Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

A three-dimensional measurement method for binocular endoscopes based on deep learning*

Hao YU[1], Changjiang ZHOU[2], Wei ZHANG[1], Liqiang WANG[1,2], Qing YANG[1,2], Bo YUAN[†‡1]

[1]*State Key Laboratory of Modern Optical Instrumentation, College of Optical Science and Engineering, Zhejiang University, Hangzhou 310027, China*

[2]*Research Center for Intelligent Sensing, Zhejiang Lab, Hangzhou 311100, China*

[†]E-mail: yuanbo@zju.edu.cn

Received Dec. 7, 2020; Revision accepted Mar. 22, 2021; Crosschecked Sept. 26, 2021; Published online Jan. 24, 2022

Abstract: In the practice of clinical endoscopy, the precise estimation of the lesion size is quite significant for diagnosis. In this paper, we propose a three-dimensional (3D) measurement method for binocular endoscopes based on deep learning, which can overcome the poor robustness of the traditional binocular matching algorithm in texture-less areas. A simulated binocular image dataset is created from the target 3D data obtained by a 3D scanner and the binocular camera is simulated by 3D rendering software to train a disparity estimation model for 3D measurement. The experimental results demonstrate that, compared with the traditional binocular matching algorithm, the proposed method improves the accuracy and disparity map generation speed by 48.9% and 90.5%, respectively. This can provide more accurate and reliable lesion size and improve the efficiency of endoscopic diagnosis.

Key words: Binocular endoscope; Three-dimensional measurement; Deep learning; Disparity estimation
https://doi.org/10.1631/FITEE.2000679 **CLC number:** TN29; TP391.4

1 Introduction

It is tough to perceive spatial positioning with conventional endoscopes that provide only two-dimensional (2D) images, which will eventually affect the accuracy and safety of diagnosis and treatment. The binocular endoscope has the ability of three-dimensional (3D) imaging and measurement, which can provide depth information to assist endoscopists in operating the endoscope more accurately, efficiently, and safely (Ogino-Nishimura et al., 2015; Cai et al., 2018; Nomura et al., 2019; Omori et al., 2020). In clinical gastrointestinal endoscopy, the precise size of lesions such as polyps is of great significance for diagnosis (Furukawa et al., 2014; Dimas et al., 2020). When operating 2D endoscopes, endoscopists mainly use subjective vision or specific measurement tools, such as endoscopic measurement rulers, to estimate the polyp size. Consequently, the endoscopist's subjective assessment often deviates greatly from the accurate value due to polyp morphology and optical distortion (Ahmad et al., 2016; Anderson et al., 2016; Shaw and Shaukat, 2016; Sakata et al., 2018) and because endoscopic measurement rulers produce inaccurate results and are inefficient (Anderson et al., 2016). Therefore, it is necessary to develop an objective, accurate, and efficient 3D measurement method to improve endoscopic surgery results (Furukawa et al., 2014).

‡ Corresponding author

* Project supported by the National Key Research and Development Program of China (No. 2019YFC0119502), the Key Research and Development Program of Zhejiang Province, China (No. 2018C03064), the Fundamental Research Funds for the Central Universities, China (No. 2019FZA5016), and the Zhejiang Provincial Natural Science Foundation, China (No. LGF20F050006)

ORCID: Hao YU, https://orcid.org/0000-0001-9984-5051; Bo YUAN, https://orcid.org/0000-0002-3185-2690

At present, the binocular endoscope realizes 3D measurement mainly by determining the disparity between the left and right images using the binocular matching algorithm (Wang D et al., 2018) and calculating 3D coordinates of each pixel by triangulation combined with the binocular camera parameters. Traditional binocular matching algorithms, such as the semi-global matching (SGM) algorithm (Hirschmüller, 2008), have a better matching result on images with rich texture information. However, most endoscopic images have inadequate areas that are texture-less or have a high gloss. The robustness of the SGM algorithm in such an area is poor, and its generated disparity map often has noise and mismatching errors, which will eventually affect the accuracy and stability of 3D measurement. To overcome the shortcomings of traditional binocular matching algorithms, researchers began to explore the binocular disparity estimation model based on deep learning. Some proposed models (Žbontar and LeCun, 2014; Kendall et al., 2017; Zhang et al., 2019) can effectively improve the accuracy of disparity predictions in texture-less and high-gloss areas, and produce excellent results in the field of autonomous driving. However, because their application scenarios are quite different from the endoscopic environment, they are ineffective when directly applied in binocular endoscopic images. In addition, binocular endoscopic images with ground truth disparity are difficult to obtain, which limits the transfer learning ability of these models.

