RESEARCH THE CONSTRUCTION OF A TRANSPARENT OBJECT RECOGNITION MODEL USING COMPUTER VISION AND ARTIFICIAL INTELLIGENCE

Dien Thi Hong Ha

University of Economics - Technology for Industries, Hanoi, Vietnam

ARTICLE INFORMATION

Journal: *Vinh University Journal of Science Natural Sciences, Engineering and Technology*
p-ISSN*: 3030-4563*
e-ISSN: *3030-4180*

Volume: 53
Issue: 2A
****Correspondence:***
dthha@uneti.edu.vn

Received: *08 January 2024*
Accepted: *26 March 2024*
Published: *20 June 2024*

Citation:
Dien Thi Hong Ha (2024). Research the construction of a transparent object recognition model using computer vision and artificial intelligence. ***Vinh Uni. J. Sci.*** *Vol. 53 (2A), pp. 37-48*
doi: 10.56824/vujs.2024a004

OPEN ACCESS

ABSTRACT

The article focuses on researching the construction of a transparent object (glass) recognition model based on the application of computer vision techniques and artificial intelligence models. Stereo Matching image processing techniques have been used to build a raw depth image from a Stereo Camera. The goal is to reconstruct the depth image, recover in-depth information, and generate a complete depth image to effectively identify the position of transparent objects in reality. Additionally, the research involves designing a software interface for observing depth images, point clouds, and controlling the robotic arm for object grasping in three-dimensional space. The following results were obtained: The quality of the depth image reconstruction model is improved compared to the ClearGrasp model when evaluated on ClearGrasp datasets; Determine orientations on how to improve models and algorithms for reconstructing depth images in a more quantitative manner. The success rate of picking up a glass cup is over 90% in cases of objects on the floor; This rate reaches over 70% when objects are placed at different heights. The software interface displays detailed information and facilitates communication, controlling depth images, point clouds, and position graphs (x, y, z). At the same time, it is easy to interact and convenient during the experiment.

Keywords: Object recognition; transparent objects; stereoscopic vision; depth image; artificial intelligence.

1. Introduction

In everyday life, transparent objects such as glasses, bottles or glass panels are widely used. However, recognizing and locating the position of these objects poses a significant challenge for computer systems and autonomous robots. This challenge has spurred research on transparent object detection models to provide breakthrough solutions for automating processes such as production, transportation, or customer service. Typically, objects have Lambertian surfaces, reflecting light uniformly from all directions, resulting in even surface brightness from all angles. However, light is mostly diffused and refracted rather than reflected in transparent objects, making it difficult for image sensors to produce accurate depth images. Conventional infrared sensors often fail to receive signals at the receiver end, resulting in undefined depth values. Meanwhile, Stereo Cameras face challenges such as pixel values at the positions of transparent objects being confused with those of objects behind them. This leads to unclear and information-deprived depth images, creating difficulties in determining the positions of transparent objects in three-dimensional space. Therefore, algorithms and models are needed to process and reconstruct depth images for transparent object recognition. This model will utilize raw depth images obtained from Stereo or RGB-D Cameras, combined with deep learning models to estimate surface (Normal estimation), detect boundaries (Boundary detection), segment objects (Semantic segmentation), and recognize objects (Object detection) to reconstruct and recover depth information of transparent objects. This aims to produce accurate depth images, enhancing the ability to recognize and apply in real-world scenarios.

Some related research includes:

- Ilya Lysenkov and colleagues identified and estimated the pose of transparent objects using the Kinect Sensor [1]. Their algorithm enabled a robot to identify and grasp transparent objects with an accuracy of up to 80% in various environmental conditions and backgrounds. While their method struggles with overlapping transparent objects, it performs well with transparent and non-transparent objects scattered in the environment.

- Chen Guo-Hua and colleagues recognized and located transparent objects using an RGB-D Camera [2]. Their method achieved an accurate detection rate of up to 87.64% in an environment containing both transparent and non-transparent objects. However, the shape of transparent objects determined by this method is limited to cylindrical objects.

