Structured-Light-Based 3D Scanning System for Industrial Manipulator in Bin Picking Application

Phuc Thanh Ly, Quoc Chi Nguyen, *Member, IEEE*, Ngoc Duy Hung Nguyen,
Phuong-Tung Pham, *Member, IEEE*, and Keum-Shik Hong, *Fellow, IEEE*

Abstract— **This paper developed an industrial manipulator integrated with a structured-light-based 3D scanning system using Gray code and phase-shifting patterns, implementing the bin-picking problem. The procedure of the vision-based robotic system developed in this paper is described as follows: First, the object is reconstructed in point cloud format by projecting a series of Gray code patterns followed by four-phase shift patterns. Point clouds of the object from different views are filtered and concatenated to obtain the full cloud. The full clouds, altogether with partial clouds, then are fed to PointNet neural network as a training dataset, whose estimated grasping pose can be used to pick up the objects from a bin. By using the estimated grasping pose, the robotic arm can reach the desired position and pick the object. Experimental results indicate that the developed vision-based system can reconstruct objects and grasp different-sized objects in the bin.**

I. Introduction

Industrial robotic arms often are utilized to automate the process of picking and placing objects. In the picking tasks, the conventional robotic arm needs to pre-determine the position of the object to be picked up. It results that they cannot handle problems wherein the positions and orientations of the object are unknown, e.g., objects in a bin. In these situations, the computer vision system can be integrated into the robotic system to determine the pose of objects. Industrial applications of 3D computer vision demand robustness and high accuracy in data acquisition for many purposes. Many approaches have been considered where either laser or structured light scanner is employed. While both have their own advantages, structured light scanners have proven more effective in close-range applications where objects are small to medium [1], and obtaining sub-millimeter accuracy is a priority [2]. Therefore, implementations of stereo vision in bin picking, positioning, and extracting individual elements in a disorganized environment have gathered increasing interest due to ongoing challenges posed by manufacturing and market demands [3-4].

In the spectrum of stereo vision, there are two primary approaches, i.e., passive and active. For the former approach, ordinary images are obtained, and input images are not encoded. As a result, output data are prone to errors and heavily dependent on the prominent distinction between pairs of input images [5]. Consequently, passive stereo systems are obsolete, making way for active systems. For active approaches, structured-light-based stereo system concerns coded light patterns, which are projected onto objects, so input images for correspondence matching are substantially enhanced and optimized. This raises the opportunities for robust integration into robot manipulators where 3D reconstruction is required. This implementation means significant improvements in automation, cutting down on operation costs and human interference for tasks that demand a combination of data acquisition and process actuation from the robots.

Structured light 3D reconstruction has been investigated in the literature. The ideal of vision systems based on a combination of Gray code and phase shifting profilometry (PSP) was proposed by Gühring et al. [6] and Sansoni et al. [7]. More recently, combining RGB-D depth camera, using infrared, with laser 3D scanner was applied for the maintenance of aircraft fuselage in which the system picked up any surface deformations [2]. Structured light for forensic medicine was also studied in [8], where blue fringes were projected for the registration of body parts that could be used for autopsy or criminal investigation. Phase shifting profilometry was also implemented individually for the sake of fast, real-time performance in 3D reconstruction [9], using 4-step PSP patterns. Though robust and easy to deploy, Gray code and PSP suffer when specular surfaces are present because projected patterns are not distinguishable, this problem was addressed and studied in [10], where maximum min-SW Gray code was implemented to overcome reflective light. Cuc et al. [11] also proposed another approach to overcome reflective surfaces using histogram thresholding for optimal exposure time. Furthermore, novel structured light patterns have also been studied. In particular, Zhang [12] proposed a circular PSP pattern to increase the multiplicity and decoding accuracy, especially for specular surfaces, and maintain sub-pixel accuracy with the limited number of images.

For bin picking purposes, objects are complex and require a combination of many segmentation methods, including instance, semantic and part segmentations for successful localization and positioning [13]. Gou et al. [14] proposed an approach using improved density-based spatial clustering of

*This work was supported by the National Research Foundation (NRF) of Korea under the auspices of the Ministry of Science and ICT, Korea (grant no. NRF-2020R1A2B5B03096000)

Q. C. Nguyen is with the Department of Mechatronics, Faculty of Mechanical Engineering, Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, Ho Chi Minh City, Vietnam (corresponding author) (e-mail: nqchi@hcmut.edu.vn).

