Machine Learning-based Depth Prediction of End-Effector for 3D Robotic Micromanipulation

Jiaqi Wang, Jiaqi Chen, Chenjie Wang, Zhuoran Zhang

Abstract— **Visual servo control of end-effectors is a crucial step in robot micromanipulation. In the three-dimensional positioning problem of end-effector, while methods have been developed for visually detecting the x-axis and y-axis positions of the end-effector tip, it remains challenging to obtain visual feedback of the z-axis positions. In this paper, a new strategy is proposed to estimate the z-axis position of the end-effector. Instead of using depth-from-focus and depth-from-defocus methods, we transform z-axis positioning problem into a multiclass classification problem. Our strategy takes a monocular image of the end-effector as input, classifies it into different depth intervals, and outputs the focal plane z-axis position for that interval. A deep learning model is developed to solve the multiclass classification problem. Considering of the shallow depth of field of an optical microscope, a novel loss function is proposed to penalize misclassification. Using glass micropipettes as an example, the deep learning model achieves an accuracy of 96.1% for depth prediction/classification. The proposed strategy provides a new method for locating the out-of-focus depth of the end-effector and for providing 3D visual feedback for robotic micromanipulation.**

Keywords: End-effector manipulation, Automation at micro-scale, Robot Vision

I. Introduction

The past decades have witnessed significant development of robotic micromanipulation techniques. Under the visual guidance of an optical microscope, an end-effector is automatically controlled for the assembly of microparts, material characterization, and manipulation of biological cells [1]. In robotic micromanipulation, obtaining the visual feedback of end-effector position is an essential step in visual servo control. The positioning of the end-effector is performed in three-dimensional space, thus requiring the robot system to have three-dimensional (x-y-z) perception of the location of the end-effector.

Within the focal plane (x-y plane), the object being imaged (i.e., end-effector) is clearly visible, and the visual detection algorithms for its position have been relatively mature, such as the recognition algorithm based on template matching [2], active contour [3], and feature points [4]. Outside the focal plane, the end-effector is blurred and it is difficult to obtain the depth information in z-direction (Fig. 1). On the macroscopic domain, depth information in robotic manipulation can be obtained by different imaging modalities, such as binocular stereo vision depth estimation [5], structured light-based 3D depth information estimation [6][7], holographic imaging [8][9], etc. Among them, binocular vision has also been introduced to microscopic operations, such as adding a side-view microscope to the conventional top-view microscope to provide direct visual feedback in the z-direction [10][11]; however, the implementation of such methods all require the support of additional hardware. In applications of robotic micromanipulation such as in clinical assisted reproductive treatment, manipulation of cells is usually performed under a conventional monocular optical microscope, which does not support hardware expansion.

In monocular microscopic vision, z-direction position estimation algorithms are usually divided into two categories: depth from focus and depth from defocus. The technical nature of depth of focus is auto-focusing, i.e., first focusing on the object by trial-and-error and up-and-down adjustment of the focal plane, and then inferring the 3D position of the object based on the z position of the current focal plane [12][13]; however, this method requires repeated trial-and-error adjustment of the focal plane until it is focused on the end-effector, limiting the speed of such methods. Depth from defocus estimates the out-of-focus depth of an object by establishing an out-of-focus model of the object being imaged, such as a look-up table between the out-of-focus distance of the object and the out-of-focus image feature (such as object area) [14][15]. Depth of defocus does not require adjusting the focal plane to focus on the object, thus avoiding the problem of reduced efficiency due to repeated trial-and-error searching for the focus position. However, the accuracy of depth of defocus relies heavily on the accurate establishment and calibration of the off-focus model. End-effectors have various geometry of its projection on the focal plane, making it challenging to accurately establish and calibrate such off-focus models.

Different from existing depth-from-focus methods and depth-from-defocus methods, this paper utilizes the characteristics of microscope depth of field and transforms the continuous depth estimation problem into a depth classification problem. The multi-class depth classification problem is then solved by machine learning, which takes a single end-effector image as input and predicts/classifies its corresponding depth of the originating imaging plane. Using glass micropipettes as an example, the developed machine learning model achieved a prediction accuracy of 96.1%.

The authors are with School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, 517182, Guangdong Province, China (e-mail: 119010017@cuhk.edu.cn, wangjiaqi@cuhk.edu.cn). Corresponding author: Zhuoran Zhang (zhangzhuoran@cuhk.edu.cn).