In an effort to solve the above-mentioned problems, we propose a 3D measurement method for binocular endoscopes based on deep learning. A simulated binocular image dataset is created for a gastroscope scenario to train a disparity estimation model for 3D measurement. The experimental results demonstrate that the proposed method is superior to the traditional binocular matching algorithm in terms of stability, accuracy, and real-time performance.

The contributions of this work are summarized as follows: (1) Given that the ground truth disparity of real binocular endoscopic images is difficult to obtain, we propose a new method for creating simulated binocular images that are very similar to real endoscopic images in terms of content and structure. (2) An end-to-end disparity estimation network, mainly based on Google's StereoNet (Khamis et al., 2018) with minor adjustments, is trained on virtual binocular images and used for estimating the disparities of real binocular endoscopic images. The 3D measurement result calculated from the estimated disparities is more stable and reliable.

2 Three-dimensional measurement method based on deep learning

The process of the proposed 3D measurement method based on deep learning is shown in Fig. 1a. The left and right images taken by the binocular camera of the 3D endoscope are used as the input of the disparity estimation model, which will output the disparity map corresponding to the left image. Then, 3D coordinates of each pixel in the disparity map in the camera coordinate system can be acquired by triangulation combined with the binocular camera parameters. Finally, the target size is calculated using the 3D coordinates. The training process of the disparity estimation model is shown in Fig. 1b. A 3D scanner is used to scan the stomach model to obtain a 3D model file, which is then imported into the 3D rendering software. By using the simulated binocular system established in the 3D rendering software, a simulation dataset is created to train the deep convolutional network for disparity estimation.

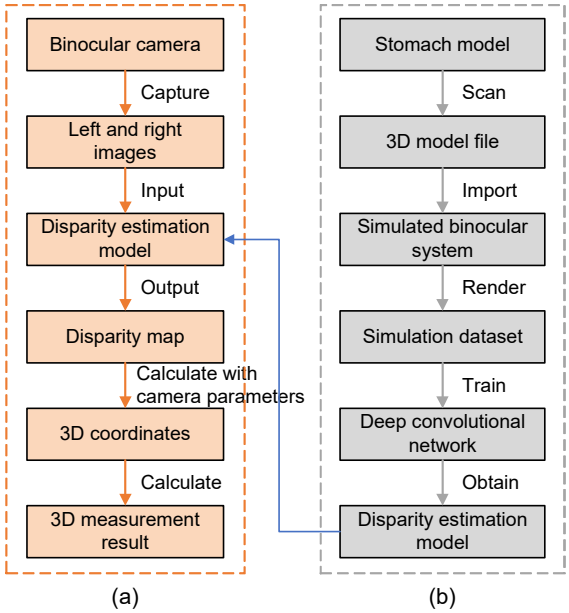

Fig. 1 The overall process of our 3D measurement method (a) and the training process of disparity estimation model (b)

2.1 Simulation dataset

Training a deep learning network requires a lot of data with labels. However, it is difficult to obtain endoscopic binocular images with ground truth disparity, which limits the transfer training, and application of deep learning models in the actual endoscopic environment. Therefore, a method for creating a simulated binocular endoscopic dataset is proposed in this paper.

In previous research (Wang XZ et al., 2020), Blender, an open-source 3D animation production software, was used for building a simulated gastrointestinal model to generate a binocular endoscopic dataset. However, because the gastrointestinal model established by them is a pure virtual model with some disadvantages such as limited texture, the dataset was not sufficiently similar to the real gastroscopic image. By contrast, the simulated binocular dataset created in this study is more similar to the images taken by actual binocular endoscopes. The specific generation process is as follows: First, a 3D scanner is used to scan the actual stomach model to obtain its 3D model file, which is then imported into Blender. After that, a simulated binocular camera system similar to the actual 3D gastroscopic binocular camera is established in Blender. Next, the relative position of the simulated binocular camera to the scanned stomach model is set close to the working state of the actual 3D gastroscope. Finally, the binocular image dataset with ground truth disparity is generated through the Blender rendering. Figs. 2c and 2d show the left and right images taken by the actual 3D gastroscope, respectively. Figs. 2e and 2f show the left and right images from the simulation dataset, respectively. It can be seen that the simulated images are very similar to the real images in terms of the gastric rugae texture and the disparity relationship. In particular, because only the depth of the scanned stomach model relative to the binocular camera can be obtained from Blender, the disparity d needs to be calculated as

$$d = fb / z, \tag{1}$$

where f is the focal length of the camera, b is the baseline of the binocular camera, and z is the distance from the scanned model to the plane of the camera's optical center. The simulation dataset results are shown in Fig. 3, in which the image resolution is

