Proceedings of the 9th International Conference of Control Systems, and Robotics (CDSR'22)
Niagara Falls, Canada – June 02-04, 2022
Paper No. 112
DOI: 10.11159/cdsr22.112

A Vision-based Object Detection and Localization System in 3D Environment for Assistive Robots' Manipulation

Md Ishrak Islam Zarif[1], Md Tanzil Shahria[2], Md Samiul Haque Sunny[2], Md Mahafuzur Rahaman Khan[3], Sheikh Iqbal Ahamed[1], Inga Wang[4], Mohammad H Rahman[3]
[1]Department of Computer Science, Marquette University, Wisconsin, United States
[2]Department of Computer Science, University of Wisconsin-Milwaukee, Wisconsin, United States
[3]Department of Mechanical Engineering, University of Wisconsin-Milwaukee, Wisconsin, United States
[4]Department of Rehabilitation Sciences & Technology, University of Wisconsin-Milwaukee, Milwaukee, United States
ishrak.zarif@marquette.edu, mshahria@uwm.edu, msunny@uwm.edu, khan45@uwm.edu, sheikh.ahamed@marquette.edu, wang52@uwm.edu, rahmanmh@uwm.edu

Abstract **–** Robots are used today in many different fields and for many different tasks. However, they quickly reach their limits without a "sense of sight," especially when working directly with humans as assistive robots. Applications that require sophisticated intelligence must work under flexible conditions that are usually not feasible without vision systems. This study proposes a vision-based system for object detection and localization that can potentially be used for assistive robots. The goal is to facilitate Activities of Daily Living (ADL) tasks in an unstructured environment for individuals who use wheelchairs. For vision-based manipulation in an unstructured environment, interest object features and homography analysis are used to get the necessary information for controlling the robotic arm. SSD MobileNet V2, a pre-trained inference model using TensorRT optimized network along with a RealSense camera is used in this study to detect, recognize, and localize objects in the 3D environment. Experiments and results have shown that the system can be utilized to perform object grasping tasks robustly.

Keywords**:** Computer Vision, Object Recognition, 3D Localization, Assistive Robot, SSD MobileNet, RealSense, Jetson Nano

1. Introduction

Robotics technology has already encountered all aspects of work as it has the potential to support lives and work practices, boost productivity and safety levels, and offer enhanced levels of service [1, 2]. The application of robotics nowadays is not limited to the research lab or automated industry. These technologies are now successfully practiced in food sorting application, medical training, police training, extracting poison, exploring the sewer, etc. [3].

While performing different services, a robot needs to interact with various objects as well as manipulate them in the environment around them. A prerequisite for large-scale robot applications is to detect and recognize targets, their location, and the generation of appropriate grasp strategies so they can grasp objects as precisely and quickly as humans. In industrial settings, grasping is mainly utilized to solve pick and place issues, whereas in the assistive application, the main goal is to facilitate the user with the help in Activities of Daily Living (ADL) tasks in an unstructured environment for individuals.

To manipulate any object, a robot needs to read data from its surroundings to make a decision about its next action. Among different sensory inputs, vision input plays a vital for robots to operate in a complex and unstructured environment [4, 5]. Robots can operate without any vision in industrial applications as the desired task and object to handle is predetermined. But for a more complex application like assistive robots, vision-based input is a must as the Robot interacts with the human subjects directly, and the tasks/environments are often unknown. Thus vision-based approaches are more vital for assistive robots.

In this research, we propose a vision-based guidance system that includes object detection, recognition, and 3D localization of any object in the surrounding environment in real-time. The developed vision-based guidance system can be

used for assistive robots for object manipulation in 3D space. The system can recognize any object from the dataset and generate the 3D coordinates for further analysis and tasks execution, such as manipulation of any object.

The rest of this paper is organized as follows: Section 2 presents a few related works in this field, Section 3 discusses the approach applied in this study, Section 4 demonstrates the experimental setup of the overall research, Section 5 illustrates the result of this study, and finally, Section 6 draws the conclusion of the study.

