Detection of Piled Objects Using RGB-D Faster R-CNN

Antonio P. Doroliat, Chih-Hung G. Li, *IEEE Member*, Wan-Jen Li, and Jenn-Jier J. Lien, *IEEE Member*

***Abstract*— Modern manufacturing turns to automation for efficient and accurate production. Advances in machine vision technology and deep learning algorithms bring automation to another level. Further, the use of a visually guided robot arm enhances the effectiveness of automation implementation. In this paper, we developed an RGB-D Faster R-CNN algorithm with a multi-modal and 1x1 Convolution fusion layer to accurately detect piled objects and generate relevant data for appropriate action by a robot arm. We used an RGB-D instead of an RGB camera to collect the features of the samples. Moreover, the model was evaluated on three different sets of samples, such as screws, blocks, and workpieces, to determine its detection performance. The performance is then compared to that of the other state-of-the-art models. Results show that among the samples considered, the block dataset provides the highest image quality for accurate detection. Moreover, for all datasets, the proposed model outperformed other R-CNN models considered in the study.**

I. Introduction

Mass manufacturing involves efficient execution at every stage of the process. Delay and wastage of material equate to additional manufacturing cost. In certain process stages, materials need to be loaded or inspected one by one. In this process, it is important that each piece is fed one at a time and the entire object is free of any occlusion. If materials are stacked together or piled on top of the other, possible interruption of the process may take place. To eliminate this situation, the industry turns to manual labor to pick and arrange the objects. However, as a human worker has variable efficiency and reliability, the job cannot be perfectly done. Another common technique for solving the problem is by using vibratory feeders [1] [2]. This system effectively prevents objects from piling together but poses potential damage to relatively fragile objects during vibration. To address the aforementioned problem, this study introduces a system that employs a deep learning algorithm for object detection and then consequently controls a robot arm in implementing the necessary intervention.

The use of machine vision for object detection of piled objects is not entirely new. Generally, object recognition and estimation of its position and orientation may be determined by obtaining data range images of the object [3]. In a prior study, traditional machine vision with RGB single input was used [4]; the algorithm uses template matching and Scale Invariant Feature Transform (SIFT) for detection and feature definition. However, it failed to distinguish handcrafted features from real ones and demonstrated poor performance for severe occlusion and the presence of reflection. The RGB-D multi-input method utilizes a slicing depth image and contour matching; it improved performance for severe occlusion and reflection but still failed for the handcrafted feature. Another study considered a deep learning algorithm Faster R-CNN with RGB as input [5]. It demonstrated fast detection but still performed poorly for objects of severe occlusion and reflected images. Gupta et al. [6] managed to solve the occlusion issue by adding a depth input in their RGB-D Faster R-CNN algorithm; expectedly, the performance was affected by the depth image quality. Additionally, both RGB and RGB-D Faster R-CNN demand higher hardware graphics capability. Though of varied performance, these algorithms were successfully implemented for identifying piled objects. Motivated to achieve a more accurate and efficient performance, we developed a modified detection architecture. It employs an RGB-D Faster R-CNN to identify piled objects to provide information to a robot arm which can be stationary or be mobilized as part of an autonomous mobile robot (AMR), as illustrated in Figure 1.

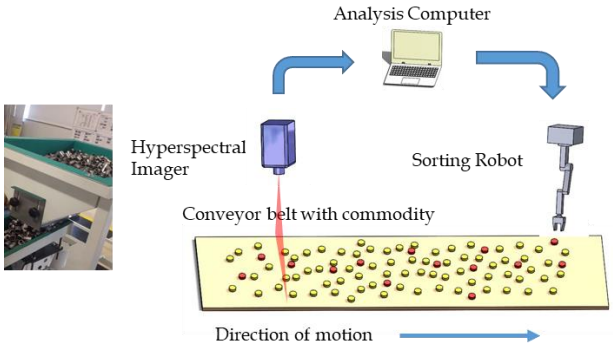

Figure 1: Autonomous system with machine vision and robot arm

A. Related Works

Computer vision methods for object detection have been developed, e.g., SIFT adopting handcrafted local features for

This work was supported by the Ministry of Science and Technology of the Republic of China, Taiwan under contract no. MOST109-2637-E-027-008

A.P. Doroliat, is with the International Graduate Institute of Mechanical and Electrical Engineering, National Taipei University of Technology, Taipei City, 106 Taiwan (e-mail: t107a89403@ntut.org.tw)

C.G. Li, is with the Graduate Institute of Manufacturing Technology, National Taipei University of Technology, Taipei City, 106 Taiwan (e-mail: cL4e@ntut.edu.tw).