P. T. Ly, N. D. H. Nguyen, and P.-T. Pham are with the Department of Mechatronics, Faculty of Mechanical Engineering, Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, Ho Chi Minh City, Vietnam (e-mails: thanh.ly0610@hcmut.edu.vn; pptung@hcmut.edu.vn hung.nguyen3112@hcmut.edu.vn).

K.-S. Hong is with the School of Mechanical Engineering, Pusan National University, Busan 46241, South Korea (e-mail: kshong@pusan.ac.kr).

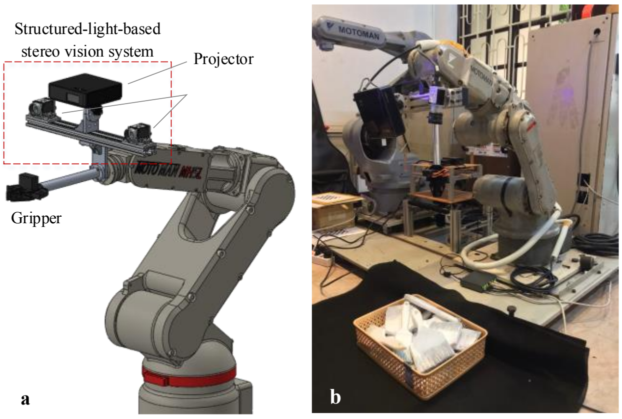

Figure 1. **a** Robot arm integrated with a structured-light-based 3D scanning system; **b** Bin picking experiment setup

applications with noise (DBSCAN) algorithm together with random sample consensus (RANSAC) for removing bin outliers and grouping closely related points to form a singular grasping object. Moreover, machine learning algorithms are also employed for these tasks. Qi et al. [15] proposed a novel neural network named PointNet applying a Multi-Layer Perceptron directly on the points to learn features for object classification, part segmentation, etc., which are all subsidiaries of bin picking task fulfillment. In addition, Xu et al. [16] proposed a high-speed instance segmentation scheme for 3D point clouds and a novel neural network named FPCC-Net with new clustering algorithms.

Thus far, to the best of our knowledge, PointNet segmentation has not been widely applied bin-picking problems. Ni et al. [17] created a customized dataset and used PointNet++ for pose estimation; however, they must rely on a predetermined grasping dataset, i.e., marking possible grasping positions before training. As a result, the outputs are not robust enough and heavily reliant on predetermined markings, rendering it unsuitable for bin picking where object orientation is not known beforehand. To overcome these issues, in this paper, we proposed a novel, end-to-end solution for a vision-based automatic bin-picking robot based on PointNet segmentation. First, the vision system is calibrated individually for each camera; then, stereo calibration is performed using the data acquired in the previous stage. While capturing images for camera calibration, the system is moved accordingly to waypoints in a predetermined trajectory optimized for accuracy. The pose at each waypoint is recorded and used in the Daniilidis hand-eye calibration stage. After the calibration steps are finished, the targeted objects are reconstructed in point cloud format by projecting a series of six Gray code patterns followed by four-phase shift patterns at 0, π/2, π, and 3π/2. Then, point clouds from different views, hereafter partial/view clouds, are subject to statistical and radius outlier removal filters. Filtered view clouds are then concatenated using transformation matrices of hand-eye calibration (camera to end effector) and model poses (end effector to base). The finished concatenated clouds, hereafter full clouds, altogether with partial clouds, are fed to PointNet [7] neural network as a training dataset, whose estimated grasping pose can be used to pick up the objects from a bin. The experiment is conducted to verify the proposed solution.

This paper is organized as follows. Section 2 presents the vision-based automated bin picking system. Section 3 introduces the process of 3D construction and 3D segmentation. Section 4 shows several experimental results and discussions. Finally, conclusions are given in Section 5.

II. System description

In this research, we integrate a structured-light-based stereo vision system into a 6-DOF Motoman HP3 manipulator (Fig. 1a). The robotic system aims to grasp different-sized paintbrushes randomly placed in the bin.

