Learning-Based 3D Vision Sensor for Robotic Peg-In-Hole Insertion of Deformable Objects

Sher Hazan, Eylon Cohen, Ronit Schneor, Anath Fischer, and Miriam Zacksenhouse *Member, IEEE,*

***Abstract*—The integration of precise object positioning systems in robotic assembly tasks, particularly Peg-in-Hole (PIH) operations, is crucial for enhancing automation efficiency in industrial environments. The challenge intensifies when dealing with deformable objects. Advanced sensors and learning methods are recruited to face such challenges. This research presents a robust learning-based 3D vision sensor which is designed for estimating the relative position between both rigid and deformable objects in real-time. It is planned to be embedded into a robotic PIH insertion pipeline in the industrial environment. Our 3D vision sensor incorporates capturing the 3D scene, predicting a probability vector indicating the relative position between the peg and hole. It uses a late fusion convolutional neural network (CNN) architecture for classification, extracting the continuous relative position from the probabilities vector. The estimated relative position is then utilized to refine the peg's position during the insertion process. The efficiency of the 3D vision sensor was evaluated across various PIH tasks involving both rigid and deformable objects. Moreover, it was embedded into an industrial robotic cell without additional training in order to validate its robustness and generalization capabilities. The approach demonstrated high success rate, highlighting its effectiveness and applicability in real-world industrial settings.**

***Index Terms*—3D vision sensor, peg-in-hole, robotic insertion, deformable objects, sensor fusion, RGB-D fusion.**

I. Introduction

ROBOTIC integration across industrial applications has long been a main goal, aiming to enhance production efficiency and reduce costs. During production, assembly operations account for 50% of the total production time and 30% of the entire production cycle cost [1].

Peg-In-Hole (PIH) is the most common task in assembly processes, accounting for approximately 40% of total assembly tasks [1]. In addition to the uncertainties in the robotic environment, the PIH problem involves further challenges such as the need for high repeatability, managing surface contact issues, and handling tight assembly tolerances.

To address these challenges, advanced perception and control strategies, particularly for achieving precise alignment, compensation for surface irregularities, and real-time adaptation to sensory, feedback is required. There are several methods for overcoming these challenges in performing PIH tasks. Initial PIH methods relied on remote center compliance to provide appropriate compliance for peg insertion. With the advance in robotic control, recent methods are based on measuring the forces and torques and enforcing the end effector to behave according to the desired compliance or impedance, with parameters determined by reinforcement learning [2], [3].

Other approaches use vision-based methods that utilize cameras and convolutional neural networks (CNNs) to analyze the scene and identify objects [4], [5]. Some techniques also use objects' 3D models to gain additional information such as location and orientation [6], [7].

Recent advancements in 3D camera technologies, driven by affordable depth cameras like Microsoft Kinect [8] and Intel RealSense [9], have enhanced 3D environment perception by adding depth information to RGB images.

The adjustment of state-ot-the-art learning based RGB analysis methods to the additional depth data, involves fusion techniques. These techniques fall into four categories, early and parallel fusion which are more naïve and later and intermediate fusions which are more complex [10].

While existing methods reduce position uncertainties during PIH assembly of rigid objects, they struggle to deal with deformable objects. The assembly of deformable objects introduces additional complexity and requires more advanced control strategies. Unlike rigid objects, deformable objects might change shape during the assembly process, which poses significant modeling, control, and performance challenges [11]. Accurate modeling of deformations, dynamic control of applied forces, and real-time adjustments for precise insertion are among the main obstacles. In response, advanced sensory feedback systems and machine learning algorithms are increasingly used to overcome those challenges. For example, the assembly of rigid pegs to deformable surfaces using learning methods was presented in [12].

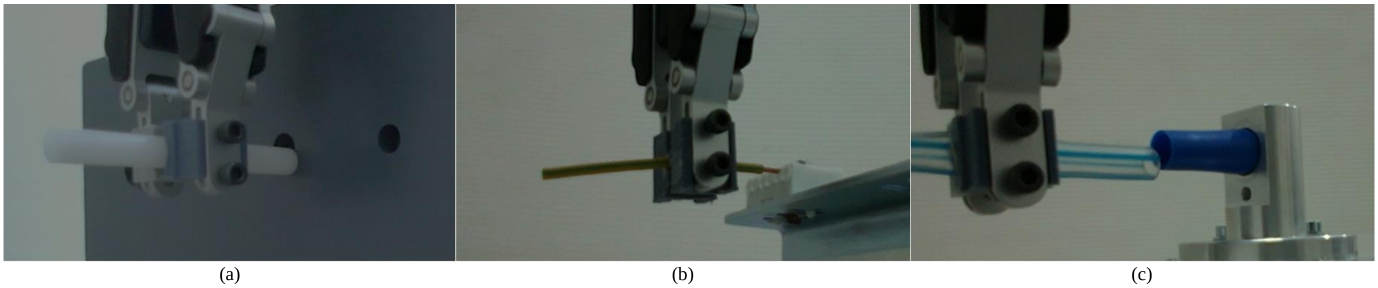
In this work, a learning-based 3D vision sensor is proposed to be integrated into a robotic PIH insertion pipeline for both rigid and deformable objects in real industrial environments. This vision sensor utilizes 3D cameras and learning classification methods for real-time estimation of the relative position between the peg and the hole in the insertion plane.