W. Li is with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, 70101 Taiwan. (e-mail: j6079633@gmail.com).

*J.J. Lien is with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, 70101 Taiwan, (e-mail: jjlien@csie.ncku.edu.tw).

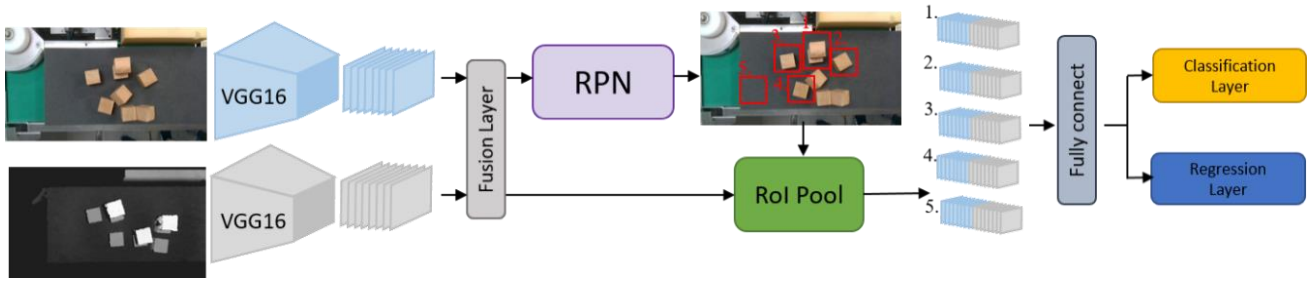

Figure 2: Proposed architecture which involves modified RGB-D Faster R-CNN

template matching [4]. It can be easily implemented without a large amount of training data; however, it failed to deal with severe occlusion and reflection problems. By incorporating RGB-D cameras, the depth information can be obtained. One may slice the depth image and attempt to match the target with the contours in the sliced image and the HSV feature in the color image; an occlusion-robust method can be acquired.

Recently, the deep learning-based detectors exhibited superior performance; they are classified into two types, the one-stage detectors and the two-stage detectors such as the region CNN (R-CNN) series [7]. The main difference between the one-stage and the two-stage detectors is whether there are defining proposals or not. The R-CNN [8] series begins with Selective Search. It uses Selective Search to generate proposals; then computes CNN features of each proposal and classifies them with SVM [9]. Fast R-CNN [10] reduces the computation time by using shared global CNN features and using CNN to replace SVM. Faster R-CNN [5], the following version, developed Region Proposal Network to substitute Selective Search; the whole process becomes an end-to-end network. Another study introduces a multi-modal CNN architecture which implements a medium feature level [11]. The multi-modal characteristic of the architecture proves to be advantageous for object recognition.

B. Contributions

The current study considered the work of Schwarz et al. as the reference architecture for RGB-D object detection and semantic segmentation for autonomous manipulation in clutter [12]. They used a modified Faster R-CNN to detect the bounding box of objects called DenseCap [13]. Then, to segment the probability map of objects, a fully convolutional neural network called OverFeat [14] [15] was utilized. Additionally, the bounding box was used to find coarse information of the object combined with semantic segmentation to achieve precision. For the depth data, the HHA encoding [6] method was used to enhance depth information. Furthermore, the pre-trained model of HHA data was trained with Cross-Modal Distillation [16]. However, encoding and training of the pre-trained model takes great effort and is time-consuming. Thus, the current study presents a modified method of topology used in the reference study. It specifically presents the detection of piled objects using RGB-D Faster R-CNN for faster and more accurate detection.

There are three salient features emphasized in this study. Firstly, an RGB-D-based Faster R-CNN is introduced, which uses RGB images, and raw depth images as input. In this model, encoding the raw depth image into HHA (Height above ground, Horizontal disparity, Angle to the gravity) format is not necessary. Secondly, adding a fusion layer improves the performance of the multi-modal model. The traditional models concatenate RGB and depth features [17], [18], to do the multi-modal task. However, the feature increases that it becomes difficult to converge and the model does not learn its correlation. In the presented model, we use a 1×1 convolution layer after concatenated RGB and depth features in the RGB-D fusion layer [19] to find the linear-nonlinear correlation of multi-modal features. Lastly, we develop a detection and classification model for an object pile grasping system.