- Thomas Weng and colleagues developed a well-known robotic arm for object grasping [3]. The introduced method is a diverse recognition model learned based on an existing single model. This transfer learning approach relies on an already available single model and a diverse dataset of images for training, with labels indicating success in accurately capturing objects and grasping efforts. Their system is not yet truly perfected as the accuracy has not reached the desired high level.

- The research by Shreeyak S. Sajjan and colleagues introduced a method for estimating the 3D spatial shape of transparent objects from a single image without knowledge of the object's internal structure [4]. The ClearGrasp model demonstrated relatively accurate estimations of the 3D spatial shape of transparent objects in different datasets.

Through a survey of typical research projects and based on hardware and software constraints, the ClearGrasp model has been chosen for enhancement in this paper. These enhancements aim to improve the model's quality on the ClearGrasp dataset, even with significantly less training data than other methods. Additionally, when used with the authors' experimental hardware model, the success rate and accuracy are expected to be 3-5% higher.

2. Theoretical background and implementation

2.1. Stereo Matching model

Stereo Matching is a technique in the field of image processing used to compute the disparity between two images captured from different perspectives of the same scene [5-6]. This technique relies on comparing corresponding pixels in both images to identify pixels with the same depth. To create a *Stereo Matching* model, two steps are required: generating a disparity map that shows the differences between the same point in the real world in the left and right frames of the Stereo Camera, and creating a depth map with each pixel value representing the distance from the Stereo Camera to that point.

2.2. Depth Image Reconstruction models

2.2.1. ClearGrasp model

Early studies on reconstructing depth from sparse depth captured by RGB-D sensors for common objects [7-8] revealed that training a neural network based on RGB color images is more effective than using sparse depth images. This result suggests that One can train a network to predict normal surface and boundary rules solely from color and use the observed depth as a normalization factor when inverting depth from the rules. ClearGrasp model consists of three smaller models: surface estimation, boundary detection, and segmentation, along with a global optimization algorithm.

2.2.2. ClueDepth Grasp model

ClueDepth introduces several improvements over ClearGrasp [9-11]: In the input depth image, the model utilizes a ClueDepth Map module to retain depth values considered valuable for reuse within the transparent object region, instead of completely eliminating regions identified as transparent objects, as done by ClearGrasp. After passing through CNN edge detection and surface estimation networks, the results are based on RGB images for prediction, without calculating Occlusion Weights or converting results to the HDF5 format. The ClueDepth model provides better results than ClearGrasp in terms of both accuracy and performance.

2.2.3. DFNet model

Another approach to solving the transparent object depth reconstruction problem is applying a CNN-based model similar to the U-Net architecture [12] but with more complex Encoder and Decoder sections, including larger-scale features to extract lower-scale features.

2.2.4. TODE-Trans model

Another model following the Encoder-Decoder approach is TODE-Trans. Its novelty compared to DFNet lies in applying the Swin Transformer in the Encoder section to extract global features of the object.

2.2.5. Object Detect3ion and Segmentation model (YOLOv7)

ClearGrasp's deep learning models lack an object detection model. They use segmentation and edge detection models to determine the position of a cup during grasping. However, when testing in real-world scenarios, accurately determining the cup's position using only these two models is challenging, especially in cases of overlapping objects. Especially in cases where multiple objects overlap, the edge detection model needs to be highly accurate to determine the precise position of each object. In this scenario, the segmentation model is Semantic Segmentation, which can only distinguish between transparent objects and the background but cannot differentiate between individual objects.

Therefore, the authors decided to enhance the model by adding an object detection model to increase accuracy in determining the position of transparent objects. The selected object detection model is YOLOv7 [8], [13], the latest version of YOLO at the time of the research.

A potential future development of the depth image reconstruction model, the Yolov7 Segmentation model, is used to replace two models in the proposed model based on ClearGrasp: the semantic segmentation model and the object detection model. This replacement aims to increase accuracy and, notably, improve the model's processing time [14-20].