The work is motivated by the challenge of assembling flexible medical tubes (MD industries [13]). The experiments were conducted in an industrial robotic cell, UR5e, in collaboration with industry. Therefore, it is expected to meet industrial standards of real-time operation (30 FPS), accuracy (maximum error of 1 mm in radius and 10 degrees in angle), and robustness in handling environmental changes and variations in camera viewpoints.

The work was supported by the Israel Innovation Authority as part of the ART (Assembly by Robotic Technology) consortium (grant number 74885).

Sher Hazan, Eylon Cohen, Ronit Schneor and Anath Fischer with the Faculty of Mechanical Engineering, Technion, Israel. sherhazan1115@gmail.com, ela3135@gmail.com, schneor@me.technion.ac.il, meranath@technion.ac.il

Miriam Zacksenhouse are with the Faculty of Mechanical Engineering, Technion, Israel and the Technion Autonomous Systems Program. mermz@technion.ac.il


(a)

(b)

(c)

Fig. 1. The PIH test cases: (a) a classical PIH with rigid objects; (b) wiring a 1 mm thick deformable electric wire to a rigid connector; (c) inserting a deformable medical pipe into a deformable connector.

II. PIH Insertion Pipeline

Our research focuses on estimating the relative position between rigid and deformable objects in PIH tasks performed in real industrial environments and embedding it as a 3D vision sensor in the PIH insertion pipeline in order to improve success rates.

The efficiency of our vision sensor was evaluated through various PIH tasks (Figure 1). These tasks involved handling both rigid and deformable objects, demonstrating the versatility and robustness of the vision sensor.

The proposed pipeline, embedded with the 3D vision sensor, is illustrated in Figure 2. The pipeline includes the following iterative stages:

3D vision sensor: It captures the current 3D scene and estimates the relative position between the peg and the hole in terms of relative radius ($\hat{R}$) and relative angle ($\hat{\theta}$).

Virtual F/T sensor: It transforms the force F_S, and torque T_S measured by OnRobot HEX-E F/T sensor at point P_S, to a virtual point P_V, located at the center of the peg's tip with its z axis aligned in the direction of insertion, to obtain virtual force F_V and virtual torque T_V according to Equation 1.

$$\begin{bmatrix}F_V\\T_V\end{bmatrix}=\begin{bmatrix}F_S\\(P_S-P_V)\times F_S+T_S\end{bmatrix} \tag{1}$$

Virtual friction force sensor: The friction components of F_V, which are used for the robot navigation are replaced with a virtual vision-based friction force $\hat{F}$ representing the direction of the estimated relative angle ($\hat{\theta}$)) extracted by the 3D vision sensor according to Equation 2.

$$\hat{F}=-[sin(\hat{\theta}),cos(\hat{\theta})] \tag{2}$$

Decision Making: Based on the estimated relative radius ($\hat{R}$) from the 3D vision sensor, a decision procedure is performed. While the estimated relative radius ($\hat{R}$) exceeds a predefined threshold, the peg's location is improved. Once the estimated relative radius condition is met, the insertion is initiated.

The threshold value is set to the maximum radial error that ensures successful insertion. For rigid objects, the threshold value is the difference between the radii of the peg and the hole. On the other hand, for deformable objects, insertion experiments with various radii were conducted to determine the appropriate threshold value.

Location improvement: Position based impedance control is implemented to adjust the robot's location in response to the virtual forces and torques $[\hat{F},F_{V,z},T_V]$, with impedance parameters learned using reinforcement learning [3]. The impedance control performs a trade-off between (a) minimizing the position error between the peg and the assumed position of the hole, and (b) minimizing the virtual forces and torques.

An updated 3D scene is then captured, and a new iteration begins.

Insertion: Once the peg and the hole are aligned, the robotic arm performs the insertion.