The structured light stereo vision system consists of two Basler cameras and one ASUS ZenBeam projector. Cameras are used to capture images, whereas the projector is utilized to project a series of six Gray code patterns. Basler cameras are connected to the central processing unit through a Power over Ethernet (PoE) switch that can transmit image information and serve as a source for camera operations. Additionally, this transmission protocol benefits high-speed data acquisition and robot communication. Waypoint data are sent to the robot through the PoE switch as a predetermined trajectory optimized for high calibration accuracy and the most well-rounded surface scan. The robot's end effector is a gripper mounted on a cylindrical aluminum rod. The experiment is set up as shown in Fig. 1b.

III. 3D Reconstruction and 3D Segmentation

A. 3D Reconstruction

Calibration: To perform 3D reconstruction, the calibration processes, including camera calibration, stereo calibration, and hand-eye calibration, are required.

The camera calibration is the process of calibrating each camera individually. Based on this process, the extrinsic matrix, intrinsic matrix, and distortion coefficients of a camera are obtained. For industrial purposes, calibration is performed offline, using parameters acquired under specific conditions. This calibration method is preferred because it offers high accuracy and noise resistance. This thesis implemented the calibration method, which makes use of a planar pattern or checkerboard. This method only requires at least two different views where both camera and the pattern can be moved arbitrarily. The camera's intrinsic and extrinsic parameters (position and orientation relative to the calibration board coordinate system) can be computed using the maximum likelihood inference. Once each set of parameters is obtained for each camera, the stereo calibration can be performed. The stereo calibration determines the geometric relationship between the right and left cameras, i.e., a transformation matrix indicates the rotation and translation that transforms the left camera coordinate to the right camera coordinate and vice versa.

Hand-eye calibration is computing the relationship between the image sensor (vision system) and the end effector (robot gripper). Consider the task of grasping an object at an unknown position relative to the robot tool end. This calibration step maps the detected object's grasping point to the tool endpoint. Firstly, a separate vision algorithm determines the position and orientation of the object with

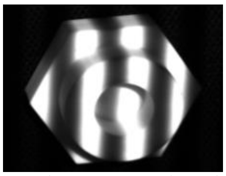

Figure 2. Gray code patterns projected to the object.

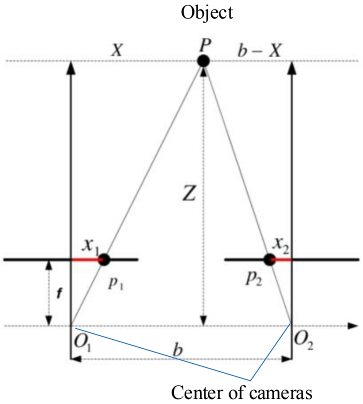

Figure 3. Triangulation scheme of stereo vision [20]

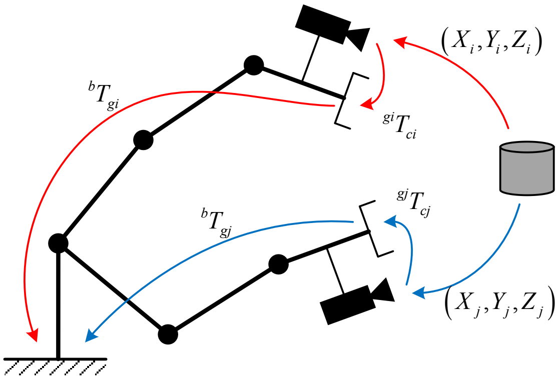

Figure 4. Eye-in-hand configuration

respect to the sensor or camera image plane. Secondly, the object grasping parameters (positions and orientations), which have been calculated in the previous stage, is mapped from the image plane to the end effector (gripper) frame. Consequently, a robot can understand an object's position through a hand-eye transformation matrix, move its gripper towards the object, and grasp it. This paper uses the Daniilidis hand-eye calibration method [18].

Pattern Generation: Finding correspondences from two images is the primary problem of 3D reconstruction using the active stereo vision system. In such systems, the correspondences are controlled by projecting specific patterns (structured light) and affixing the information of each pattern onto the object, see Fig. 2. The information of each pixel in the first image plane is then used to derive its correspondence on the other image plane. This process is referred to as pattern encoding and decoding. Structured light serves to encode input images whose correspondences can help obtain better stereo matching in the decoding process and triangulation for 3D depth.