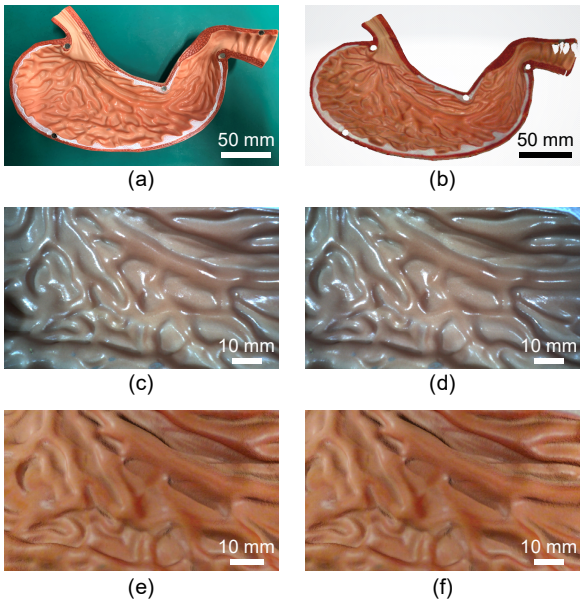

Fig. 2 Comparison of real and scanned models: (a) actual stomach model; (b) image obtained using the 3D scanned stomach model; (c) left image obtained using the actual stomach model; (d) right image obtained using the actual stomach model; (e) left image obtained using the scanned model; (f) right image obtained using the scanned model

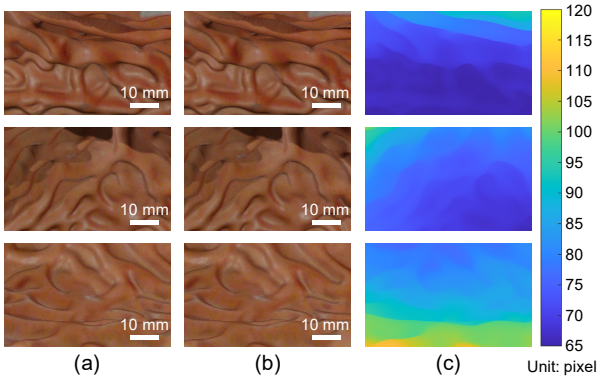

Fig. 3 Examples of the simulation dataset: (a) left camera image; (b) right camera image; (c) ground truth disparity

1280×800, and the disparity image corresponds to the left image.

2.2 Disparity estimation network

Our end-to-end disparity estimation network is mainly based on the StereoNet proposed by Google researchers (Khamis et al., 2018), which has the advantages of good performance in the weak area, low structure complexity, and high computational efficiency. The pipeline mainly includes the feature extraction network, cost volume filter network, and

details refinement network (Fig. 4). Here, we do not use the Siamese network in the cost volume filter and refinement stage, but only generate the disparities corresponding to the left image.

At the feature extraction stage, we use the Siamese network, which first employs three 5×5 convolution layers with a stride of 2 to down-sample the input images and then employs 6 residual blocks composed of 3×3 convolution, batch regularization, and leaky ReLu activations, to extract the features.

A rough cost volume is obtained by subtracting feature vectors of the binocular images. Next, it is filtered through four 3D convolutions with a size of 3×3×3, batch normalization, leaky ReLu activations, and one final 3×3×3 convolution layer. Then the coarse disparity map is generated from the disparity with the minimum cost at each pixel using differentiable arg min.

At the disparity optimization stage, the bilinear interpolation method is used to up-sample the coarse disparity map, and then the color information of the input RGB image is used for refining the high-frequency details. The details refinement network takes the concatenated color and disparity as input and generates the 32-dimensional tensor through a 3×3 2D convolution. Then the tensor is passed through 6 residual blocks in which the dilation convolution is used and the dilation is set as (1, 2, 4, 8, 1, 1). By applying a final 3×3 2D convolution layer, we finally obtain the refined disparity map.