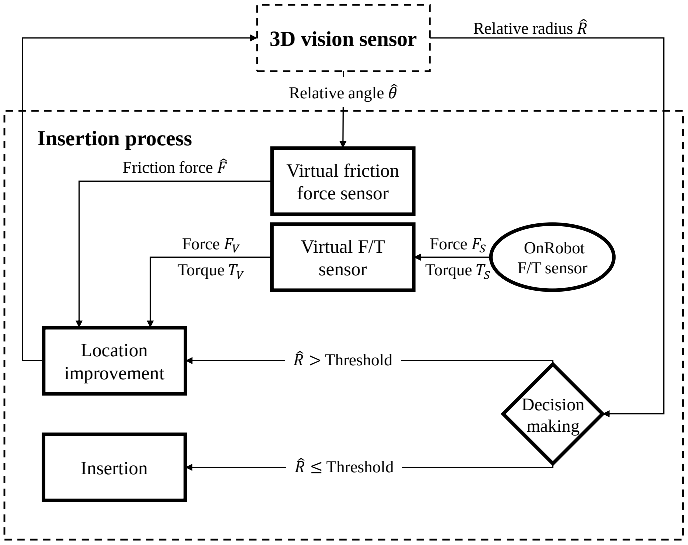


Fig. 2. Block diagram of the entire PIH insertion pipeline embedded with the 3D vision sensor.

III. 3D vision Sensor

The presented 3D vision sensor (Figure 3) utilizes RGB-D data, classification CNN, and linear interpolation to estimate the relative angle ($\hat{\theta}$) and radius ($\hat{R}$) between the peg and the hole in the insertion plain.

In order to apply classification methods to assembly tasks within a continuous domain, a discretization of the scene space is required. This is achieved by dividing the scene space into equal sub-areas as classes where each class corresponds to a range of values in the domain (Figure 4).

Following the discretization, a classification CNN is utilized to predict the probabilities vector of the peg being in each

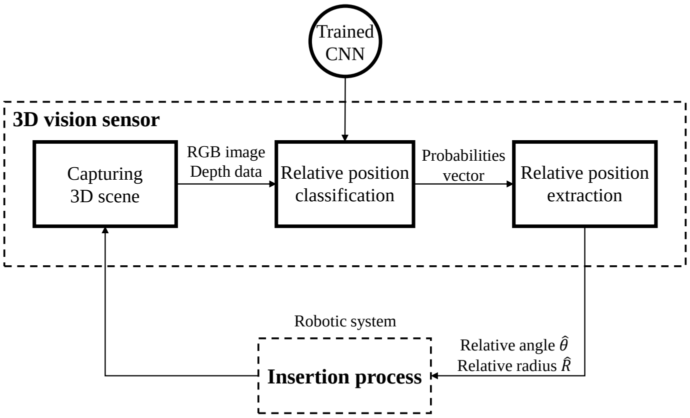


Fig. 3. Block diagram of the proposed 3D vision sensor for estimating relative positions in Peg-in-Hole tasks.

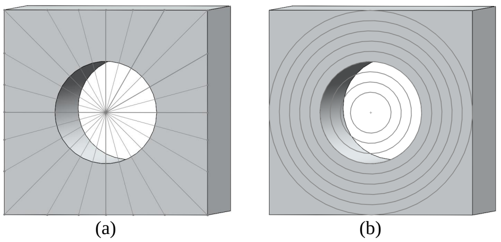


Fig. 4. Discretization of the scene space into sub-areas: (a) according to angle, (b) according to radius.

sub-area. For this task, a late RGB-D fusion architecture was proposed (Figure 5) utilizing two separate ResNet-18 [14] CNNs, one for RGB images and one for depth data. The features from the convolutional layers of both CNNs are concatenated before reaching the fully connected layer. This concatenated feature vector is then inserted into one fully connected layer.

The performance of the proposed model was compared to early and parallel fusion architectures (Figure 6) as well as separated RGB and depth architectures trained on the same dataset. All architectures were based on Resnet-18 architecture due to its advantages of real-time performance, small training dataset requirements, suitability for classification and regression tasks, and ease of adaptation for RGB-D data.

The training dataset consisted of realistic 3D scenes captured in the UR5e robotic cell using a RealSense L515 3D camera mounted on the robotic arm's gripper, focusing on the peg's tip. The capturing procedure included initial calibration where the peg and the hole are aligned, followed by random movements of the robotic arm (angle and radius) within a predefined range, capturing the scene and labeling based on the new position. The known depth range of the scene was used to focus on the region of interest (ROI) by filtering out background objects and noise. For each task, 2500 3D scenes (RGB images and depth data) were captured (Figure 7 (a)). In order to expand the dataset, patches of various sizes and shapes were extracted from these scenes. The ROI was detected from the depth data and the scene was centered accordingly to ensure information retention. From each scene, three unique patches were extracted (Figure 7 (b)), resulting in 7,500 scenes for each task. Additionally, to improve the model generalization, various augmentations were applied during training, including color, lighting, and translation augmentation (Figure 7 (c)).