Gray code and phase shifting patterns are selected for 3D reconstruction. Theoretically, Gray code can encode images at the pixel level (210 for 1024x768). In practice, it is more robust than binary encoding, with wrong decoding resulting in misplacement of at most one resolution unit. Gray code bright and dark lines width with the finest resolution is twice as wide as compared to that of binary code. This facilitates analysis, especially at steep object surfaces where the code appears to be compressed. This paper employs a 5-bit Gray code pattern instead of creating the pixel-level pattern. Despite being theoretically feasible, Gray code in practice suffers from light interference as reflection and camera sensors' incapability to capture the fine, minute distinction between large bit pattern fringes.

The phase shifting profilometry is used to overcome this issue by wrapping the phase into the encoded pictures. To generate phase-shift patterns, sinusoidal values are firstly generated. After which, the values are normalized such that they are bound by [0, 1]. Consequently, the values in each phase multiply with the array [0, 255] that represents black and white pixel values. Accordingly, they become discrete columns of black and white pixels, representing sinusoidal fringe patterns. The number of fringe patterns depends on system requirement, where more patterns mean higher accuracy but at the expense of processing time. In this paper, a four-step pattern (four fringe images with phase shift 0, $\pi/2$, π, and $3\pi/2$) is chosen due to higher accuracy, error tolerance, and low pattern count than the three-step pattern [19].

Point Cloud Registration, Filter: Applying stereo calibration parameters obtained in the previous steps, encoded input images are rectified. Encoded rectified images then go through a triangulation process where 3D data is derived, see Fig. 3. Assuming camera intrinsic and baseline distance b are known, the depth of image point can be computed.

$$\frac{b}{Z} = \frac{b + x_1 - x_2}{Z - f} \Rightarrow Z = \frac{bf}{x_1 - x_2}, \tag{1}$$

where f is the camera focal length, x_1 and x_2 are the distances demonstrated in Fig. 3. Once Z has been determined, X and Y are easily computed using a similar triangle theorem. Accordingly, a raw point cloud is obtained. However, noise is still present in clouds. Two cloud filters are employed in this research based on their merits and performance. Radius outlier removal filter is powerful and intuitive. However, it leaves large residues when applied to dense clusters, plus it is very difficult to tune and only ideal for post-processing where dense clusters have been removed. The other approach, statical outlier removal, uses the statistical method of distance mean and standard deviation to identify outliers: those outside one standard deviation will be removed. This is an aggressive filtering method. Still, it leaves small residues with dense clusters and usually mistakes inliers for outliers because sparse clouds are difficult to tune, and dense clouds reserve the most clustered points. In this paper, a combination of statical and radius outlier removal filters is used, which offers great compensation for each's drawbacks.

Cloud concatenation is the process of concatenating two different view clouds. As the robot moves the stereo system for scanning, the pose at each view cloud is recorded, and the transformation matrix from the end effector to the base is derived from joint angles. Figure 4 illustrates the vision system that captures the object in the two different views, i.e., i- and j-view. Initial matrix ${}^{cj}\mathbf{T}_{ci}$, representing the coordinate transformation from camera position i to j, is derived with the following formula.

$$ {}^{cj}\mathbf{T}_{ci} = {}^{g_j}\mathbf{T}_{cj}^{-1}\,{}^{\mathrm{b}}\mathbf{T}_{gj}^{-1}\,{}^{\mathrm{b}}\mathbf{T}_{gi}\,{}^{gi}\mathbf{T}_{ci} \tag{2} $$

where ${}^{\mathrm{b}}\mathbf{T}_{gi}$ and ${}^{\mathrm{b}}\mathbf{T}_{gj}$ are the transformation matrices that transform the gripper coordinates corresponding to the i and j point clouds, respectively, to the base coordinate; ${}^{gi}\mathbf{T}_{ci}$ and ${}^{gj}\mathbf{T}_{cj}$ denote the matrices that transform the camera coordinates to the gripper coordinates corresponding to the i and j point clouds, respectively. Matrices ${}^{gi}\mathbf{T}_{ci}$ and ${}^{gj}\mathbf{T}_{cj}$ can be obtained from hand-eye calibration, the same for all positions because the camera system is fixed on the end effector.

Since point clouds are relative to the camera image plane, multiplying them with the obtained initial matrix allows all view clouds to be concatenated approximately close to each other. For refined cloud alignment, the iterative closest point (ICP) algorithm is employed with point-to-plane metrics.