C. Wang is with the Biomedical Engineering Department, The Chinese University of Hong Kong, Shenzhen, 517182, Guangdong Province, China (e-mail: 119020396@link.cuhk.edu.cn).

This work is supported by the University Development Fund of CUHKSZ (UDF01002141), National Natural Science Foundation of China under Grant (62203374), and Guangdong Basic and Applied Basic Research Foundation (2021A1515110023).

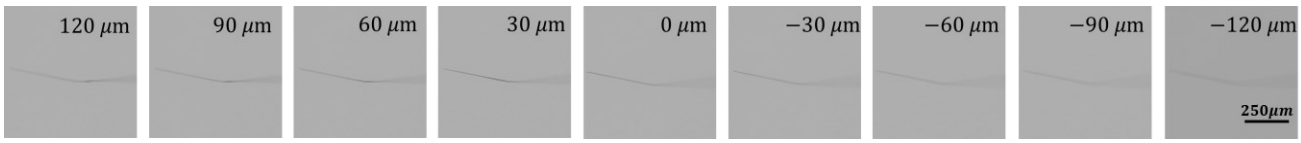

Figure 1. Images of the same pipette at different z-positions, marked on top of corresponding images.

II. METHOD

A. System overview

The robotic micromanipulation system consists of a standard inverted microscope (Eclipse Ti2, Nikon), a CMOS camera (Basler MED ace 23 MP 164 color, Balser) and a motorized micromanipulator (Sutter MP-285, Sutter) (Fig.2). The end-effector is a glass micropipette used for cell injection in clinical infertility treatment, which is controlled by the sutter micromanipulator to move the z-axis and adjust the out-of-focus distance after finding the in-focus plane. The micropipette image with different out-of-focus distances was taken with under a 4x objective.

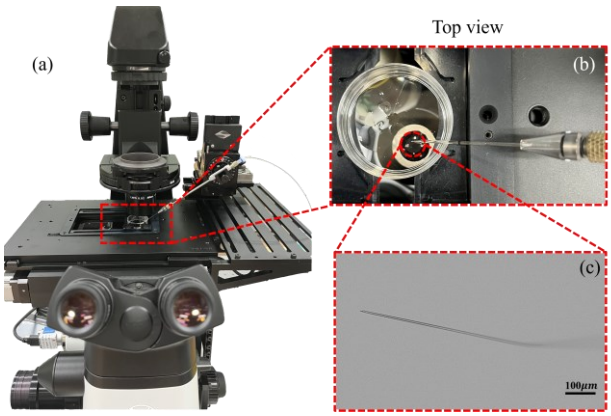

Figure 2. (a) Setup of the robotic micromanipulation system. (b) The top view of the set up. (c) Micropipette imaging at 4× magnification

B. The Depth Prediction of Micropipette under Microscope

The traditional methods cannot meet the demand of dynamic depth prediction of the micropipette under microscope in terms of both speed and accuracy, so a new method for estimating the out-of-focus depth of pipette is proposed in this paper.

The depth of field of an optical microscope is limited, and objects within the depth of field are simultaneously in focus. Hence, the continuous out-of-focus depth can be approximated by dividing it into multiple intervals, and the observed micropipette images in each interval show an approximate state of the focal plane with the same degree of out-of-focus as that interval. This could naturally transform the depth estimation problem into a classification problem: to which interval (originating imaging focal plane) does the micropipette image belong to. Therefore, the continuous depth prediction problem of the pipette under the microscope is transformed into a classification problem: classify the pipette image to different focal plane intervals, input an image, predict which focal plane interval it is in and output its Z-axis position (focal plane). Then, the multi-class classification problem can be solved by using machine learning methods. In this paper, a classification method based on ResNet-34 is proposed and compared with other machine learning algorithms and existing traditional methods based on focus measure.

To train the depth classification model, a dataset consists of 900 images (100 z-stacks and each z-stack containing 9 images) of micropipette was collected. The depth of field of the 4x objective (numerical aperture: 0.13) used in this paper is 55.5 μm, then the range of sharpness variation of the observed objects is small [Fig.3(a)] and the image information changes less in the range of 27.75 μm above and below the focus point. Therefore, we divided the continuous depth information and acquired images every 30 μm with the in-focus position of the pipette as the origin, and the image out-of-focus degree changed significantly. Pipette images acquired at -120 μm are classified as class 1, while images acquired at +120 μm are classified as class 9. The size of each image is 798 × 798 pixels for each image. To enable the algorithm to fully learn the depth information of the images, we rotate the pipette to acquire images from multiple angles (Fig. 3(b)).