2.3. Hardware and software design

2.3.1. Stereo camera

The stereo camera was constructed by combining two identical USB cameras (Figure 1). Each camera measures 70x30 mm with a resolution of 640x480 pixels. The distance between the two camera lenses is 70 mm, roughly equivalent to the distance between the eyes of an adult.

2.3.2. Robot arm

Designed to experimentally evaluate the three-dimensional spatial accuracy of detected transparent objects. The chosen robot arm model is the Community robotic arm (Figure 2) with three degrees of freedom and a gripper, sufficient for grasping transparent objects on a flat surface such as transparent glass or plastic cups.

Figure 1: *Stereo camera*

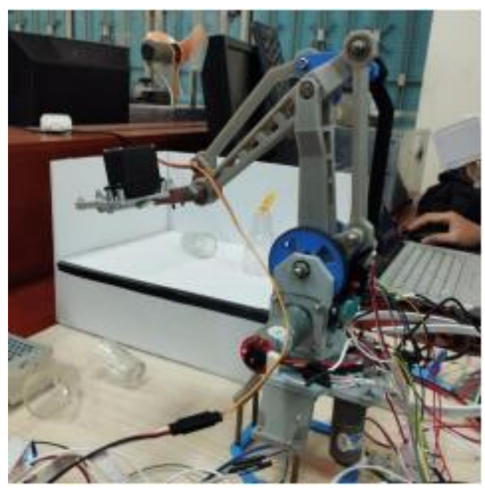

Figure 2: *Robot arm*

2.3.3. Experimental environment

The experimental environment involves a transparent object-grasping robotic arm, a three-degree-of-freedom robot arm, a stereo camera, and the space where the object is placed within the observation range of the stereo camera with dimensions of 30x20 cm, featuring a background in either white or black, and illuminated (Figure 3 (a, b)).

Two transparent object samples were used in the experiment, both being small-sized glass cups suitable for the robot arm to grasp.

(a)

(b)

Figure 3 (a, b): *Experiment setup of the transparent object-grasping robot arm*

2.3.4. Software

a. Backend - server running the recognition model: This research utilizes Google Colab, a free service from Google that provides a Jupyter Notebook environment for easy writing, execution, and sharing of Python source code on the cloud platform.

b. Frontend - observation and control interface: The program on the computer is developed using the Python language and the pyQt5 library. This software includes display and interaction features.

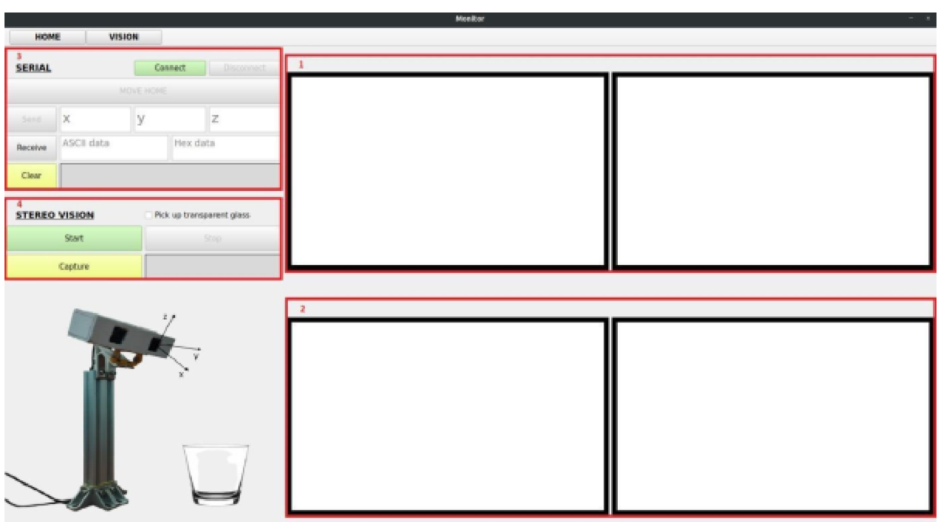

Figure 4: *Homepage interface*

Figure 4 depicts the display of images from two cameras, depth images before and after reconstruction, the serial communication interface with the Robot controller, and the display interface.