B. 3D Segmentation

In the paintbrush picking task, it is necessary to prevent the robot from grasping along the curvature of the paintbrush or along its fragile hairs, either of which leads to highly unstable grasping. Therefore, 3D segmentation is performed. Image segmentation means partitioning an image or video frame into multiple image regions, with each region carrying a certain meaning. In this paper, a deep learning-based part segmentation model, called PointNet, is implemented, which instructs the robot to grasp along a pre-defined part of a paintbrush point cloud.

PointNet segmentation receives 3D data points as inputs. Those inputs go through a joint alignment network to normalize features and positions to local coordinates, as mentioned above. As dimensions expand, max-pooling using symmetric function on unordered point sets extracts global features of the clouds. The results of which are aggregated and concatenated with local features to create a multi-lay vector. This vector is then subject to Multilayer Perceptron (MLP) to derive cloud index scores into two classes: Graspable and ungraspable. These scores then go through a thresholding process to determine true graspability. The loss function used in this segmentation network is categorical cross-entropy.

CloudCompare's Interactive Segmentation Tool is utilized to define ground-truth labels for graspable and ungraspable parts, which are used for the training data. In CloudCompare, filtered view clouds and full clouds are annotated for grasping area, the thinnest part of the paintbrush handle, by segmenting the proposed cloud clusters. PointNet requires point clouds to be of certain symmetrization, so view clouds with different perspective coordinates are aligned to full clouds "coordinate perspective."

The dataset comprises 20 different paintbrushes of five dimensions. To reduce overfitting when training the model, data augmentation is performed to increase the amount of data. The training point clouds are augmented by applying random jitter and noise. Additionally, the dataset is enriched by cutting out certain parts of the clouds. This exposes the model to cases where the objects are not scanned from end to end.

The dataset is divided into a train-validation-test ratio of 208-52-40. This is to ensure that every brush's full cloud and view cloud are included in the test dataset while the train and validation test follows an 80-20 ratio of the remaining clouds. Then, we train for 50 epochs using stochastic gradient descent and Adam optimizer with 0.001 initial learning rate and 0.9 momentum value. The learning rate is halved every 15 epochs.

IV. Results and Discussion

A. 3D Reconstruction

The 3D reconstructing task of the robotic system using the proposed method is verified via reconstructing two different objects. Fig. 5 shows the point cloud of a plaster statue. By subjectively visual evaluation, fine surface details are all captured and registered in the frontal view. Superficial details such as the nose bridge, eye sockets, lips, and cheekbones are all prominent in the view cloud. Furthermore, the 3D reconstructed model shows certain creases on the face and neck of the statue clearly, see Fig. 5.b. In Fig. 6, the point cloud of a 3D printed block is demonstrated. The precision of the 3D reconstruction is evaluated based on actual and registered distance discrepancy. This quantitative evaluation experiment delivers satisfactory results, i.e., the error rate is less than 0.6%.

Figure 7 illustrates the process of full cloud concatenation. Full clouds are obtained by aligning view clouds, using the initial matrix in (2) and the ICP algorithm. They are subject to subsampling and normalization before being fed to PointNet. When initial matrices do not align view clouds to approximate positions, manual transformation using CloudCompare is employed. This means using CloudCompare to find the initial matrix by rough aligning them in the application, then using that transformation matrix for the ICP algorithm. The fully reconstructed point cloud of the object is nearly identical with minor errors. The full cloud concatenation experiment also produces expected results. In-depth analysis shows that the error rate maintains less than 0.3% for all dimensions. This plays an important role in preparing the dataset in a timely manner where the whole system takes less than 5 minutes from start to end, including camera calibration (mono, stereo, hand-eye), view cloud registration, cloud filtering, and full cloud concatenation. This aspect can be scaled to a great extent, reducing manual interference substantially, automating all steps in the workflow, and deploying the system for bin picking tasks. It is noted that the full clouds of the brushes are subject to subsampling and normalization before being fed to PointNet.

