25, 0.5, 1\}). The threshold is based on the gray histogram of the image to determine the image boundary. When the value of a part of pixels in the gray histogram is greater than the threshold set, it is considered as the depth estimation boundary. The calculation formula is given as follows:

$$P = \frac{t_p}{t_p + f_p} \tag{4}$$

$$R = \frac{t_p}{t_p + f_n} \tag{5}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{6}$$

Based on these three parameters, a rough estimation of the quality of image generation is presented in the following table 2. It can be observed from the table that, compared to the existing methodology, the proposed approach in this article demonstrates significant advantages in accuracy and comprehensive parameters, resulting in good image depth estimation.

Table 2. Comparison of boundary accuracy of different methods.

Method	Accuracy rate	Recall rate	Comprehensive parameter
Threshold value>0.25			
Dharmasiri	0.577	0.626	0.591
This article	0.672	0.527	0.575
Threshold value>0.5			

Table 2. (continued).

Dharmasiri	0.531	0.509	0.506
This article	0.685	0.513	0.592
	Threshold value>1		
Dharmasiri	0.617	0.489	0.533
This article	0.736	0.509	0.617

6. Conclusion

By adding the step of pyramid split attention in feature extraction, the network enhances the information correlation between scales by using multi-scale convolution kernels and global average pooling operation, which makes the network extract multi-scale information with better consistency. It shows good performance in edge areas and the deepest depth areas.

However, when extracting features from images, the pyramid split attention module is required to improve the quality of pictures. Compared to the original approach, although the proposed solution has improved image quality to some extent, it requires a longer response time and may cause delays in practical operations. In the experimental process, despite achieving favorable results in vitro environment due to limitations of the experimental equipment only general vitro datasets could be utilized for verification. Because of this, there were constraints on conducting in vivo cavity experiments during the validation phase.

In the future, it still needs to be further verified by in vivo experiments or using models. Furthermore, incorporating split and extraction during image processing led to process slowly and significant delays during practical operations. Therefore, it is necessary to improve the image response time for actual applications.

The quantitative analysis of the experimental data proves the Pyramid Split Attention network can better estimate the depth of low-feature pictures in the reconstruction of monocular camera images. This model is very suitable for application in the narrow and less feature information parts such as the nasal cavity. It is of great significance in the picture's reconstruction of the human endoscope.

References

[1] Jianfei Shi & Jia Li. (2018). Efficacy analysis of functional endoscopic sinus surgery in the treatment of chronic rhinosinusitis with nasal polyps. Hebei Medical Journal (03),437-441.

[2] Xiaoyu Peng. (2017). Master of endoscopic visual SLAM method research in minimally invasive surgery (Dissertation of degree, University of Electronic Science and Technology). Master's degree https://kns.cnki.net/kcms2/article/abstract?v=gisQO9UvOsYV4fM8BvH7T7JBrmjZDKUqfRT-FFavuxS98BUi66bdINY6bJvDDJsQadDssVlPZggC0x5Tt0haY7mRQK8nu1N_SxMkknPJNbZGmgm4EF13dJi8BiUkiW2WTTPAg_SwoXHpCuO6L8_aaq8Dh_uFLDkjua_G8yNk-PQJLmOVt7SVhyu5TXFw0Gg4tzo863CT3gc=&uniplatform=NZKPT&language=CHS

[3] Snavely, N., Seitz, S. M., & Szeliski, R. (2008, June). Skeletal graphs for efficient structure from motion. In 2008 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). IEEE.

[4] Zhang, R., Tsai, P. S., Cryer, J. E., & Shah, M. (1999). Shape-from-shading: a survey. IEEE transactions on pattern analysis and machine intelligence, 21(8), 690-706.

[5] Tianteng Bi, Yue Liu, DongDong Weng & YongTian Wang. (2018). A review of single image depth estimation based on supervised learning. Journal of Computer Aided Design and Graphics (08),1383-1393.

[6] Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., & Gaidon, A. (2020). 3d packing for self-supervised monocular depth estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2485-2494).

[7] Xue, F., Cao, J., Zhou, Y., Sheng, F., Wang, Y., & Ming, A. (2021). Boundary-induced and scene-aggregated network for monocular depth prediction. Pattern Recognition, 115, 107901.

[8] Wenju Li, Mengying Li, Liu Cui, Wanghui Chu, Yi Zhang & Hui Gao. (2023). Monocular depth estimation method based on pyramid split attention network.Journal of Computer Applications(06),1736-1742.

[9] Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., & Navab, N. (2016, October). Deeper depth prediction with fully convolutional residual networks. In 2016 Fourth international conference on 3D vision (3DV) (pp. 239-248). IEEE.

[10] Zhang, H., Zu, K., Lu, J., Zou, Y., & Meng, D. (2021). Epsanet: An efficient pyramid split attention block on convolutional neural network. arXiv preprint arXiv:2105.14447.

[11] Chen, X., Chen, X., & Zha, Z. J. (2019). Structure-aware residual pyramid network for monocular depth estimation. arXiv preprint arXiv:1907.06023.