To adapt to the endoscopic images and improve the accuracy of disparity prediction, we adopt the invariant loss function proposed by other researchers (Wang XZ et al., 2020). The loss function is defined as follows:

$$L=\frac{1}{n}\sum\nolimits_i p_i^2-\frac{1}{2n^2}\left(\sum\nolimits_i p_i\right)^2, \tag{2}$$

$$p_i=\ln d_i-\ln d_i', \tag{3}$$

where n is the total number of pixels in the disparity map, d_i is the predicted disparity, and d_i' is the ground truth value at pixel i.

2.3 Evaluation indicators

To quantitatively evaluate the effect of the disparity map generated from simulated binocular images, end-point-error (EPE) and K-pixel-error (KPE) are adopted as metrics (Wang XZ et al., 2020). EPE is the average Euclidean distance between the estimated disparity and ground truth, and KPE is the percentage of the EPE that exceeds K pixels. EPE and KPE are calculated as

$$\mathrm{EPE}=\frac{1}{n}\sum\nolimits_n|d_i-d_i'|, \tag{4}$$

$$\begin{aligned}&\mathrm{KPE}=\frac{1}{n}\sum\nolimits_n f_i\times 100\%,\\&f_i=\begin{cases}1, & |d_i-d_i'|>K,\\0, & |d_i-d_i'|\le K,\end{cases}\end{aligned} \tag{5}$$

where f_i is counted only when the EPE of pixel i exceeds K pixels. However, in the evaluation of the disparity map generated from real binocular endoscopic images, EPE and KPE are not suitable for evaluating the results because of the lack of ground truth disparity. Therefore, we propose an indicator to evaluate the stability of 3D measurement achieved using the disparity map, which is the standard deviation of the multiple measured lengths for the same 3D

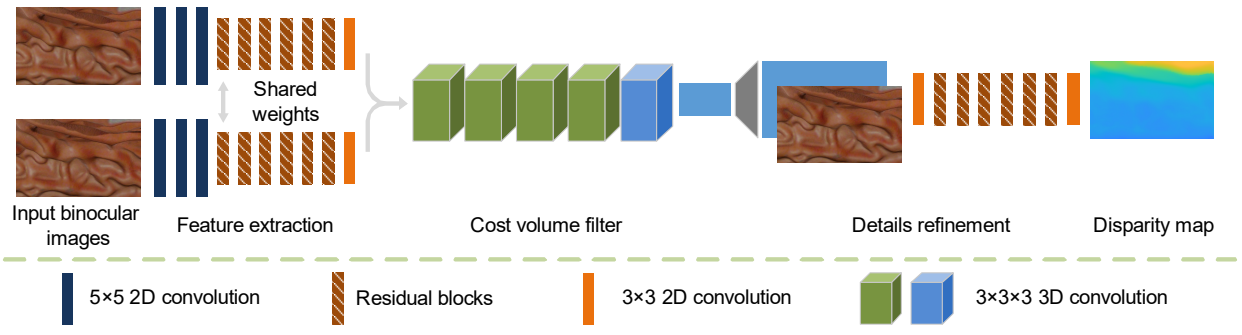

Fig. 4 The network architecture
The binocular images flow into the network and output the disparity map

curve. The specific process is as follows: First, the gastric rugae with clear texture in the left image taken by a 3D gastroscope are selected as the curves to be measured, and the curves are fitted with a constant number of points. Then the 3D coordinates of these fitting points in the camera coordinate system are calculated from the generated disparity map using the following equations:

$$\begin{cases} X = \dfrac{x_i - c_x}{f} Z, \\ Y = \dfrac{y_i - c_y}{f} Z, \\ Z = \dfrac{fb}{d_i}, \end{cases} \tag{6}$$

where x_i and y_i are the x- and y-coordinate of pixel i in the pixel coordinate system, respectively, f is the focal length of the camera, b is the baseline of the binocular camera, d_i is the disparity of pixel i, c_x and c_y are the x- and y-coordinate of the principal point of the camera in the pixel coordinate system, respectively, and X, Y, and Z are the 3D coordinates in the camera coordinate system. Finally, the curve length is acquired from the 3D coordinates of the fitting points.