B. Bin picking task

In a bin-picking environment, the objects are disorganized. Therefore, the first scan of the bin (Fig. 8a) returns a noisy cloud with objects intertwined with each other (Fig. 8b). This cloud cannot run inference on PointNet because the model only understands singular brushes. Therefore, DBSCAN and RANSAC are used for removing

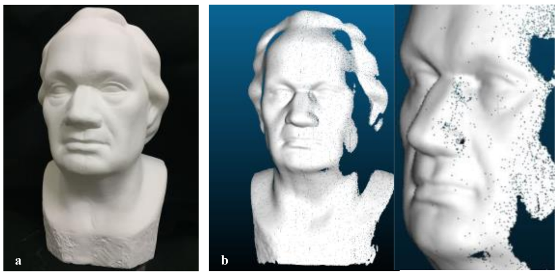

Figure 5. **a** Plaster statue; **b** Reconstructed view cloud

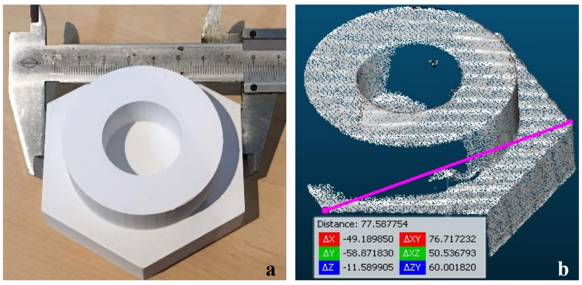

Figure 6. 3D printed block: **a** Actual block; **b** Reconstructed view cloud.

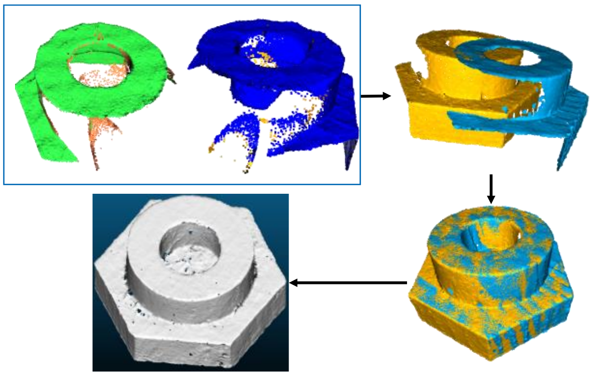

Figure 7. Full cloud concatenation.

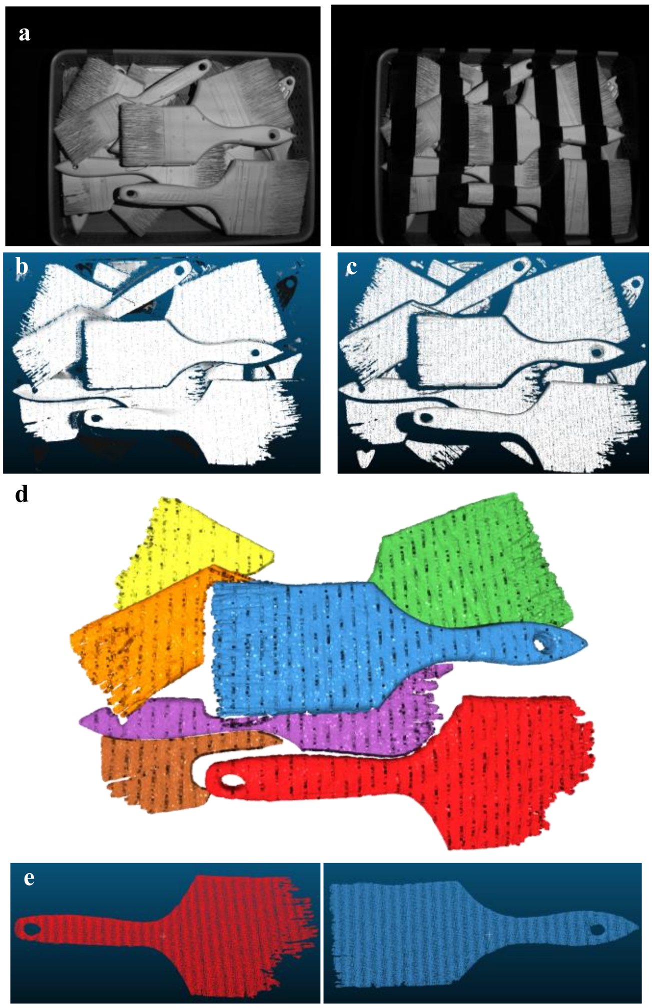

Figure 8. **a** Scanning the bin; **b** Noisy bin cloud; **c** Bin cloud after statistical and radius filters; **d** Segmented individual brushes using DBSCAN and RANSAC; **e** Segmented clouds ready for PointNet inference.

bin outliers and grouping closely related points to form a singular grasping object (Fig. 8d). As a result, these segmented clouds can run inference on PointNet (Fig. 8e).