2. Literature Review

Research on geometric primitive object detection and robust recognition systems is getting popular among researchers. Global and local feature-based techniques are the literature's primary 3D object identification techniques [6, 7]. The global features approach calculates a collection of global features that effectively describe the complete 3D object [8] whereas local approaches, calculate features around specific vital points in the neighborhood. Compared to global approaches, these strategies are superior at handling occlusion and chaos [7].

Vision-based Measurement (VBM) has recently been utilized to study a variety of phenomena, including robot navigation [9, 10], and facial recognition [11], among others. Stereoscopic vision, in particular, is a dependable technique for extracting data from the environment [12, 13]. The precision of the findings is determined on the stereo camera system chosen, and the stereo correspondence method applied. Biological observation suggests that two slightly shifted views of the same landscape give sufficient information to interpret object depth. Two cameras on pan-tilt units [14] or one omnidirectional camera system with epipolar geometry [15] are used in stereo vision approaches. Based on the standard 4-dimensional point pair feature, Choi and Christensen [15] created a color point pair feature (CPPF) to individually record the geometric surface form and photometric color features. Moreover, the camera might view featureless scenes (e.g., a white wall). Real-time appearance-based mapping (RTAB-MAP) [16] can leverage external odometry as motion estimates for more robust odometry.

In industrial settings, grasping is mainly utilized to solve pick and place issues. They demonstrated a robot control technique and a computer vision algorithm. A recognition method that relies only on 3D geometry information and an efficient localized random sample consensus (RANSAC)-like sampling strategy in household environments and a combination of perception and manipulation components such as grasp planning and motion planning. In all situations, simply a depth map was used to extract geometry information from the images. According to Verma et al. [17], the algorithm of density clustering and homograph transformation can acquire the object's maximum stable extremal approach and then realize the object's exact placement, which gives substantial help for the manipulator's successful grabbing. Because of its solid mathematical foundation and flexibility to solve multiclassification challenges, the Bayesian method is frequently employed in noise reduction, servo control, and grasping probability prediction in target detection and identification and Robot grasping research.

Another study [18] employed 3D mapping to combine a structured light sensor and a Time of Flight (ToF) camera for item localization in industrial settings. Based on RGB and point cloud pictures, the SVM-rank algorithm was utilized to detect object features [19], produce the grasping strategy, and finally accomplish accurate item grabbing with a five-fingered dexterous hand. Budiharto [20] devised a rapid stereo vision-based object recognition technique that employed the Bayesian approach to decrease camera noise and achieve robust tracking. Zhang et al. [21] suggested a shared control wheelchair manipulator that can identify a water cup automatically based on vision and assist the impaired in completing the job of drinking water. GoogLeNet has achieved excellent accuracy in robot target detection due to its good performance in image recognition [22].

3. Methods

To detect, recognize, and localize objects in the 3D environment, the proposed system uses a pre-trained inference model using TensorRT Optimized Network [23], OpenCV, and libraries & packages provided by Jetson. Using this

method, the proposed system can detect and recognize any objects from the COCO dataset [24] within the image frame and calculate the x and y coordinates of the center point of that object. Using these coordinates and the depth channel of the RealSense camera [25], the system calculates the depth (third coordinate) of different objects which are used for 3D localization of objects.

3.1 System Architecture

At first, the system loads the pre-trained SSD MobileNet V2 inference network using TensorRT Optimized Network and initializes the display and RealSense depth camera. NVIDIA TensorRT is a deep learning inference SDK with exceptional performance [26]. It provides a deep learning inference optimizer and runtime for deep learning inference applications with low latency and high throughput. TensorRT is based on CUDA which is NVIDIA's parallel programming architecture, and it allows users to improve inference using CUDA-X libraries, development tools, and technologies for artificial intelligence, vision-based work, autonomous machines, and high-performance computing. During inference, TensorRT-based apps are up to 40 times faster than CPU-only systems. Fig. 1 presents the basic structure of the applied TensorRT model. Here, CBR is a single kernel that combines the convolution, bias, and ReLU layers of varied sizes.