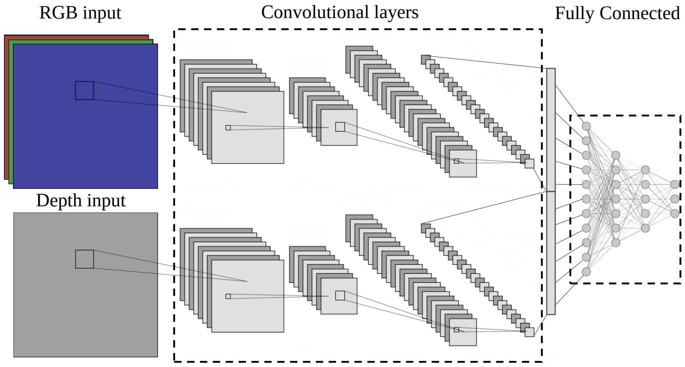


Fig. 5. Architecture of the proposed RGB-D late fusion model using separate RGB and depth ResNet-18 CNNs, followed by feature concatenation and a fully connected layer.

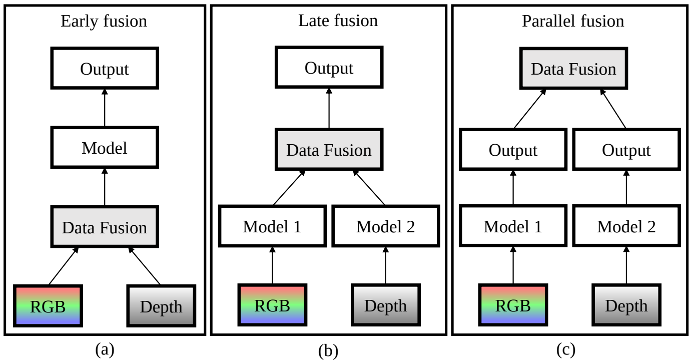


Fig. 6. Illustration of different RGB-D fusion methods: (a) Early fusion, where RGB and depth data are fused at the input level, (b) Late fusion, where features from separate CNNs are concatenated at a later stage, (c) Parallel fusion, where separate CNNs process the data in parallel and the output is fused.

The datasets were utilized to train the classification CNN architectures to predict the probability of being in each sub-area. To ensure comparable results, identical setups and hyperparameters were used for training all architectures.

The predicted probability vector is employed to estimate the continuous relative angle ($\hat{\theta}$) and radius ($\hat{R}$) between the peg and the hole in the insertion plain. This estimation is achieved through linear interpolation, which includes thresholding negligible values to suppress noise, normalizing the filtered probability vector, and calculating the mean value across all classes using the probability vector (Equation 3).

$$\hat{X} = \sum_{i=1}^{n} \frac{D}{n} \left(i - \frac{1}{2} \right) P(x_i) \qquad (3)$$

Where D is the domain, n is the number of classes, and $P(x_i)$ is the probability for class i.

Since the angular domain is circular, the mean values were calculated separately for $sin(\theta)$ and $cos(\theta)$ instead for θ, and the estimated angle $\hat{\theta}$ was derived from these values.

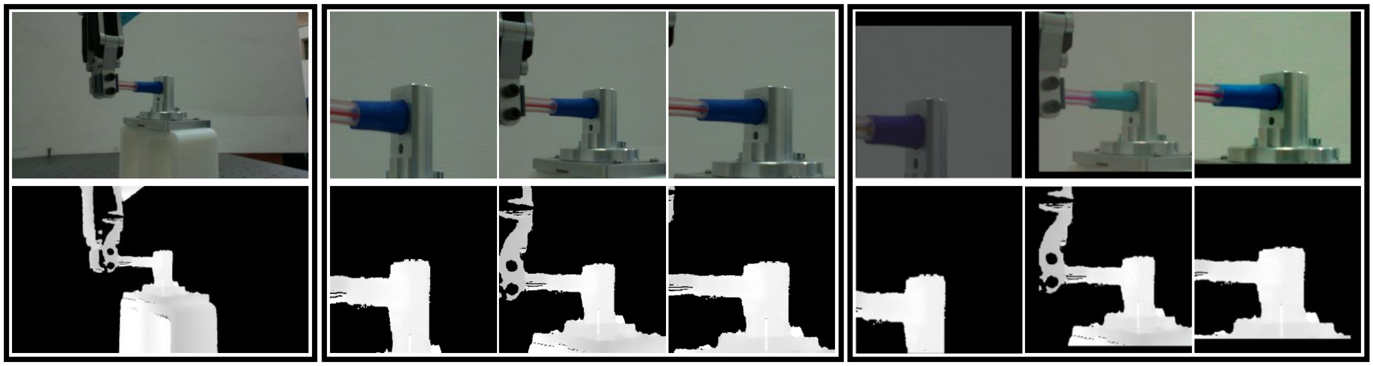

(a) (b) (c)

Fig. 7. (a) Sample of depth data and RGB images of the captured 3D scenes. (b) The patches were extracted from the scene sample. (c) The same patches after applying the augmentations.