After inference, the segmented part returns a cluster of graspable points in the cloud. The grasping area is determined to be the thinnest part of the handle (Fig. 9). The centroid of the cluster of graspable points in the cloud can be determined as follows:

$$\text{Centroid} = \left[\text{average}(x), \text{average}(y), \text{average}(z)\right] \quad (3)$$

Finally, this estimated centroid multiplied with the hand-eye matrix will return the grasping point in the robot coordinate. This allows the robot to move to the position and grasp the paintbrush (Fig. 10). The experiment shows the success rate of grasping the paintbrush is 64%. This success rate is not high.

C. Discussions

The success rate of grasping the paintbrush is limited due to the calibration issue. The hand-eye calibration algorithm used in this paper has low accuracy. Since the accuracy of hand-eye calibration heavily affects cloud concatenation and bin picking tasks, a better calibration algorithm should be considered. Our current scheme still lacks a cloud concatenation process, which stands as a bottleneck. This leads to poor, unsatisfactory initial view clouds alignment, resulting in erroneous cloud registration. Therefore, we suggest a more robust hand-eye calibration method to achieve higher accuracy for both grasping and point cloud aligning problems.

Furthermore, though efficient and adequately fast, the point cloud registration process still consumes more than 60 seconds. Industrial applications may require less than 5-10 seconds for every generation, including robot actuation time. This can be overcome by using fewer fringe patterns, triggering capturing process with a change in pattern projection instead of delaying for a fixed duration.

V. Conclusion

This paper develops an industrial robotic arm integrated with a structured-light-based stereo vision system for the bin-picking task. In this system, the solution for 3D reconstruction is developed based on the combination of the Gray code and phase shifting profilometry. Point clouds of

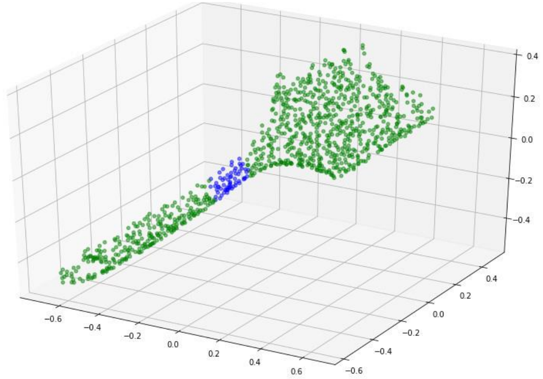

Figure 9. Cluster of graspable points in the cloud (blue dots).

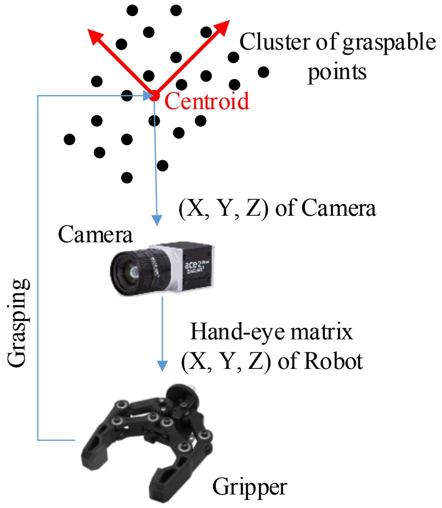

Figure 10. Grasping calculation sequence

the object are determined and then concatenated to obtain the full clouds. The full clouds altogether with partial clouds are fed to PointNet neural network as a training dataset, whose estimated grasping pose can be used to pick up the objects from a bin. The experimental results show that the developed system can reconstruct objects with high precision. Additionally, the system can extract each individual paintbrush cloud from the bin environment cloud, and the robotic arm can grasp along a pre-defined part of a paintbrush point cloud. However, the success rate of grasping the paintbrush is still limited. Several probable solutions proposed to improve the system include calibrating the system with a more robust method, using multiple angles to get redundant calibration parameters, equipping the system with a better gripper, and developing collision avoidance algorithm before grasping.

REFERENCES

[1] R. Singh, B. Baby, A. Suri and S. Anand, “Comparison of laser and structured light scanning techniques for neurosurgery applications,” in *International Conference on Signal Processing and Integrated Networks (SPIN)*, Noida, India, 2016.