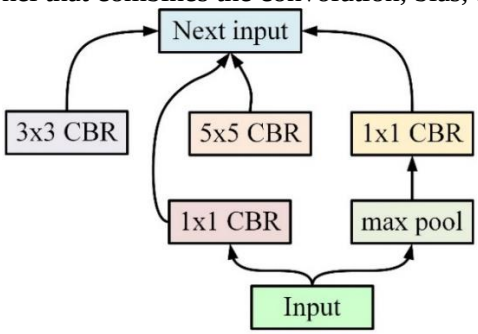

Fig. 1: Structure of the applied TensorRT model.

The SSD MobileNet architecture is a single convolutional network that learns to predict and classify bounding box locations in a single pass [27]. A baseline architecture is followed by numerous convolution layers in the SSD MobileNet network. Only one shot is needed in SSD MobileNet to recognize several items within an image, but techniques based on regional proposal networks (RPNs), such as the R-CNN series, require two shots, one for generating region proposals and the other for detecting the object of each proposal. As a result, SSD MobileNet is faster and less expensive in terms of computing than other networks. It does not need much computation power to run the SSD MobileNet. As the whole system is mounted on a wheelchair and Jetson Nano is used for computation with limited computation power, SSD MobileNet architecture is best suited in this regard. Fig. 2 represents the basic structure of the applied SSD MobileNet architecture.

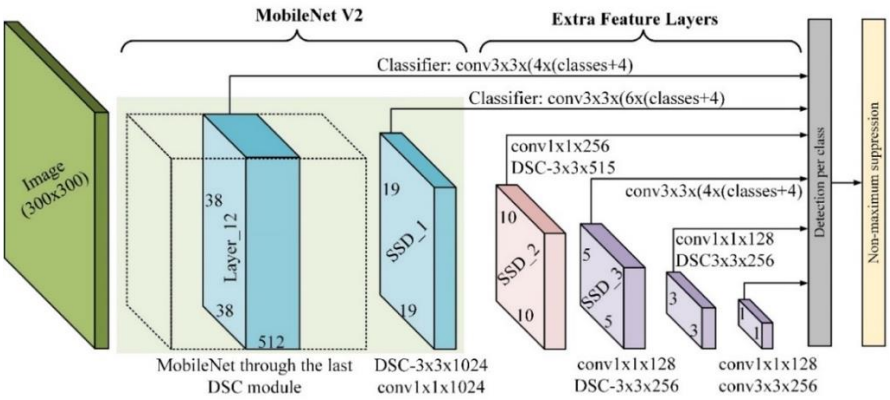

Fig. 2: System architecture of Single Shot Detector (SSD) model.

The system then checks if the depth camera is attached or not and receives the RGB and depth frames from the camera as input. After that, the model converts the RGB frame into a CUDA image, as the detection method in the inference of Jetson takes only CUDA image as input. Then the system detects and recognizes different objects along with the confidence value from the input frame using the detection model. The adopted MobileNet model can successfully detect objects from the COCO dataset in real-time. The system then computes the 2D coordinates of the center of different objects in terms of pixel values. The model then uses the depth frame of the camera and the 2D coordinates of each object to estimate the distance of that object from the camera. After all computations, the system previews the RGB frame along with the estimation value of all the objects and coordinates of the center point of objects. The system then passes the 2D coordinates and distance value of each object to the following system as a 3D coordinate. The next step uses the coordinate values and applies inverse kinematics along with the PID controller to estimate trajectory. Fig. 3 presents the overall flowchart of the proposed system.