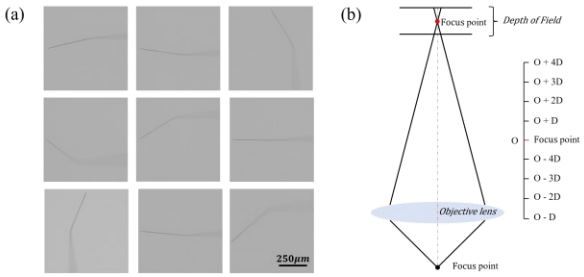

Figure 3. (a) Schematic diagram of the depth of field of the objective lens. (b) Pipette images of multiple angles in the dataset.

C. Loss function design with penalty coefficients

The cross-entropy loss function is typically applied to multiclassification problems, and its expression is

$$CELoss = -\frac{1}{N}\sum_{n=1}^{N} Log(P_n, i) \quad (1)$$

In order to improve the classification accuracy of the model, we add penalty coefficients to the cross-entropy loss function. The images in the pipette dataset collected based on depth gradient are strongly correlated, and there is a correspondence between the difference in depth gradient and the difference in category. The larger the category difference is, the larger the depth gradient difference is. In order to improve the classification accuracy of the model for similar

depth images, we increase the loss for the case of wrong prediction, which makes the algorithm pay more attention to it. The new loss function is designed as

$$WLoss = -\frac{1}{N}\sum_{n=1}^{N} W_n Log(P_n, i) \quad (2)$$

$$W_n = \left|L_i{}^2 - L_n{}^2\right| + 1 \quad (3)$$

where L_i is the ground truth, L_n is the predicted label value, and P_n is the predicted probability of SoftMax output.

III. RESULT AND DISCUSSION

A. *Performance of the depth prediction model*

The depth prediction model was evaluated by its prediction accuracy, model size, training time and inference time. As summarized in Table 1, the model performance was compared with that of conventional deep learning models, including VGG, GoogLeNet, ResNet-50, and ResNet-101. The highest accuracy is ResNet34 with WLoss with 96.1%, followed by ResNet34 with 80.2%. The fastest FPS is AlexNet with 42 images per second, and the smallest model size is 41.4 M which is GoogLeNet. Typically, a frame rate not lower than 30 FPS is regarded as real-time for potential robotic micromanipulation tasks [13]. Considering the trade-off between accuracy and inference time, ResNet-34 with WLoss was finally chosen as the depth prediction model.

TABLE 1. PERFORMANCE OF Z-POSITION PREDICTED BY DIFFERENT DNN-BASED MODELS.

Model	*Accuracy*	*Model Size*	*Training Time*	*FPS*
ResNet-34	80.20%	85.3 M	2999	31
ResNet-50	78.90%	94.4 M	3927	32
ResNet-101	76.60%	170.7 M	5962	27
GoogLeNet	75.30%	41.4M	3429	32
AlexNet	61.70%	58.4M	2241	42
ResNet-34+WLoss	**96.10%**	**85.3 M**	**2999**	**30**

By adding penalty coefficients to the cross-entropy loss function, the algorithm's loss in case of misclassification is increased to improve the model's focus on error cases, and the parameters are dynamically updated by backward propagation to adjust the model's weights and biases. When the gap between the predicted class and the ground truth is larger, the penalty coefficient is larger, making the misclassification interval of the model smaller. Compared with the model trained using cross entropy as the loss function, the accuracy is improved by 15.9%, the precision is improved by 16.08%, the recall is improved from 80.30% to 96.29%, and the F1-score is improved by 16.17% (Table.2). Overall, using WLoss (Eq. 2) improved the model performance.