3 Experiments and results analysis

3.1 Experiment description

The binocular endoscopic images were all taken with a self-developed 3D gastroscope prototype (Fig. 5). The prototype consisted of the 3D gastroscope body, the LED light source, and the signal processing circuit. At the tip of the gastroscope were two cameras, two lights, one forceps channel, and one air-water channel. The calibrated focal length of the binocular camera was 1059.6 pixels, the calibrated baseline was 5.9 mm, and the calibrated x- and y-coordinate of the principal point of the camera in the pixel coordinate system was 633.6 pixels and 367.1 pixels, respectively.

In this study, the Einscan-pro-2x-plus 3D scanner from Hangzhou SHINING 3D Technology Co., Ltd. was used to scan the stomach model. The simulation datasets generated by Blender contained the left image, the right image, and the ground truth disparity corresponding to the left image. Two thousand sets were selected as the training set and the remaining 100 sets as the testing set. The deep learning model used the RMSprop optimizer, and the learning rate was 1e−3. The training process lasted 40 epochs, and the maximum disparity value of the model was set as 192.

A computer with an Intel Core i9-9900K CPU, an NVIDIA GeForce RTX 2080Ti GPU, and 32 GB of memory was used in our experiment. For comparison, we implemented the traditional semi-global block matching (SGBM) algorithm, fine-tuned for optimal performance. The deep learning algorithm and SGBM algorithm were realized by the PyTorch 1.9.1 deep learning framework in Ubuntu 19.04 and Microsoft Visual Studio 2017 in Windows 10, respectively.

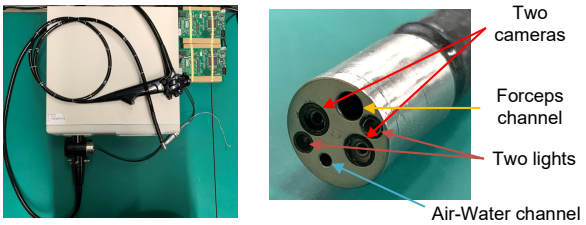

Fig. 5 The 3D prototype endoscope (a) and the tip of the 3D endoscope (b)

3.2 Experimental results and analysis

First, the simulated binocular images were used to evaluate the effectiveness of the disparity map, and the results are shown in Fig. 6. It can be seen from Fig. 6 that the disparity map produced by the SGBM algorithm had a lot of noise in inadequate areas such as shadows, while the disparity map produced by the proposed method was closer to the ground truth disparity map. KPE, EPE, and the single image runtime of each method were calculated as shown in Table 1. Both the KPE and EPE of the proposed method have been greatly improved compared to the SGBM algorithm. The accuracy measured by EPE increased by 48.9%. In particular, when K equaled 50, the KPE of the disparity map produced by the SGBM algorithm and our method were 2.22% and 0.46%, respectively, which shows that the probability that extremely serious errors occurred in our method is much lower

than that with the SGBM algorithm. The runtime of the proposed method was 34 ms, which is 90.5% less than that of the SGBM algorithm. Hence, the proposed method can be used for real-time processing.

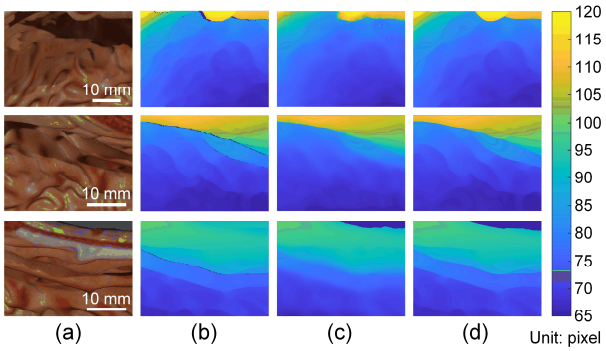

Fig. 6 Disparity results of simulated binocular images: (a) input left image; (b) disparity map produced using the SGBM algorithm; (c) disparity map produced by the proposed method; (d) ground truth disparity map

Table 1 Disparity map performance comparison among different methods

Method	KPE (%)			
	5 pixels	7 pixels	9 pixels	50 pixels
SGBM	2.57	2.48	2.42	2.22
Ours	2.20	1.41	1.07	0.46
Method	EPE (pixel)		Runtime (ms)	
SGBM	2.25		357	
Ours	1.15		34	