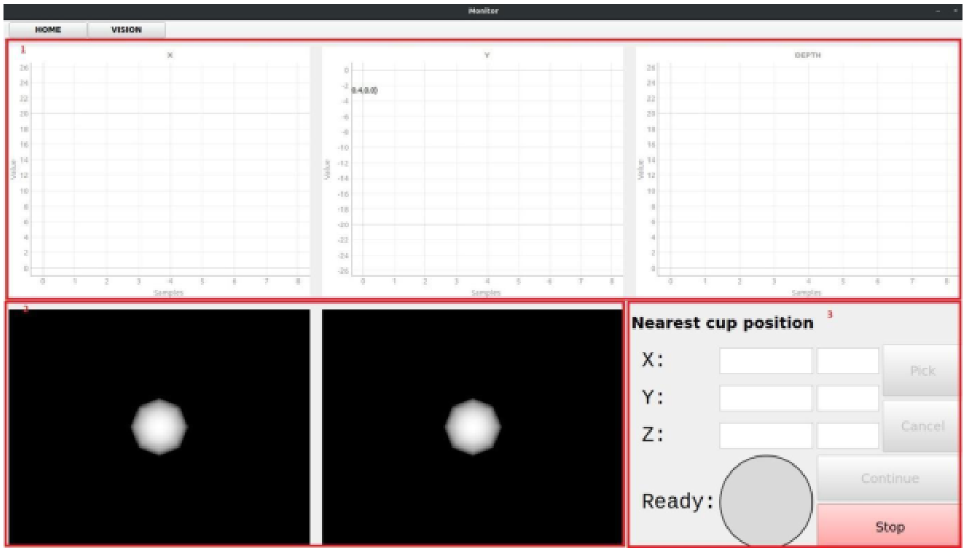

Figure 5: *Visionpage interface*

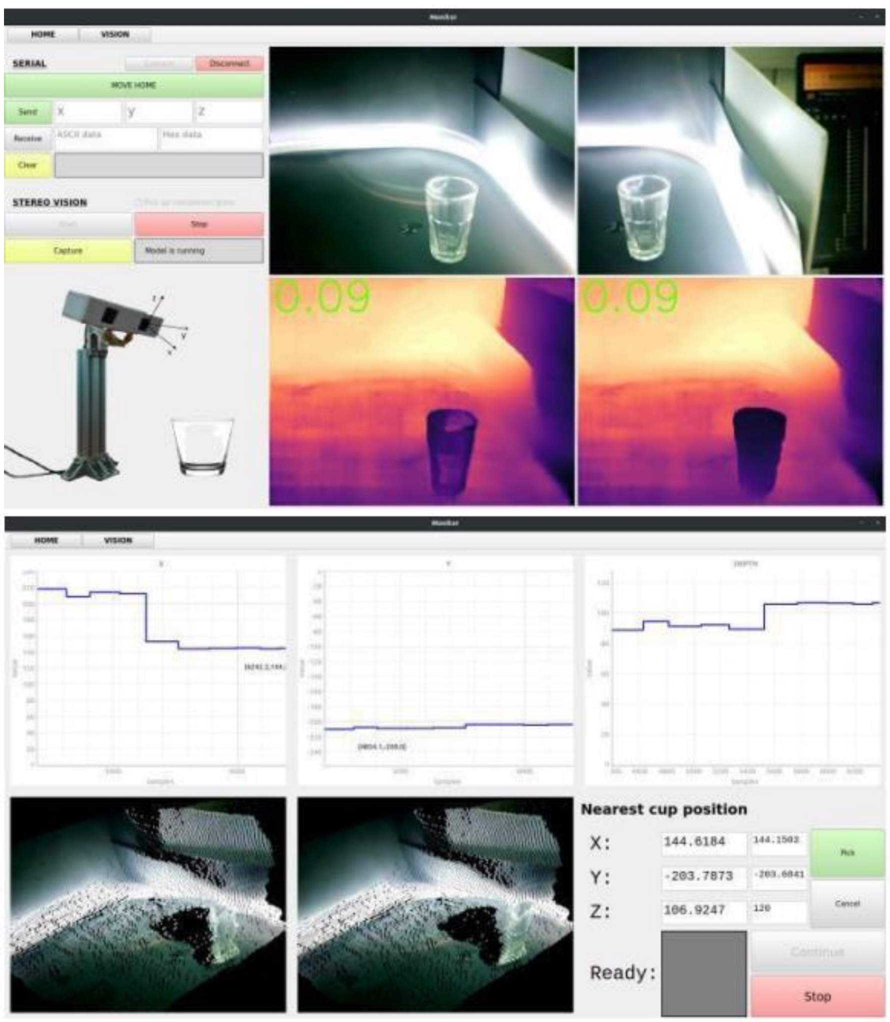

Figure 6: *Interface during real-time operation*

Figure 5 represents the Vision page interface, which includes graphs of position coordinates x, y, z over time and displays the Point Cloud. Figure 6 is the interface when in actual operation.

3. Model construction and results evaluation

3.1. Building the proposed model based on ClearGrasp

Through literature review and experimentation, our research utilizes ClearGrasp to propose an improved model (Figure 7). Subsequently, the detailed implementation involves three small models and further enhancements by incorporating external models. Specifically, these include: *Surface Estimation Model; Boundary Detection Model; Semantic Segmentation Model, and Object Detection Model*. Among