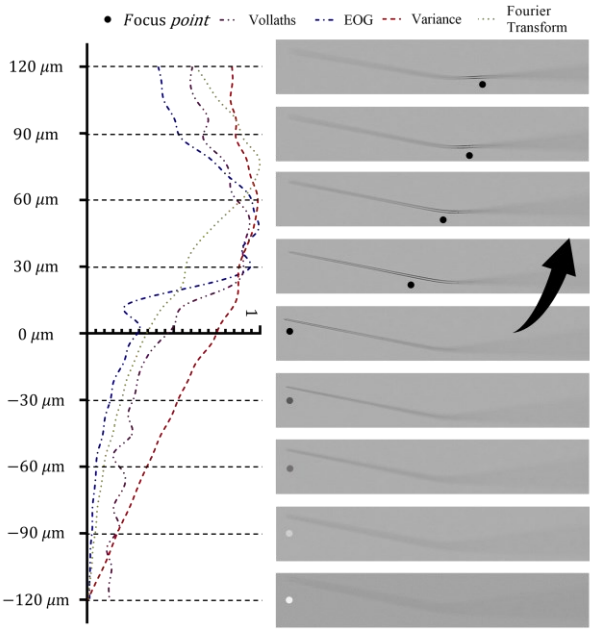

Figure 4 . Performance of in-focus image selection and the actual pipette image corresponding to the Z-axis. The black dot indicates the position of the focus, and the arrow indicates the trend of the movement of the focus.

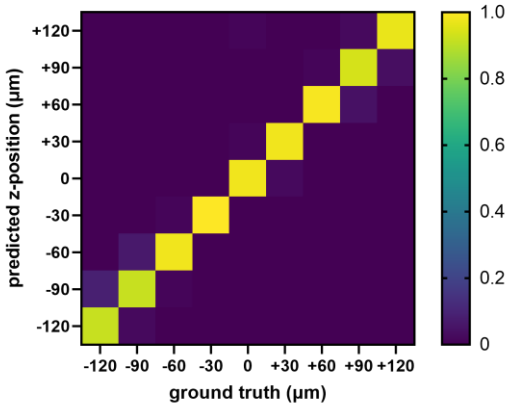

Figure 5 . The confusion-matrix, of Zpositions predicted by the classification model on the same z-stacks of pipette images. The color bar shows the color coding of classification probability.

B. *Comparison of depth estimation using the prediction model versus conventional focus measure methods*

Further we compared the improved ResNet-34 algorithm with the traditional methods. The traditional methods use different focus measure for in-focus image selection, we selected four more representative methods as baseline to compare with the machine learning based methods:

1. Variance-based method. The variance function is used to represent the dispersion degree of the image grayscale distribution. In-focus images have a large range of grayscale value changes, a high degree of dispersion, and a large variance.

TABLE 2. COMPARISON BETWEEN RESNET-34 AND RESNET-34+WLOSS PERFORMANCE INDICATORS.

Model		-120μm	-90μm	-60μm	-30μm	0μm	+30μm	+60μm	+90μm	+120μm
ResNet-34	Precision	0.977	0.807	0.852	0.841	0.818	0.795	0.693	0.568	0.864
	Recall	0.956	0.826	0.721	0.851	0.923	0.824	0.735	0.633	0.760
	F1-score	0.966	0.816	0.781	0.846	0.867	0.809	0.713	0.599	0.809
ResNet-34+WLoss	Precision	0.968	**0.937**	**0.989**	**0.979**	**0.979**	**1.000**	**0.979**	**0.916**	**0.916**
	Recall	**0.968**	**0.957**	**0.959**	**0.989**	**0.979**	**0.990**	**0.939**	**0.906**	**0.978**
	F1-score	**0.968**	**0.947**	**0.974**	**0.984**	**0.979**	**0.995**	**0.959**	**0.911**	**0.946**

2. Method based on two-dimensional discrete Fourier transform. the distance from the pixel in the Image to the center pixel is used as the high-frequency component in the emphasis spectrum, and the more high-frequency components, the clearer the focus of the image.
3. Method based on autocorrelation function. The lower the correlation between pixels, the clearer the in-focus image edge.
4. Gradient-based method. The sharper the image edge, the larger the gradient.

Results of the focus-measure methods are shown in Fig.4, where z = 0 μm is the ground truth value of the in-focus micropipette image. However, the four focus-measure methods all reached their peak focus-measure scores at -60 μm and -90 μm, resulting in an average absolute error of 75 μm. Obviously, the position of the corresponding micropipette tip is not in the in-focus position (see black dots in Fig. 4). The recognition error of the traditional algorithm is between 50% and 75%.