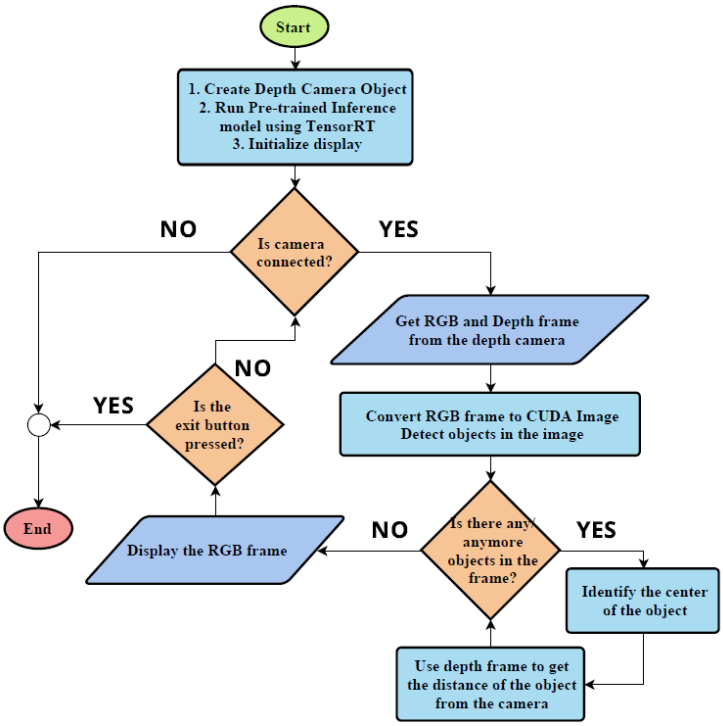

Fig. 3: Workflow of the proposed model.

There are five libraries in total that we used in this system: OpenCV, pyrealsense2, NumPy, jetson.inference, and jetson.utils. The pseudo-code of the presented system is given below:

```
SET detectNet USING SSD-MobileNet_V2
SET display
SET config
SET pipeline_wrapper
SET pipeline_profile
SET DepthCamera

DEFINE FUNCTION DepthCamera:
   SET frames FROM pipeline
   SET color_frame, depth_frame TO pipeline_frames

   IF NOT depth_frame OR NOT color_frame:
      RETURN False
   ELSE
      RETURN distance_value, depth_image, color_image

WHILE display Is Open:
   GET RGB_frame, depth_frame FROM DepthCamera
   CONVERT RGB_frame TO CUDA_image
   GET objects and confidence_value FROM CUDA_image USING detectNet

   FOR each object IN objects:
      GET x and y coordinate of the object
```

```
    GET distance USING depth_frame and coordinates
    SET (confidence_value, x, y, distance) as OUTPUT

DISPLAY RGB_frame until EXIT
```

4. Experimental Setup

For our experiment, we are using the NVIDIA Jetson Nano Developer Kit 4GB version [28] for detection as well as recognition and the Intel RealSense Depth Camera D435 [29] to capture the RGB and depth frames from the image. The depth camera is attached using a camera stand with the 6 degrees of freedom (DoF) robotic arm alongside the gripper from UFACTORY [30]. The xArm is mounted on a power wheelchair from Permobil [31]. A control box is attached at the back of the wheelchair for controlling both the wheelchair and xArm. The control box, xArm, and Jetson Nano all use the power from the power wheelchair battery. We also attached an emergency switch at the side of the wheelchair to avoid unwanted circumstances.

The NVIDIA Jetson Nano is a development kit and embedded system-on-module (SoM) from the NVIDIA Jetson family [28]. It features a compact form factor and a lot of processing power, so it's perfect for computer vision and deep learning applications. Jetson Nano has a 128-core Maxwell GPU, quad-core ARM A57 64-bit CPU, 4GB LPDDR4 memory, MIPI CSI-2 and PCIe Gen2 high-speed I/O, and MIPI CSI-2 and PCIe Gen2 high-speed I/O. The NVIDIA JetPack SDK is used to run Linux on the Jetson Nano, which provides 472 GFLOPS of FP16 computation performance while requiring 5 to 20W of power.