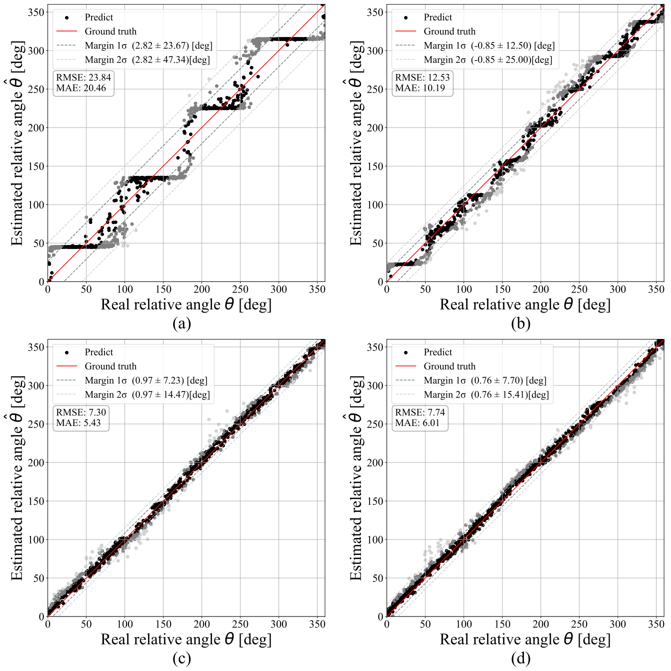
IV. Results

In order to fine-tune the division parameter, which divides the scene into sub-areas, a selected model was trained with various division parameters, and the estimation error was evaluated. The performance is summarized in Table I. It is indicated that dividing the scene into 16 sub-areas produced the best results for our dataset size. Smaller division parameters resulted in high estimation error due to high precision classification, leading to high probability identification of the sub-areas. In such cases, the linear interpolation is ineffective, and the estimation becomes the class itself, as shown in Figure 8 as steps in the graphs. Conversely, larger division parameters led to overfitting due to fewer examples per class in the training set, causing increased error.

TABLE I
Performance comparison based on MAE of relative angle estimation for the Medical pipe task using the RGB model trained with different scene division parameters.

No. of sub areas	4	8	16	32
MAE [deg]	20.46	10.19	**5.43**	6.01

The evaluation of the relative position estimation involves analyzing the performance of the various CNN architectures for each task and assessing the accuracy of the relative radius and angle estimations. The results are summarized in the following tables, are based on mean absolute error (MAE) as the evaluation metric. The error in the estimated angle (Table II) is measured in degrees and the error in the estimated radius (Table III) is measured in millimeters. The proposed 3D vision sensor demonstrates promising results, as both angle and radius errors are below the predefined industrial error (1 [mm] in radius and 10° in angle). Moreover, it was found that the late fusion architecture yields the best performance.



(a) (b) (c) (d)

Fig. 8. Angle estimation results of the RGB model with scene divisions into (a) 4, (b) 8, (c) 16, and (d) 32 sub-areas. See Figure 9 for notation explanations.

TABLE II
Performance comparison based on MAE of relative angle estimation across the various CNN architectures and tasks using our vision sensor.

	RGB	Depth	Early fusion	Late fusion	parallel fusion
Classical PIH	6.65	25.38	8.43	**5.30**	11.78
Electric wire	25.13	22.72	11.04	**6.88**	12.09
Medical pipe	5.43	10.86	5.37	**4.75**	6.96

In addition, a comparison to regression CNN architectures was performed. The Medical Pipe task was selected for this comparison, and regression CNN models were trained on the same Medical Pipe dataset used to train the vision sensor. The performance of the models is summarized in the following tables. As can be seen, the 3D vision sensor estimates the angle more accurately than the regression model (Table IV). Conversely, the regression model provides a better estimation of the radius (Table V). Since the estimated angle is a crucial parameter in the insertion process and the radius serves as an indicator, the vision sensor proves its effectiveness in the PIH insertion tasks.

The performance of the relative position estimation was evaluated for all models. The performance graphs for the Medical Pipe task, our primary task, are presented below.

TABLE III
PERFORMANCE COMPARISON BASED ON MAE OF RELATIVE RADIUS ESTIMATION ACROSS THE VARIOUS CNN ARCHITECTURES AND TASKS USING OUR VISION SENSOR.