II. The RGB-D Faster R-CNN for Piled Object Detection

Figure 2 presents the modified network architecture of RGB-D Faster R-CNN proposed in this study. It involves inputs of two VGG16 convolutional neural networks for RGB and depth feature extraction. The fusion layer is applied to fuse the extracted features in the two different models. Then, Region Proposal Network (RPN) learns how to generate a region proposal from the fused features. From the generated proposals, a region of interest (RoI) is identified and is used to crop and resize the local feature map on the global feature map. Finally, the fully connected layer encodes the local feature map and lets the classification layer and regression layer classify the category and refine the offset of proposals respectively.

III. Experimental Results

The deep learning model introduced in this study demonstrates promising results in object detection at relatively high precision. Experimental evidence shows that it outperformed other systems for all datasets and levels of difficulty. Although there is a corresponding trade-off in computational time, the proposed system provides accurate object detection using RGB and depth image as input.

The experiments involved an examination of the performance in three levels of difficulty such as easy, moderate, and hard. Easy means all objects are not occluded, moderate means some objects are occluded but not over 50%, and hard means some objects are covered over 50%. The obtained datasets for “screw”, “block”, and “workpiece” are compared with each other and to the NYUv2 dataset [20]. The

TABLE I
THE mAP OF DETECTION

		Easy			Moderate			Hard			Total		
	r	0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9
RGB	1) NYUv2										0.712	0.442	0.036
	2) Screw	0.909	0.613	0.030	0.815	0.570	0.015	0.713	0.389	0.012	0.811	0.474	0.012
	3) Block	1.0	1.0	0.169	1.0	1.0	0.152				1.0	1.0	0.202
	4) Workpiece	0.909	0.764	0.018	0.909	0.777	0.106	0.904	0.766	0.045	0.908	0.768	0.021
RGB-D without fusion	1) NYUv2										0.139	0.047	0.004
	2) Screw	0.898	0.541	0.045	0.801	0.459	0.003	0.694	0.253	0.002	0.796	0.421	0.010
	3) Block	1.0	1.0	0.109	0.909	0.879	0.011				0.909	0.894	0.105
	4) Workpiece	0.998	0.873	0.025	0.909	0.651	0.009	0.893	0.651	0.023	0.906	0.752	0.030
RGB-D	1) NYUv2										0.810	0.552	0.042
	2) Screw	0.909	0.655	0.007	0.816	0.559	0.015	0.715	0.504	0.091	0.812	0.565	0.045
	3) Block	1.0	1.0	0.276	1.0	1.0	0.226				1.0	1.0	0.267
	4) Workpiece	1.0	0.771	0.030	0.909	0.781	0.091	0.980	0.733	0.011	0.907	0.753	0.091

test results specifically indicate that the proposed model delivers better performance than a single modal model on the NYUv2 dataset, the screw dataset, and the block dataset (see Table I). Interestingly, the single input model outperformed the RGB-D w/o 1×1 Conv. Fusion. In general, the proposed model of RGB-D w/ 1×1 Conv. Fusion demonstrates superior performance over the other systems for all datasets and levels of difficulty.

The proposed model demonstrates specific levels of improvement on detection precision. Results show that the depth with fusion increased the precision. Nevertheless, poor depth map quality and discontinuity affect the sufficiency of depth information. The inferior performance on the workpiece can be attributed to the poor depth map quality, which results in insufficient depth information. Its depth map is broken at multiple locations due to reflection, while that of the blocks is smooth and continuous. Furthermore, the depth value changes significantly at the boundary. The blocks produces identifiable edges in contrast to that of the screws, which are small and fuzzy. These justify the higher precision obtained for the block datasets.

IV. Conclusion and Future Works

This study successfully implemented a deep learning algorithm using RGB-D Faster R-CNN for object detection. We introduced a modified Faster R-CNN model that enables the system to process multiple inputs. Specifically, a multi-modal fusion layer is employed and a 1×1 convolutional fusion layer is added to extract useful representation in the multi-modal features. This in turn increases the detection performance as indicated by our experimental results. Using screws, blocks, and workpiece as sample objects for performance testing, the topmost object IoU ratios (denoted as r in Table I) were obtained and improvement by the RGB-D Faster R-CNN over RGB Faster R-CNN was witnessed. It was found that the system works steadily well for piled blocks that produce high depth image quality. In contrast, the discontinuity in the depth images of the workpiece dataset and the blurry edges of the screw dataset may be attributed to the inferior performance. On the other hand, it is believed that the performance can be improved by modifying the hyper-parameters of the Faster R-CNN and increasing the number of datasets.