The Intel RealSense Depth Camera D435 has four lenses, an RGB module for conventional photo and video, an IR projector, as well as a right and left imager for depth sensing [29]. With the stereo depth approach, this USB-powered camera produces depth data. With its global image shutter and wide field of view, the Intel RealSense Depth Camera D435 provides precise depth perception even when the object or device is moving. In the camera module, two depth sensors are separated by a little distance. The two pictures obtained by these two sensors are compared using a stereo camera. Because the distance between the sensors is known, these comparisons can yield depth data.

The RealSense camera records video of whatever is in front of it and feeds it to the Jetson Nano. The Jetson Nano then began analyzing the picture, splitting frames from the video. It determines coordinates, distance value, information, as well as placements on the frame, following the recognition and detection of objects in 3D space. The findings are then sent to the display that is attached.

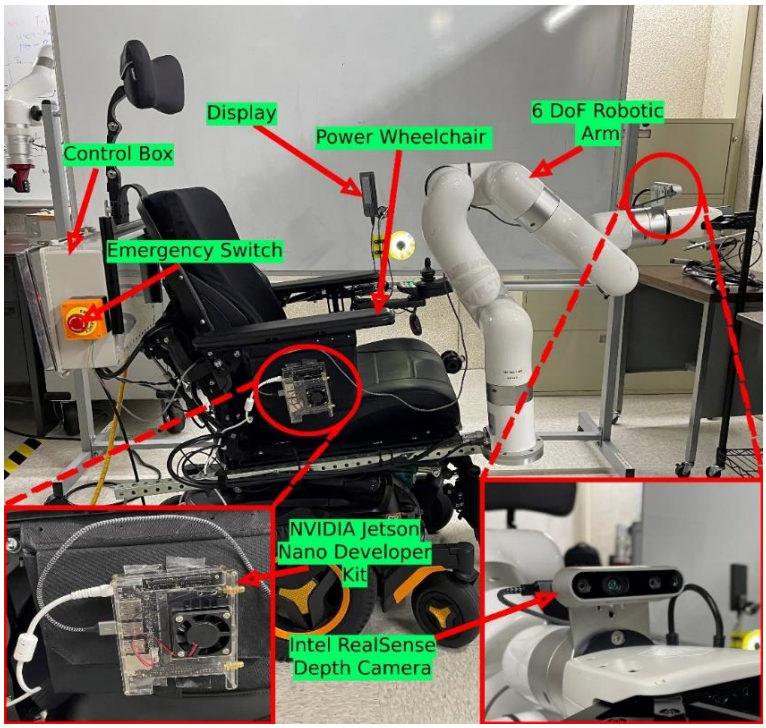

Fig. 4: Experimental setup for the proposed system.

5. Results and Discussion

In this experiment, NVIDIA Jetson Nano is used as the development tool and Intel RealSense Depth Camera as the input device. The proposed system can successfully detect and recognize objects in the input image frame from the COCO dataset with an accuracy of 79.84%. Among the recognized objects, the model can precisely estimate the 3D coordinates of the center of detected objects with an accuracy of 92%. This accuracy is measured by validating the position of the center of each object with respect to the camera. Fig. 5 displays the recognizing and coordinate estimation of different objects in real-time: (a) cup and bottle, (b) book and apple, (c) laptop, keyboard, and mouse, (d) backpack and microwave, (e) person and laptop, (f) scissors and cell phone.

Overall, the proposed system works really well with the surrounding environment, yet there are some issues with the experimental setup. The depth value is not steady and fluctuates sometimes. In the future, we will explore other commercially available depth camera to enhance the performance of the proposed system. We will also explore other possible strategies to design a more robust model for our application.