	RGB	Depth	Early fusion	Late fusion	parallel fusion
Classical PIH	0.59	1.03	0.64	0.61	**0.53**
Electric wire	0.36	0.39	**0.31**	0.32	0.35
Medical pipe	0.56	0.78	**0.45**	0.49	0.57

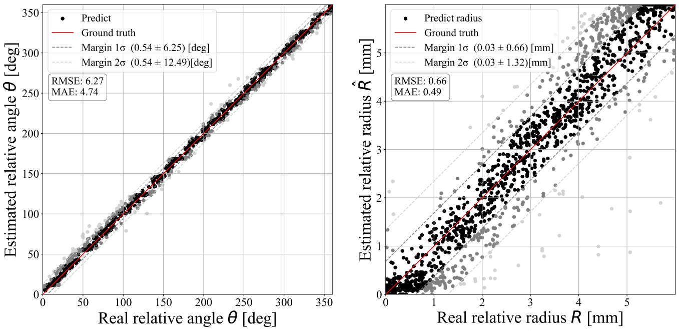
TABLE IV
COMPARISON BETWEEN OUR VISION SENSOR AND REGRESSION MODELS, BASED ON MAE OF RELATIVE ANGLE ESTIMATION FOR THE MEDICAL PIP TASK ACROSS THE VARIOUS CNN ARCHITECTURES.

	RGB	Depth	Early fusion	Late fusion	parallel fusion
Vision sensor	5.43	10.86	5.37	**4.75**	6.96
Regression	17.16	27.38	28.93	17.59	**16.80**

TABLE V
COMPARISON BETWEEN OUR VISION SENSOR AND REGRESSION MODELS, BASED ON MAE OF RELATIVE RADIUS ESTIMATION FOR THE MEDICAL PIP TASK ACROSS THE VARIOUS CNN ARCHITECTURES.

	RGB	Depth	Early fusion	Late fusion	parallel fusion
Vision sensor	0.56	0.78	**0.45**	0.49	0.57
Regression	0.31	1.01	0.39	**0.29**	0.51

Figures 9 (a) and (b) showcase the estimated relative angle and radius with respect to the ground truth (GT). As can be seen, the angle estimation yields high performance with an MAE of less than 5 degrees. Additionally, while the radius estimation is less accurate, it still meets the predefined goal of less than 1 [mm] error.

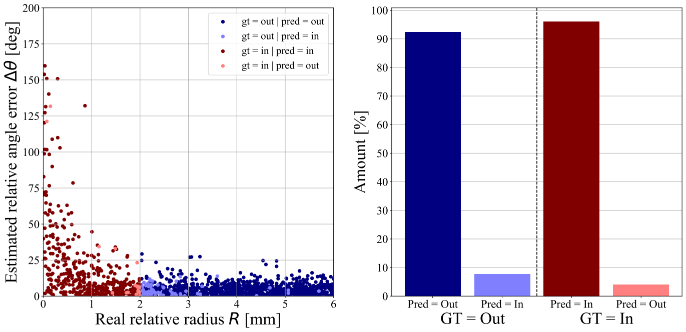


(a) (b)

Fig. 9. Comparison of ground truth and estimated values for the Medical Pipe task: (a) Relative angle estimation, (b) Relative radius estimation. The red line represents the ground truth, indicating where the points should ideally lie. The gray lines indicate the first and second margins, encompassing where 95% and 68% of the points are located, respectively.

Since the estimated relative radius was used as an indicative value, it was converted into a binary value *In*/*Out*. If the radius exceeds the threshold, the binary value is classified as *Out*, indicating the peg is outside the hole. Conversely, if it is below the threshold, the binary value is classified as *In*, signifying the peg is aligned with the hole and the insertion can be performed. Figure 10 (a) showcases the error in the estimated angle relative to the actual radius. The graph indicates that for small radii, a high error in the estimated angle is obtained. Additionally, the graph shows that the error in the *In*/*Out* binary value is distributed around a 2 mm radius, reflecting the 2 mm threshold set for this task.

Another perspective on this data is provided in Figure 10 (b), showing the prediction accuracy of the *In*/*Out* binary value as a percentage. While approximately 5% of the *In* samples were incorrectly predicted, these results do not impact the process. The robotic arm will continue to improve the peg location until an *In* prediction is received. Conversely, approximately 8% of the *Out* samples were incorrectly predicted. In these cases, the robotic arm will initiate the insertion outside the hole, leading to a failed insertion. To minimize this error, the insertion is initiated only after receiving two consecutive *In* predictions, reducing the chance of inserting the peg outside the hole to below 1%.



(a) (b)

Fig. 10. Analysis of errors in the estimated: (a) Angle estimation error versus actual radius, (b) In/Out binary classification accuracy for the insertion task. In both graphs: The blue and red colors indicate *In* and *Out* ground truth samples respectively. The Dark and Bright colors indicate correct and incorrect predictions respectively. (For example, dark red color indicates *In* ground truth sample that correctly predicted as *In*).