Similarly, the real binocular endoscopic images were used to evaluate the effectiveness of the disparity map, and the results are shown in Fig. 7. Due to the more complicated imaging conditions in the actual environment, in the disparity map produced by the SGBM algorithm, a large amount of noise is densely distributed in the shadow and overexposed areas. By contrast, the disparity map generated by the proposed method had less noise and was smoother. In addition, the stability of 3D measurement by the proposed method and SGBM was compared. Three gastric rugae of the stomach model were selected as the curves to be measured (Fig. 8). According to their disparity maps, the length of each curve was measured 10 times, and the corresponding average and standard deviation were calculated as shown in Table 2, where the incorrect measurement results obtained using the SGBM algorithm were highlighted and eliminated when calculating the average and standard deviation. From Table 2, the measurement error rate of the SGBM algorithm was approximately 50%, and the standard deviation was extremely large, which indicates that the noise generated by the SGBM algorithm will seriously affect the stability and reliability of 3D measurement. In contrast, the proposed method had no serious errors and the standard deviation was smaller, which demonstrates that the 3D measurement results of the proposed method are more stable and reliable.

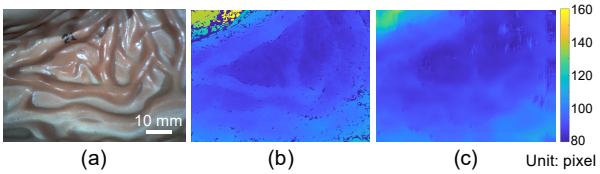

Fig. 7 Disparity results of real endoscopic images: (a) input image (left); (b) disparity map produced by the SGBM algorithm; (c) disparity map produced by our method

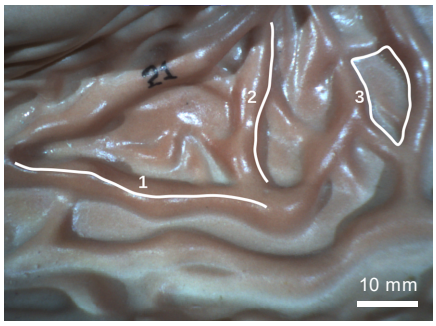

Fig. 8 Three-dimensional measurement curves in the experiment

The proposed method has the potential for further optimization and there are two possible major sources of the remaining errors in our approach: First, due to the unsatisfactory imaging conditions in the actual environment, there is a high probability that the real endoscopic images will contain texture-less areas such as overexposure and shadows, resulting in loss of image features in those local areas. Second, the discrepancy in shallow features, such as color and exposure, between the real endoscopic images and the virtual ones in the training sets affects the accuracy of network prediction.

Table 2 Three-dimensional measurement performance comparison among different methods

Measurement No.	Length (mm)					
	SGBM			Ours		
	Curve 1	Curve 2	Curve 3	Curve 1	Curve 2	Curve 3
1	**14 906.38**	**13 046.76**	44.56	47.09	31.70	53.26
2	46.49	33.96	**14 042.41**	48.24	31.25	53.18
3	44.60	**13 051.37**	47.77	46.85	31.38	51.74
4	**14 920.92**	32.30	**13 693.33**	46.94	31.84	53.07
5	45.60	**13 049.21**	44.28	46.51	31.71	52.46
6	52.70	32.46	**13 774.25**	45.83	31.27	51.99
7	**14 942.16**	32.82	44.07	46.48	31.98	52.88
8	**14 910.01**	**13 046.72**	**14 038.81**	46.77	31.10	53.46
9	43.78	32.76	47.65	47.67	30.76	50.93
10	43.39	34.69	**41 627.01**	46.33	31.65	50.07
Average value (mm)	46.09	33.17	45.67	46.87	31.46	52.30
Standard deviation (mm)	3.43	0.95	1.87	0.69	0.38	1.12

The incorrect measurement results obtained using the SGBM algorithm are in bold and eliminated when calculating the average and standard deviation

4 Conclusions

In this paper, a simulated binocular dataset was generated by establishing a simulated gastroscope scenario through Blender to train a disparity estimation model for 3D measurement. The effectiveness of the disparity maps generated from simulated images was evaluated, and the results demonstrated that compared with the SGBM algorithm, the proposed method improved EPE by 48.9% and reduced the runtime by 90.5%. In addition, the proposed method did not create serious errors and had better 3D measurement stability. The method developed in this study further promotes the potential application of 3D measurement in the endoscopic environment and can provide accurate and reliable sizes to help endoscopists make accurate diagnoses.

Contributors

Hao YU and Bo YUAN designed the research. Hao YU, Changjiang ZHOU, and Wei ZHANG processed the data. Hao YU drafted the paper. Liqiang WANG and Qing YANG helped organize the paper. Hao YU revised and finalized the paper.