these, the Yolov7 segmentation model is used to replace two small models in the initial depth image reconstruction model, namely the segmentation model and the object detection model. ClearGrasp provides data for training surface estimation, boundary detection, and segmentation models. The dataset comprises 74 GB of data, featuring 5 transparent object samples, totaling over 50 thousand images generated by Blender software and 286 real images captured by Intel Realsense D415, D435 RGB-D cameras.

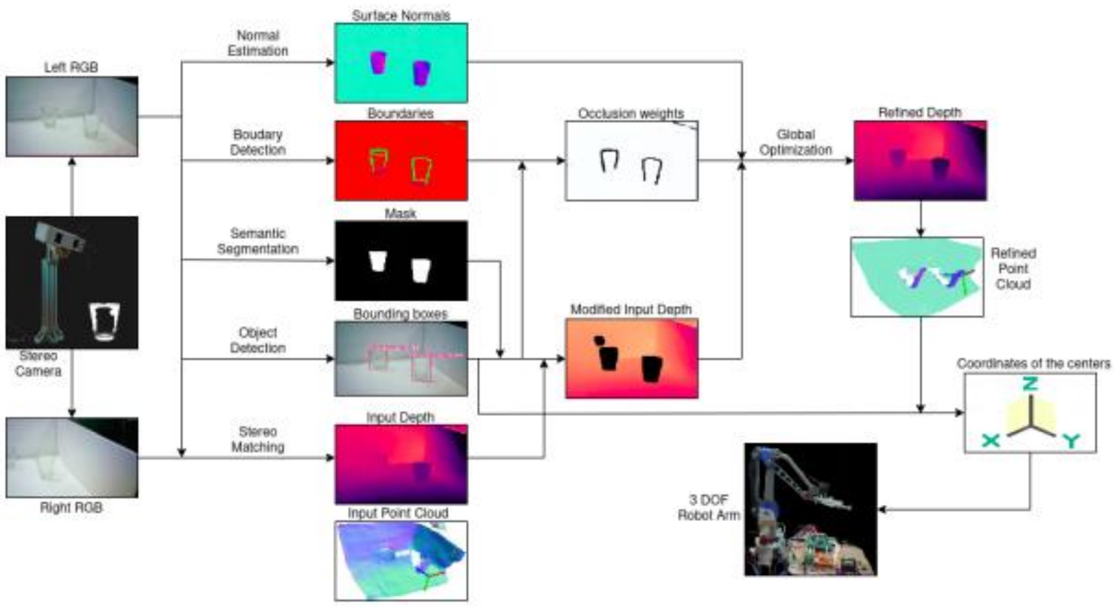

Figure 7: *Model structure based on ClearGrasp*

▪ **Dataset:** The group utilizes the Validation and Test sets of the ClearGrasp dataset to evaluate the networks. The predicted outputs of the CNN networks are resized to 256x256, while the final output is further adjusted to 144x256 to ensure fairness.