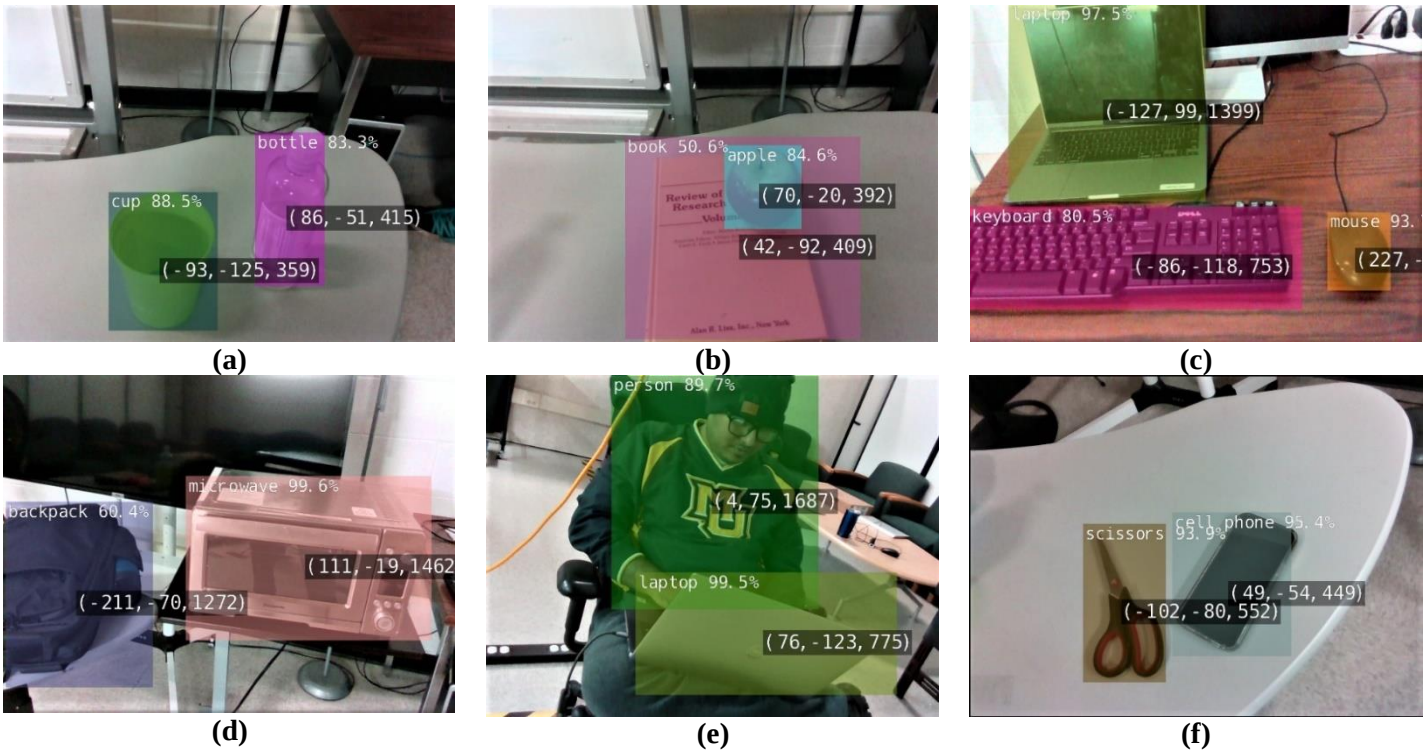

Fig. 5: The system recognizing different objects ((a) cup and bottle, (b) book and apple, (c) laptop, keyboard, and mouse, (d) backpack and microwave, (e) person and laptop, (f) scissors and cell phone) in real time

6. Conclusion

This study describes the development of vision-based object detection, recognition, and localization system in a 3D environment. The developed system can be used for Assistive Robot to manipulate objects in performing activities of the daily living task. SSD MobileNet V2, along with OpenCV, is used to extract information from the objects in a 3D environment. Results have shown that we can get the coordinates of an object with depth information in an unknown environment which can be mapped with end effector for manipulation. Cameras as the primary sensing method are much more cost-effective than the other image sensors, such as lidar used in many vision-based robot manipulations that use SLAM algorithms. Much remains to be done to create the complete system that will perform vision-based navigation in the way we are envisioning. In the future, we will investigate other possible input sensors to get more stable and accurate performance.