Another way to represent the estimation error of the radius and angle is in the Cartesian coordinate system. The following graphs showcase the error in the estimation of the relative angle (Figure 11 (a)) and the estimation of the relative radius (Figure 11 (b)) based on the peg location on the insertion plain, where the origin represents the center of the hole, and the points indicate the center of the peg. The estimated relative radius error graph shows that the error in the estimated relative radius does not correspond to the peg location. The error may arise from factors such as image quality and capturing angle. Additionally, as previously shown, the estimated relative angle error is high for small radii. However, in such a case, the peg is already initiated. Examining the rest of the estimated relative angle errors, it can be seen that it is distributed similarly to the radius estimation error, as poor image quality would affect both parameters.

To validate the real-world applicability and the generalization of the vision sensor, a performance test was conducted in industrial settings. This test focused on the Medical Pipe task and was executed without additional training at Polygon Technologies [15]. The results of these experiments are depicted in Figures 12 (a) and (b). As evident, the results were promising even without additional training, highlighting the generalization capability of the proposed vision sensor.

The crucial test of the 3D vision sensor lies in validating it across the insertion pipeline. For this validation, insertion experiments were conducted both with and without the vision sensor. The results are summarized in Table VI. As shown in

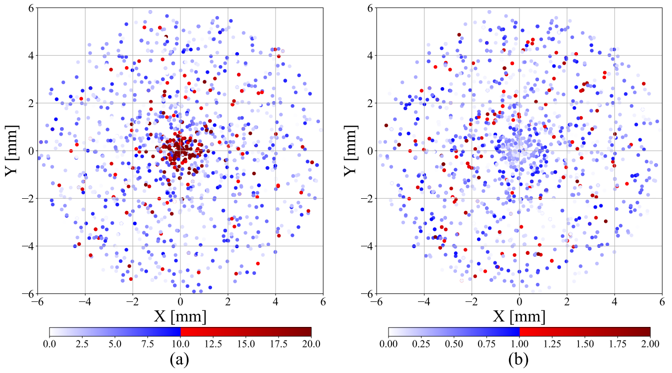


Fig. 11. Estimation errors based on the peg location in the insertion plane: (a) Relative angle error, (b) Relative radius error. The blue and red points indicate estimation error below and above the predefined industrial error respectively (1 [mm] in radius and 10° in angle).

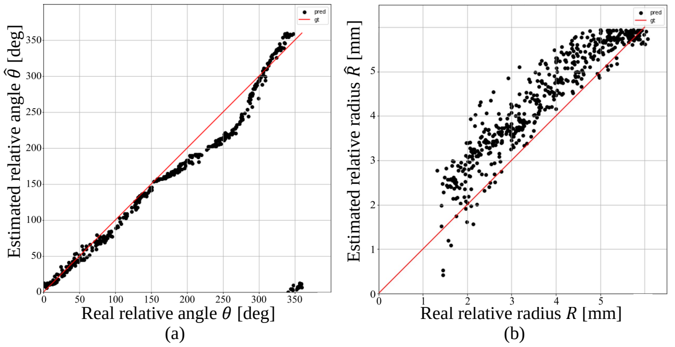


Fig. 12. Performance of the vision sensor estimation for the Medical Pipe task in industrial settings: (a) Relative angle estimation, (b) Relative radius estimation. See Figure 9 for notation explanations.

the table, for the classic PIH task with a 4.5 mm peg and a 6 mm hole, the vision sensor enhanced the success rate from 83% to 100%. Similar improvements were observed for an 8 mm peg and a 10 mm hole. For a 16 mm peg and a 20 mm hole, the success rate rose from 51% to 90%. Additionally, for the Medical Pipe task, the success rate in the insertion process increased from 80% to 98%. Furthermore, the experiments with the vision sensor were conducted with increased initial radial uncertainty (R_{err}) in the hole position, which further validates the effectiveness of the vision sensor.

TABLE VI

SUCCESS RATE OF PIH INSERTION EXPERIMENT WITH AND WITHOUT OUR VISION SENSOR, ON A UR5E ROBOTIC CELL EQUIPPED WITH ONROBOT F/T SENSOR AND REALSENSE L515 CAMERA. SUCCESS RATES WERE EVALUATED FROM 100 TRIALS. ERRORS WERE UNIFORMLY DISTRIBUTED IN THE INDICATED RANGES IN THE RADIAL POSITION OF THE HOLE R_{err}.