▪ **Evaluation metrics:** For the Semantic Segmentation task, the following metrics are employed: Accuracy, F1 Score, Precision, Recall, IOU, with IOU and F1 Score being the two most critical metrics. For Boundary Detection, commonly used evaluation metrics are ODS and OIS. However, with the addition of predicting a new class, the 'Contact Edge,' predicting boundaries can be considered equivalent to semantic segmentation for the three classes.

3.2. Evaluation of the quality of Depth Image Reconstruction model based on ClearGrasp

The quality evaluation results for each sub-model are presented in Tables 1, 2, and 3.

a. Surface estimation model (Normal estimation model)

Table 1: *The quality of the surface estimation model*

Models	Dataset training of ClearGrasp	Test	Scores achieved				
			Mean	Med	11.25^0	22.25^0	30^0
Bae_et_al	17483	Real Novel	24.28	19.19	31.73	60.05	72.10
DeepLab V3+	>50000		22.29	18.09	31.63	63.44	76.06
Bae_et_al	17483	Real Known	21.80	18.31	33.59	64.82	76.84
DeepLab V3+	>50000		21.93	18.72	32.82	64.39	76.05

b. Boundary detection model

Table 2: *Quality semantic of edge detection model*

Models	Dataset Training of ClearGrasp	Test	Scores Achieved			
			F1 Score	Precision	Recall	IOU
BiDiNet	15540	Real Novel	23.85	49.12	73.60	22.83
DeepLab V3+	>50000		24.06	49.12	74.02	23.21
BiDiNet	15540	Real Known	23.96	49.02	73.91	23.01
DeepLab V3+	>50000		24.08	49.02	74.16	23.24

c. Semantic segmentation model

Table 3: *The quality of segmentation model*

Models	Dataset Training of ClearGrasp	Test	Scores Achieved				
			Accuracy	F1 Score	Precision	Recall	IOU
MKDCNet	17483	Real Novel	96.09	63.66	51.82	97.73	50.56
DeepLab V3+	>50000		95.84	61.79	49.28	97.63	48.16
MKDCNet	17483	Real Known	97.68	71.18	71.18	93.08	67.06
DeepLab V3+	>50000		96.78	72.37	64.01	95.35	61.72

Where:

▪ ***Real Novel*** refers to test data on which the model has never been trained. This helps assess the model's generalization ability, i.e., its capability to apply learning outcomes from the training data to new data that the model has not seen before.

▪ ***Real Known*** represents test data that the model has been trained on or similar data to the training data. This helps evaluate the model's performance on data it has been trained on and also helps determine whether the model is overfitting.

4. Conclusion

The article presented the results of research on building a Stereo Matching model to construct raw depth images from a Stereo Camera. The study includes constructing the model, depth image reconstruction algorithm, depth information recovery, and generating complete depth images, successfully identifying the positions of two transparent glass samples prepared in real-world scenarios. An interface has also been designed for observing depth images, point clouds, and controlling the robotic arm for object grasping in three-dimensional space. Regarding the algorithm for transparent object detection using the Stereo Camera, the study is divided into two parts: Stereo Matching and depth image reconstruction. Since the research uses two single cameras instead of a Depth Camera (RGB-D Camera) like most previous studies, a Stereo Matching model is needed to create depth images from left-right image pairs. Additionally, a depth image reconstruction model helps improve the accuracy of depth images from the Stereo Matching model to precisely identify the depth of transparent objects.

- Stereo Matching model: Constructs depth images from the Stereo Camera. Images from two cameras are preprocessed to align frames. The group then explores and experiments with various stereo matching models, including image processing algorithms and deep learning models like HITNet, CRNet, and PCW-Net.