Acknowledgment

This material is based upon work supported by NASA under Award No. RIP23_1-0 issued through Wisconsin Space Grant Consortium. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Aeronautics and Space Administration.

References

[1] “Why is robotics important?,” *ONE Only Natural Energy*, 19-Jun-2017. [Online]. Available: https://www.onlynaturalenergy.com/why-is-robotics-important/. [Accessed: 20-Nov-2021].

[2] Sunny, M. S. H., Zarif, M. I. I., Rulik, I., Sanjuan, J., Rahman, M. H., Ahamed, S. I., ... & Brahmi, B. (2021). Eye-gaze control of a wheelchair mounted 6DOF assistive robot for activities of daily living. *Journal of NeuroEngineering and Rehabilitation*, *18*(1), 1-12.

[3] J. Roberts, "Six ways robots are used today that you probably didn't know about," *The Conversation*, 28-Sep-2021. [Online]. Available: https://theconversation.com/six-ways-robots-are-used-today-that-you-probably-didnt-know-about-82067. [Accessed: 20-Nov-2021].

[4] H. Koichi. "A review on vision-based control of robot manipulators." *Advanced Robotics* 17, no. 10 (2003): 969-991. [Accessed: 20-Nov-2021].

[5] Y. Ren, H. Sun, Y. Tang and S. Wang, "Vision Based Object Grasping of Robotic Manipulator," 2018 24th International Conference on Automation and Computing (ICAC), 2018, pp. 1-5, doi: 10.23919/IConAC.2018.8749001.

[6] E. Akagund ¨ uz and I. Ulusoy, "3D object recognition from range images ¨ using transform invariant object representation," Electronics Letters, vol. 46, no. 22, pp. 1499–1500, 2010.

[7] N. Bayramoglu and A. A. Alatan, "Shape index SIFT: Range image recognition using local features," in Proceedings of International Conference on Pattern Recognition, 2010, pp. 352–355.

[8] U. Castellani, M. Cristani, S. Fantoni, and V. Murino, "Sparse points matching by combining 3D mesh saliency with statistical descriptors," Computer Graphics Forum, vol. 27, no. 2, pp. 643–652, 2008.

[9] H. Koch, A. Konig, A. Weigl-Seitz, K. Kleinmann, and J. Suchy, "Multisensor contour following with vision, force, and acceleration sensors for an industrial robot," IEEE Trans. Instrum. Meas., vol. 62, no. 2, pp. 268–280, Feb. 2013.

[10] K. Das Sharma, A. Chatterjee, and A. Rakshit, "A PSO–Lyapunov hybrid stable adaptive fuzzy tracking control approach for vision-based robot navigation," IEEE Trans. Instrum. Meas., vol. 61, no. 7, pp. 1908–1914, Jul. 2012

[11] G. Betta, D. Capriglione, M. Corvino, C. Liguori, and A. Paolillo, "Face based recognition algorithms: A first step toward a metrological characterization," IEEE Trans. Instrum. Meas., vol. 62, no. 5, pp. 1008–1016, May 2013.

[12] F. Kazemzadeh, S. A. Haider, C. Scharfenberger, A. Wong, and D. A. Clausi, "Multispectral stereoscopic imaging device: Simultaneous multiview imaging from the visible to the near-infrared," IEEE Trans. Instrum. Meas., vol. 63, no. 7, pp. 1871–1873, Jul. 2014.

[13] Pollefeys, M., Nistér, D., Frahm, J. M., Akbarzadeh, A., Mordohai, P., Clipp, B., ... & Towles, H. (2008). Detailed real-time urban 3d reconstruction from video. International Journal of Computer Vision, 78(2), 143-167.

[14] S. K. Gehrig, J. Klappstein, U. Franke, and A. G. DaimlerChrysler, "Active stereo for intersection assistance," in Proc. VMV, 2004, pp. 29–35. [16] Z. Zhu, "Omnidirectional stereo vision," in Proc. Workshop Omnidirectional Vis., Budapest, Hungary, 2001, pp. 1–12

[15] C. Choi and H. I. Christensen, ''RGB-D object pose estimation in unstructured environments,'' Robot. Auto. Syst., vol. 75, pp. 595–613, Jan. 2016.