The large error of focus-measure methods is mainly because adjusting the depth of the micropipette upward does not lead to the defocusing of the whole image, but to the movement of the focus point. These focus-measure methods can only determine whether the image is in-focus by the change of the pixel gradient. Since the in-focus interval of the end-effector will move from its tip to the body part when moving upward and does not disappear, the focus area becomes larger instead. Therefore, the pixel gradient still exists, so the focus-measure methods cannot accurately judge the focus situation of the tip, and will produce a large misidentification due to the shift of the in-focus area. In contrast, using ResNet-34 can fully learn the features of each out-of-focus depth interval, and it can be seen from Fig.5 that the algorithm's classification accuracy for each class is above 90%. This can be attributed to the fact that the model learns not only the gradient change, but also the geometry of the end-effector, thus increasing its ability for distinguishing different in-focus parts within the end-effector.

REFERENCES

[1] Z. Zhang, X. Wang, J. Liu, C. Dai, and Y. Sun, "Robotic micromanipulation: Fundamentals and applications," Annu. Rev. Control Robot. Auton. Syst., vol. 2, no. 1, pp. 181–203, 2019.

[2] E. Shojaei-Baghini, Y. Zheng, and Y. Sun, "Automated micropipette aspiration of single cells," Ann. Biomed. Eng., vol. 41, no. 6, pp. 1208–1216, 2013.

[3] J. Liu et al., "Locating end-effector tips in robotic micromanipulation," IEEE Trans. Robot., vol. 30, no. 1, pp. 125–130, 2014.

[4] Y. Ma, K. Du, D. Zhou, J. Zhang, X. Liu, and D. Xu, "Automatic precision robot assembly system with microscopic vision and force sensor," Int. J. Adv. Robot. Syst., vol. 16, no. 3, p. 172988141985161, 2019.

[5] Y. Zhang, L. Zhu, R. Hamzaoui, S. Kwong, and Y.-S. Ho, "Highly efficient multiview depth coding based on histogram projection and Allowable Depth Distortion," IEEE Trans. Image Process., vol. 30, pp. 402–417, 2021.

[6] K. Wu, J. Tan, H. L. Xia, and C. B. Liu, "An exposure fusion-based structured light approach for the 3D measurement of a specular surface," IEEE Sens. J., vol. 21, no. 5, pp. 6314–6324, 2021.

[7] K. Zhang, M. Yan, T. Huang, J. Zheng, and Z. Li, "3D reconstruction of complex spatial weld seam for autonomous welding by laser structured light scanning," J. Manuf. Process., vol. 39, pp. 200–207, 2019.

[8] M. U. Daloglu et al., "Label-free 3D computational imaging of spermatozoon locomotion, head spin and flagellum beating over a large volume," Light Sci. Appl., vol. 7, no. 1, p. 17121, 2018.

[9] T.-W. Su, L. Xue, and A. Ozcan, "High-throughput lensfree 3D tracking of human sperms reveals rare statistics of helical trajectories," Proc. Natl. Acad. Sci. U. S. A., vol. 109, no. 40, pp. 16018–16022, 2012.

[10] P. Ryan and E. Diller, "Magnetic actuation for full dexterity microrobotic control using rotating permanent magnets," IEEE Trans. Robot., vol. 33, no. 6, pp. 1398–1409, 2017.

[11] S. Liu and Y.-F. Li, "Precision 3-D motion tracking for binocular microscopic vision system," IEEE Trans. Ind. Electron., vol. 66, no. 12, pp. 9339–9349, 2019.

[12] Z. Wang, C. Feng, W. T. Ang, S. Y. M. Tan, and W. T. Latt, "Autofocusing and polar body detection in automated cell manipulation," IEEE Trans. Biomed. Eng., vol. 64, no. 5, pp. 1099–1105, 2017.

[13] J. Liu et al., "Automated vitrification of embryos: A robotics approach," IEEE Robot. Autom. Mag., vol. 22, no. 2, pp. 33–40, 2015.

[14] K. M. Taute, S. Gude, S. J. Tans, and T. S. Shimizu, "High-throughput 3D tracking of bacteria on a standard phase contrast microscope," Nat. Commun., vol. 6, no. 1, p. 8776, 2015.

[15] A. Zhang and J. Sun, "Joint depth and defocus estimation from a single image using physical consistency," IEEE Trans. Image Process., vol. 30, pp. 3419–3433, 2021.