- Depth Image Reconstruction model: To enhance the accuracy of depth images for transparent objects, after considering various depth image reconstruction models, this study decides to use the ClearGrasp model as a foundation. The ClearGrasp model consists of small models for surface estimation, edge detection, and segmentation. The group's approach includes retraining the model with different datasets such as ClearGrasp and Super Caustics, testing alternative edge detection and segmentation models for those available in ClearGrasp, adding an Object Detection model (YOLOv7), and incorporating image processing algorithms to improve the overall model's recognition capabilities.

After this process, the study explores newer depth image reconstruction models like ClueDepth Grasp and DFNet to further enhance the model. The transparent object detection model, especially the depth image reconstruction model, has improved detection accuracy and compatibility with the experimental setup compared to the original model. The algorithm's quality depends on the Stereo Matching network for the surrounding environment, optimizing the use of global depth information and local information of transparent objects. The environmental context and lighting conditions also affect the output. While the quality has improved, the speed remains slow. The use of Global Optimization incurs significant computational costs compared to CNN networks, and the processing time is uneven across different runs. Additionally, deploying the model on Google Colab results in slower processing speed and is influenced by the network's quality. Google Colab also has limitations due to modest GPU and CPU quality, especially the CPU.

REFERENCES

[1] I. Lysenkov, V. Eruhimov and G. Bradski, "Recognition and Pose Estimation of Rigid Transparent Objects with a Kinect Sensor," *Conference: Robotics: Science and Systems 2012*, pp. 273-280, 2012. DOI: 10.15607/RSS.2012.VIII.035

[2] C. Guo-Hua, W. Jun-Yi and Z. Ai-Jun, "Transparent object detection and location based on RGB-D camera," *Journal of Physics: Conference Series,* vol. 1183 (1), 2019. DOI: 10.1088/1742-6596/1183/1/012011

[3] T. Weng, A. Pallankize, Y. Tang, O. Kroemer and David Held, "Multi-modal Transfer Learning for Grasping Transparent and Specular Objects," *IEEE Robotics and Automation Letters*, vol. 5 (3), pp. 3791-3798, 2020. DOI: 10.1109/LRA.2020.2974686

[4] S. Sajjan, M. Moore, M. Pan, and S. Shuran, "ClearGrasp: 3D Shape Estimation of Transparent Objects for Manipulation," *2020 IEEE International Conference on Robotics and Automation (ICRA),* pp. 3634-3642, 2020. DOI: 10.1109/ICRA40945.2020.9197518

[5] V. Tankovich, C. Häne, and S. Bouaziz, "HITNet: Hierarchical Iterative Tile Refinement Network for Real-time Stereo Matching," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14357-14367, 2021. DOI: 10.1109/CVPR46437.2021.01413

[6] Z. Shen, Y. Dai, and Liangjun Zhang "PCW-Net: Pyramid Combination and Warping Cost Volume for Stereo Matching," *European Conference on Computer Vision (ECCV 2022)*, pp. 280-297, 2022. DOI: 10.1007/978-3-031-19824-3_17

[7] M. Mousavi and R. Estrada, "SuperCaustics: Real-time, open-source simulation of transparent objects for deep learning applications," *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA),* pp. 649-655, 2021. DOI: 10.1109/ICMLA52953.2021.00108

[8] W. Chien-Yao, A. Bochkovskiy and M. L. Hong-Yuan, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *Computer Vision and Pattern Recognition*, 2022. DOI: 10.48550/arXiv.2207.02696

[9] S. Chen and V. A. Prisacariu, "DFNet: Enhance Absolute Pose Regression with Direct Feature Matching," *European Conference on Computer Vision (ECCV 2022)*, pp. 1-17, 2022. DOI: 10.1007/978-3-031-20080-91

[10] H. Yuanlin and W. Liu (2022), "ClueDepth Grasp: Leveraging positional clues of depth for completing depth of transparent objects," *Front. Neurorobot*, 16:1041702, 2022. DOI: 10.3389/fnbot.2022.1041702

[11] L. Jiankun, and L. Shuaicheng, "Practical Stereo Matching via Cascaded Recurrent Network with Adaptive Correlation," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16242-16251, 2022.