References

[1] S. Okabe and Y. Yokoyama, "Study on vibratory feeders: Calculation of natural frequency of bowl-type vibratory feeders," *ASME. J. Mech. Des.* January 1981; 103(1): 249-256. https://doi.org/10.1115/1.3254873.

[2] S. Okabe, Y. Kamiya, K. Tsujikado, and Y. Yokoyama, "Vibratory feeding by nonsinusoidal vibration—Optimum wave form," *ASME. J. Vib., Acoust., Stress, and Reliab.* April 1985; 107(2): 188–195. https://doi.org/10.1115/1.3269243

[3] G. Kordelas, A. Mademlis, P. Daras, and M.G. Strintzis, "Object recognition and pose estimation from 2.5D scenes, " *In: Furht B. (eds) Encyclopedia of Multimedia. Springer*, Boston, MA., 2008. https://doi.org/10.1007/978-0-387-78414-4.

[4] D. G. Lowe, "Object recognition from local scale-invariant features," *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999, pp. 1150-1157 vol.2, doi: 10.1109/ICCV.1999.790410.

[5] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 1 June 2017, doi: 10.1109/TPAMI.2016.2577031.

[6] S. Gupta, R. Girshick, Arbelaez P and Malik J, "Learning rich features from RGB-D images for object detection and segmentation," in *European Conference on Computer Vision (ECCV)*, 2014.

[7] K.S. Chahal and K. Dey, "A survey of modern object detection literature using deep learning," *arXiv*, 2018, arXiv:1808.07256v1

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014 pp. 580-587.

[9] J.R. Uijlings, K.E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," Selective Search for Object Recognition. *Int J Comput Vis* **104,** 154–171 (2013). https://doi.org/10.1007/s11263-013-0620-5.

[10] R. Girshick, "Fast R-CNN," in *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[11] L. Schneider et al., "Multimodal Neural Networks: RGB-D for Semantic Segmentation and Object Detection," *In: Sharma P., Bianchi F. (eds) Image Analysis. SCIA 2017. Lecture Notes in Computer Science*, vol 10269. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59126-1_9.

[12] M. Schwarz, A. Milan, A. S. Periyasamy, and S. Behnke, "RGB-D object detection and semantic segmentation for autonomous manipulation in clutter," *The International Journal of Robotics Research*, pp. 437-451, 2018. https://doi.org/10.1177/0278364917713117.

[13] J. Johnson, A. Karpathy and L. Fei-Fei, "DenseCap: Fully Convolutional Localization Networks for Dense Captioning," in *IEEE Conference on Computer Vision and Pattern Recognition (VCPR)*, 2016, pp. 4565-4574.

[14] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus and LeCun, "OverFeat: Integrated recognition, localization and," arXiv, 2016.

[15] F. Husain, H. Schulz, B. Dellen, C. Torras and S. Behnke, "Combining semantic and geometric features for object class segmentation of indoor scenes," in *IEEE Robotics and Automation Letters*, vol. 2, no. 1, pp. 49-55, Jan. 2017, doi: 10.1109/LRA.2016.2532927.

[16] S. Gupta, J. Hoffman and J. Malik, "Cross modal distillation for supervision transfer," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2827-2836.

[17] Z. Deng and L. J. Latecki, "Amodal detection of 3D objects: Inferring 3D bounding boxes from 2D ones in RGB-Depth images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5762-5770.

[18] M. M. Rahman, Y. Tan, J. Xue, L. Shao, and K. Lu, "3D object detection: Learning 3D bounding boxes from scaled down 2D bounding boxes in RGB-D images," *Information Sciences*, pp. 147-158, 2019.

[19] T. Ophoff, K. Van Beeck, and T. Goedem, "Exploring RGB+Depth fusion for real-time object detection," *Sensors* 2019, *19*(4), 866; https://doi.org/10.3390/s19040866.

[20] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," in *European Conference on Computer Vision (ECCV)*, 2012, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33715-4_54.