[2] Y. Sun, L. Zhang and O. Ma, “Robotics-assisted 3D scanning of aircraft,” in *AIAA AVIATION 2020 FORUM*, Cincinnati, 2020.

[3] J.-K. Oh, S. Lee and C.-H. Lee, “Stereo vision based automation for a bin-picking solution,” *International Journal of Control, Automation, and Systems*, vol. 10, no. 2, 2012, pp. 362-373.

[4] D. Buchholz, “*Bin-Picking – New Approaches for a Classical Problem*,” Braunschweig, 2015.

[5] H. Kawasaki, R. Furukawa, R. Sagawa and Y. Yagi, “Dynamic scene shape reconstruction using a single structured light pattern,” in *2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage*, 2008.

[6] J. Gühring, C. Brenner, J. Böhm and D. Fritsch, “*Data Processing And Calibration Of A Cross-Pattern Stripe Projector*,” International Archives of Photogrammetry and Remote Sensing, Vols. XXXIII, Part B5, 2000, pp. 327-328.

[7] G. Sansoni, M. Carocci and R. Rodella, “Three-dimensional vision based on a combination of gray-code and phase-shift light projection: analysis and compensation of the systematic errors,” *Applied Optics*, vol. 38, no. 31, 1999, pp. 6565-6573.

[8] R. Breitbeck, W. Ptacek, L. Ebert, M. Fürst, G. Kronreif and M. Thali, “Virtobot – A Robot System for Optical 3D Scanning in Forensic Medicine,” in *4th International Conference on 3D Body Scanning Technologies*, California, 2013.

[9] X. Liu, H. Sheng, Y. Zhang and Z. Xiong, “A structured light 3d measurement system based on heterogeneous parallel computation model,” in *2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, Shenzhen, China, 2015.

[10] Y. Zhang and A. Yilmaza, “Structured light based 3d scanning for specular surface by the combination of gray code and phase shifting,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vols. XLI-B3, 2016, pp. 12-19.

[11] N. T. K. Cuc, N. V. Vinh and N. T. Hung, “Solution for shiny specular 3d mechanical surface measurement using combined phase shift and gray code light projection,” *Journal of Science & Technology*, vol. 135, 2019, pp. 1-6.

[12] Y. Zhang, “A structured light based 3d reconstruction using combined circular phase shifting patterns,” *The Ohio State University*, Ohio, 2019.

[13] M. Fujitaa, Y. Domaeb, A. Nodac, G. A. G. Ricardez, T. Nagatania, A. Zeng, S. Song, A. Rodriguezf, A. Causog, I. M. Cheng and T. Ogasawara, “What are the important technologies for bin picking? Technology analysis of robots in competitions based on a set of performance metrics,” *Advanced Robotics*, vol. 34, 2020, pp. 560–574.

[14] J. Guo, L. Fu, M. Jia, K. Wang and S. Liu, “Fast and robust bin-picking system for densely piled industrial objects,” in *2020 Chinese Automation Congress* (CAC), Shanghai, 2020.

[15] C. R. Qi, H. Su, K. Mo and L. J. Guibas, “PointNet: Deep learning on point sets for 3D classification and segmentation,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Honolulu, 2017.

[16] Y. Xu, S. Arai, D. Liu, F. Lin and K. Kosuge, “FPCC: Fast point cloud clustering-based instance segmentation for industrial bin-picking,” *Neurocomputing*, 2022.

[17] P. Ni, W. Zhang, X. Zhu and Q. Cao, “PointNet++ grasping: Learning an end-to-end spatial grasp generation algorithm from sparse point clouds,” in *IEEE International Conference on Robotics and Automation* (ICRA 2020), Paris, France, 2020.

[18] K. Daniilidis, “Hand-eye calibration using dual quaternions,” The *International Journal of Robotics Research*, vol. 18, no. 3, 1999, pp. 286-298.

[19] C. Zuo and L. Huang, “Phase shifting algorithms for fringe projection profilometry: A review,” *Optics and Lasers in Engineering*, vol. 109, 2018, pp. 23-59.

[20] Y.-C. Du, M. Muslikhin, T.-H. Hsieh and M.-S. Wang, “Stereo vision-based object recognition and manipulation by regions with convolutional neural network,” *Electronics*, vol. 9, no. 2, 2020, pp. 210