Task	Sensors	R_{err} [mm]	Success rate [%]
4.5[mm] Peg	Force	[1, 2.5]	83
6[mm] Hole	Force + Vision	[1, 3]	**100**
8[mm] Peg	Force	[1.5, 3.5]	80
10[mm] Hole	Force + Vision	[1.5, 10]	**100**
16[mm] Peg	Force	[2.5, 8]	51
20[mm] Hole	Force + Vision	[4, 14]	**90**
8.5[mm] Medical pipe	Force	[1.5, 2.5]	80
8.5[mm] Connector	Force + Vision	[1.5, 2.5]	**95**

V. CONCLUSION

This research introduces a classification approach for estimating the relative position between objects, with a focus on deformable objects. The objective was to develop and integrate a 3D vision sensor within the Peg-in-Hole (PIH) insertion pipeline and to evaluate its performance across various PIH assembly tasks in a real industrial robotic environment. The proposed 3D vision sensor captures RGB-D data and estimates the relative position between the peg and the hole, subsequently guiding the insertion operations based on this estimation. The proposed vision sensor was tested in an industrial robotic cell, yielding promising results that demonstrated its effectiveness without the need for additional training.

Satisfactory results were achieved in all assembly tasks for both angle and radius estimation. The approach provided high precision in estimating the relative angle compared to conventional regression methods. However, the regression method outperformed in radius estimation, which can be attributed to the limitations of the linear interpolation implementation. While the angles domain is circular, the radius domain has bounds that cannot be fully captured. Although radius estimation served only as the insertion criterion in our tasks, it may play a more crucial role in other scenarios.

The rigid and deformable PIH insertion experiments demonstrated the applicability of the proposed approach in the real-world industrial environment. The outcomes showed a significant improvement in the success rate of the insertion operation across all tasks. This indicates that the 3D vision sensor not only enhances the precision of relative position estimation but also significantly boosts the overall efficiency and reliability of PIH insertion processes.

In conclusion, the classification-based 3D vision sensor presents a viable solution for improving PIH insertion tasks in industrial settings. Its ability to provide accurate relative position estimations and enhance insertion success rates highlights its potential for broader applications in automated assembly processes involving both rigid and deformable objects.

REFERENCES

[1] J. Jiang, Z. Huang, Z. Bi, X. Ma, and G. Yu, “State-of-the-art control strategies for robotic pih assembly,” *Robotics and Computer-Integrated Manufacturing*, vol. 65, p. 101894, 2020.

[2] Í. Elguea-Aguinaco, A. Serrano-Muñoz, D. Chrysostomou, I. Inziarte-Hidalgo, S. Bøgh, and N. Arana-Arexolaleiba, “A review on reinforcement learning for contact-rich robotic manipulation tasks,” *Robotics and Computer-Integrated Manufacturing*, vol. 81, p. 102517, 2023.

[3] S. Kozlovsky, E. Newman, and M. Zacksenhouse, “Reinforcement learning of impedance policies for peg-in-hole tasks: Role of asymmetric matrices,” *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 898–10 905, 2022.

[4] Y. Shen, Q. Jia, R. Wang, Z. Huang, and G. Chen, “Learning-based visual servoing for high-precision peg-in-hole assembly,” in *Actuators*, vol. 12, no. 4. MDPI, 2023, p. 144.

[5] M. T. Shahria, M. S. H. Sunny, M. I. I. Zarif, J. Ghommam, S. I. Ahamed, and M. H. Rahman, “A comprehensive review of vision-based robotic applications: Current state, components, approaches, barriers, and potential solutions,” *Robotics*, vol. 11, no. 6, p. 139, 2022.

[6] H.-C. Song, Y.-L. Kim, and J.-B. Song, “Automated guidance of peg-in-hole assembly tasks for complex-shaped parts,” in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 4517–4522.

[7] M. C. Leu, H. A. ElMaraghy, A. Y. Nee, S. K. Ong, M. Lanzetta, M. Putz, W. Zhu, and A. Bernard, “Cad model based virtual assembly simulation, planning and training,” *CIRP Annals*, vol. 62, no. 2, pp. 799–822, 2013.

[8] Z. Zhang, “Microsoft kinect sensor and its effect,” *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[9] L. Keselman, J. Iselin Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, “Intel realsense stereoscopic depth cameras,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 1–10.

[10] D. Ramachandram and G. W. Taylor, “Deep multimodal learning: A survey on recent advances and trends,” *IEEE signal processing magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[11] X. Yanchun, B. Yuewei, and H. Yafei, “Assembly strategy study on the elastic deformable peg in hole,” in *2010 The 2nd International Conference on Industrial Mechatronics and Automation*, vol. 1. IEEE, 2010, pp. 193–197.

[12] J. Luo, E. Solowjow, C. Wen, J. A. Ojea, and A. M. Agogino, “Deep reinforcement learning for robotic assembly of mixed deformable and rigid objects,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 2062–2069.

[13] Mdc, “Mdc industries,” http://www.mdcindustries.com, 2022.

[14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] “polygon technologies,” https://www.polygon-technologies.com/.