[16] Labbé, M., and F. Michaud. 2019. "RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation." J. Field Rob. 36 (2): 416–446. https://doi .org/10.1002/rob.21831.

[17] N. K. Verma, A. Mustafa, and A. Salour, ''Stereo-vision based object grasping using robotic manipulator,'' in Proc. 11th Int. Conf. Ind. Inf. Syst. (ICIIS). Roorkee, India: IEEE, Dec. 2016, pp. 95–100

[18] Pfitzner C, Antal W, Hess P, May S (2014) 3D multi-sensor data fusion for object localization in industrial applications. In Proceedings of ISR/Robotik, 41st International Symposium on Robotics, 1-6, ISBN: 978-3-8007-3601-0

[19] H. Yuan, D. Li, and J. Wu, ''Efficient learning of grasp selection for five-finger dexterous hand,'' in Proc. IEEE 7th Annu. Int. Conf. CYBER Technol. Autom., Control, Intell. Syst. (CYBER). Honolulu, HI, USA: IEEE, Jul. 2017, pp. 1101–1106.

[20] W. Budiharto, ‘‘Robust vision-based detection and grasping object for manipulator using SIFT keypoint detector,’’ in Proc. Int. Conf. Adv. Mech. Syst. Kumamoto, Japan: IEEE, Aug. 2014, pp. 448–452.
[21] Z. Zhang, S. Mao, K. Chen, L. Xiao, B. Liao, C. Li, and P. Zhang, ‘‘CNN and PCA based visual system of a wheelchair manipulator robot for automatic drinking,’’ in Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO). Kuala Lumpur, Malaysia: IEEE, Dec. 2018, pp. 1280–1286.
[22] W. Dongyu, H. Fuwen, T. Mikolajczyk, and H. Yunhua, ‘‘Object detection for soft robotic manipulation based on RGB-D sensors,’’ in Proc. WRC Symp. Adv. Robot. Autom. (WRC SARA). Beijing, China: IEEE, Aug. 2018, pp. 52–58.
[23] “A review on vision-based control of robot manipulators,” *Neuralet*. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1163/156855303322554382?journalCode=tadr20. [Accessed: 20-Nov-2021].
[24] “Common objects in context,” *COCO*. [Online]. Available: https://cocodataset.org/. [Accessed: 20-Nov-2021].
[25] “Jetson/libraries,” *Jetson/Libraries - eLinux.org*. [Online]. Available: https://elinux.org/Jetson/Libraries. [Accessed: 20-Nov-2021].
[26] “Nvidia TENSORRT,” *NVIDIA Developer*, 23-Nov-2021. [Online]. Available: https://developer.nvidia.com/tensorrt. [Accessed: 24-Nov-2021].
[27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg, "Ssd: Single shot multibox detector." In *European conference on computer vision*, pp. 21-37. Springer, Cham, 2016.
[28] “Jetson Nano Developer Kit,” *NVIDIA Developer*, 14-Apr-2021. [Online]. Available: https://developer.nvidia.com/embedded/jetson-nano-developer-kit. [Accessed: 20-Nov-2021].
[29] “Depth camera D435,” *Intel® RealSense™ Depth and Tracking Cameras*, 17-Jun-2021. [Online]. Available: https://www.intelrealsense.com/depth-camera-d435/. [Accessed: 20-Nov-2021].
[30] “UFACTORY xArm 6,” *UFACTORY*. [Online]. Available: https://www.ufactory.cc/products/xarm-6-2020. [Accessed: 20-Nov-2021].
[31] “Permobil M3 Corpus,” *Permobil*. [Online]. Available: https://www.permobil.com/en-us/products/power-wheelchairs/permobil-m3-corpus. [Accessed: 20-Nov-2021].