Compliance with ethics guidelines

Hao YU, Changjiang ZHOU, Wei ZHANG, Liqiang WANG, Qing YANG, and Bo YUAN declare that they have no conflict of interest.

References

Ahmad I, Levine JB, Anderson JC, 2016. Endoscopic measurement of colorectal polyps: how do we measure up? *Gastroenterology*, 150(3):769-771. https://doi.org/10.1053/j.gastro.2016.01.020

Anderson BW, Smyrk TC, Anderson KS, et al., 2016. Endoscopic overestimation of colorectal polyp size. *Gastrointest Endosc*, 83(1):201-208. https://doi.org/10.1016/j.gie.2015.06.058

Cai HF, Wang R, Li Y, et al., 2018. Role of 3D reconstruction in the evaluation of patients with lower segment oesophageal cancer. *J Thorac Dis*, 10(7):3940-3947. https://doi.org/10.21037/jtd.2018.06.119

Dimas G, Bianchi F, Iakovidis DK, et al., 2020. Endoscopic single-image size measurements. *Meas Sci Technol*, 31(7):074010. https://doi.org/10.1088/1361-6501/ab803c

Furukawa R, Aoyama M, Hiura S, et al., 2014. Calibration of a 3D endoscopic system based on active stereo method for shape measurement of biological tissues and specimen. Proc 36th Annual Int Conf of the IEEE Engineering in Medicine and Biology Society, p.4991-4994. https://doi.org/10.1109/EMBC.2014.6944745

Hirschmüller H, 2008. Stereo processing by semiglobal matching and mutual information. *IEEE Trans Patt Anal Mach Intell*, 30(2):328-341. https://doi.org/10.1109/TPAMI.2007.1166

Kendall A, Martirosyan H, Dasgupta S, et al., 2017. End-to-end learning of geometry and context for deep stereo regression. Proc IEEE Int Conf on Computer Vision, p.66-75. https://doi.org/10.1109/ICCV.2017.17

Khamis S, Fanello S, Rhemann C, et al., 2018. StereoNet: guided hierarchical refinement for real-time edge-aware depth prediction. https://arxiv.org/abs/1807.08865v1

Nomura K, Kikuchi D, Kaise M, et al., 2019. Comparison of 3D endoscopy and conventional 2D endoscopy in gastric endoscopic submucosal dissection: an ex vivo animal study. *Surg Endosc*, 33(12):4164-4170. https://doi.org/10.1007/s00464-019-06726-w

Ogino-Nishimura E, Nakagawa T, Sakamoto T, et al., 2015. Efficacy of three-dimensional endoscopy in endonasal

surgery. *Auris Nasus Larynx*, 42(3):203-207. https://doi.org/10.1016/j.anl.2014.10.004

Omori J, Goto O, Higuchi K, et al., 2020. Three-dimensional flexible endoscopy can facilitate efficient and reliable endoscopic hand suturing: an ex-vivo study. *Clin Endosc*, 53(3):334-338. https://doi.org/10.5946/ce.2019.207

Sakata S, Mcivor F, Klein K, et al., 2018. Measurement of polyp size at colonoscopy: a proof-of-concept simulation study to address technology bias. *Gut*, 67(2):206-208. https://doi.org/10.1136/gutjnl-2016-312915

Shaw MJ, Shaukat A, 2016. Does polyp size scatter matter? *Gastrointest Endosc*, 83(1):209-211. https://doi.org/10.1016/j.gie.2015.08.060

Wang D, Liu H, Cheng X, 2018. A miniature binocular endoscope with local feature matching and stereo matching for 3D measurement and 3D reconstruction. *Sensors*, 18(7):2243. https://doi.org/10.3390/s18072243

Wang XZ, Nie YF, Lu SP, et al., 2020. Deep convolutional network for stereo depth mapping in binocular endoscopy. *IEEE Access*, 8:73241-73249. https://doi.org/10.1109/ACCESS.2020.2987767

Žbontar J, LeCun Y, 2014. Computing the stereo matching cost with a convolutional neural network. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.1592-1599. https://doi.org/10.1109/CVPR.2015.7298767

Zhang FH, Prisacariu V, Yang RG, et al., 2019. GA-Net: guided aggregation net for end-to-end stereo matching. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.185-194. https://doi.org/10.1109/CVPR.2019.00027