[12] Z. Shen, Y. Dai and Z. Rao, "CFNet: Cascade and Fused Cost Volume for Robust Stereo Matching," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13901-13910, 2021. DOI: 10.1109/CVPR46437.2021.01369

[13] A. Geiger, P. Lenz and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354-3361, 2012. DOI: 10.1109/CVPR.2012.6248074

[14] C. Jia-Ren and C. Yong-Sheng, "Pyramid Stereo Matching Network," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5410-5418, 2018.

[15] X. Guo and L. Hongsheng, "Group-wise Correlation Stereo Network," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3268-3277, 2019. DOI: 10.1109/CVPR.2019.00339

[16] Y. Zhang and T. Funkhouser, "Deep depth completion of a single rgb-d image", *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. *175-185*, 2018. DOI: 10.1109/CVPR.2018.00026

[17] Z. Wu and R. Fan, "Transparent Objects: A Corner Case in Stereo Matching," *2023 IEEE International Conference on Robotics and Automation (ICRA 2023) - London, UK*, pp.12353-12359, 2023. DOI: 10.1109/ICRA48891.2023.10161385

[18] L. H. Hiep, "Study to design of automatic bean sprout growing machine ICTU_ASM_2019," *TNU Journal of Science and Technology*, vol. 204, no. 11, pp. 39-45, 2019.

[19] L. H. Hiep, "Study to improve of automatic control system in Tea black production ferment processing by applying of digital image processing technology," *TNU Journal of Science and Technology*, vol. 225, no. 06, pp. 338-395, 2020.

[20]. L. H. Hiep, "Design a robottics forearm product for bioinformatic laboratoriesh," *TNU Journal of Science and Technology*, vol. 226, no. 11, pp. 226-233, 2021. DOI: 10.34238/tnu-jst.4659

TÓM TẮT

NGHIÊN CỨU XÂY DỰNG MÔ HÌNH NHẬN DIỆN VẬT THỂ TRONG SUỐT DỰA TRÊN THỊ GIÁC LẬP THỂ VÀ TRÍ TUỆ NHÂN TẠO

Điền Thị Hồng Hà
Trường Đại học Kinh tế - Kỹ thuật công nghiệp, Hà Nội, Việt Nam
Ngày nhận bài 08/01/2024, ngày nhận đăng 26/3/2024

Bài báo trình bày kết quả nghiên cứu xây dựng mô hình Stereo Matching để xây dựng ảnh độ sâu thô từ Stereo Camera. Từ đó nhằm tái cấu trúc ảnh độ sâu, khôi phục thông tin độ sâu và tạo ra ảnh độ sâu hoàn chỉnh giúp nhận diện tốt vị trí của hai mẫu ly trong suốt trên thực tế. Bên cạnh đó nghiên cứu đã thực hiện thiết kế giao diện phầm mềm quan sát ảnh độ sâu, đám mây điểm và giao tiếp điều khiển cánh tay Robot gắp vật thể trong không gian ba chiều. Kết quả cho thấy: Chất lượng mô hình tái cấu trúc ảnh độ sâu được cải thiện so với mô hình ClearGrasp khi đánh giá trên bộ dữ liệu ClearGrasp; Có được định hướng về cách cải thiện mô hình, giải thuật tái cấu trúc ảnh độ sâu ở mức định lượng hơn; Tỉ lệ gắp ly thủy tinh thành công trên 90% trong trường hợp vật nằm trên mặt sàn; Tỉ lệ gắp ly thủy tinh thành công trên 70% trong trường hợp vật được đặt ở các độ cao khác nhau; Giao diện hiển thị tốt về chi tiết và giao tiếp, điều khiển được ảnh độ sâu, đám mây điểm, đồ thị vị trí x, y, z; Dễ tương tác và thuận tiện trong quá trình thực nghiệm.

Keyword: Nhận diện vật thể; vật thể trong suốt; thị giác lập thể; ảnh độ sâu; trí tuệ nhân tạo.