IJARETY

International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 12, Issue 2, March-April 2025

Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202067

Deep Learning-Based Stereo Vision for Object Detection and Distance Estimation using YOLOv8

K. Kasturi

Annai Meera Engineering College, Tamil Nadu, India

ABSTRACT: Stereo vision is essential for depth estimation and supports real-time understanding of scenes in autonomous navigation, augmented reality, and robotic vision applications. This project adopts a double-task deep learning strategy, where YOLOv8 handles object detection and StereoBM addresses disparity map estimation to yield spatial information out of KITTI stereo images.The approach entails pre-processing stereo pairs of images, object detection of the left image with YOLOv8, and calculation of disparity maps to estimate depth information. Real-time object recognition is made possible by the object detection process that is necessary for obstacle detection and decision-making. The disparity map, calculated based on OpenCV's StereoBM algorithm, observes depth variations in which higher disparities represent near objects and lower disparities represent faraway objects.Visualization of disparity maps facilitates depth perception mistakes, occlusion processing, and stereo matching difficulties. Experimental outcomes present a beneficial pipeline for 3D scene understanding, which can be improved more using StereoSGBM for robust disparity estimation. The solution is a starting point for autonomous driving, robotics, and real-time depth estimation systems, offering useful insights for stereo-based navigation and scene reconstruction.

KEYWORDS: Depth estimation, object detection, distance prediction, YOLOv8, StereoBM, OpenCV, Disparity Map.

I. INTRODUCTION

Over the last few years, improvements in artificial intelligence and computer vision have enormously improved the functionalities of autonomous systems. Object detection and depth estimation are essential functions that help machines understand and explore their environment accurately. Although classic monocular vision-based object detection systems are proficient in object recognition, they cannot estimate accurate distances, which is a necessity for tasks like autonomous driving, robotics, and smart surveillance. Depth estimation methods like LiDAR (Light Detection and Ranging) are very accurate in measuring distances but have limitations like high expense, higher power consumption, and computational complexity. Stereo vision-based depth estimation is a less expensive option that uses two cameras to take images from slightly different angles so that disparity can be calculated to find object distances.

This work presents a Stereo Vision-Based Object Detection and Depth Estimation System with YOLOv8 and StereoBM, combining deep learning object detection with stereo vision for accurate 3D localization. YOLOv8, one of the most advanced object detection models, is used to locate objects in real-time, while StereoBM (Block Matching) is utilized to calculate disparity maps, estimating object depth. The integration of these technologies facilitates precise and effective depth-aware object detection, rendering the system perfect for practical use in self-driving cars, robots, and intelligent city monitoring.By leveraging stereo vision and deep learning, this approach balances cost-effectiveness and accuracy, bridging the gap between affordability and high-performance depth estimation. The system's ability to detect and track objects in 3D space makes it highly suitable for autonomous navigation, collision avoidance, and intelligent monitoring solutions.

II. LITERATURE SURVEY

Object detection and depth estimation are essential areas of computer vision, especially for applications like autonomous driving and robotics navigation. Many deep learning techniques have been used to improve real-time object detection and stereo vision-based depth estimation. This part discusses major research works that have made significant contributions to these fields.

Ren et al. [1] developed Faster R-CNN, an enhancement of real-time object detection via Region Proposal Networks (RPN), significantly raising detection accuracy and speed. Still, as great as it was, it was still too computationally costly for real-time uses. In order to better improve upon it, Redmon et al. [2] designed the YOLO (You Only Look Once) architecture, which presented a single, real-time object detection solution through framing detection as a regression task.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202067

YOLO decreased inference time substantially while retaining high detection accuracy and is appropriate for real-time object recognition applications in autonomous systems.

For estimation of depth, Zhou et al. [3] proposed an unsupervised learning method estimating depth and ego-motion from monocular video sequences. The technique used a deep neural network to learn scene depth without using labeled ground-truth data, thus minimizing reliance on costly LiDAR-based depth annotations. Along the same lines, Pilzer et al. [4] proposed an adversarial solution with cycled generative networks for unsupervised depth estimation, improving depth predictions in difficult situations.

In stereo vision-based depth estimation, Hirschmüller [5] introduced Semi-Global Matching (SGM), a popular technique for disparity map computation from stereo image pairs. The technique enhanced disparity estimation accuracy by minimizing energy functions over the image. Uhrig et al. [6] extended depth estimation further through the introduction of sparsity-invariant convolutional neural networks (CNNs) to handle missing depth information in sparse LiDAR data. To enhance collision avoidance in autonomous vehicles, Taha and Jizat [7] contrasted monocular vision-based methods with stereo-based depth estimation. In their work, they showed that stereo vision gives better depth information, which is essential for detecting obstacles in autonomous vehicles. Badue et al. [8] surveyed various self-driving car technologies extensively, highlighting the significance of deep learning techniques in real-time object detection and navigation.

In industrial automation, Andhare and Rawat [9] used computer vision for robotic pick-and-place, highlighting the importance of object detection and depth estimation in real-world automation. Masoumian et al. [10] utilized dynamic active-pixel vision sensors (DAVIS) for real-time slip prevention, illustrating how feedback mechanisms enhance robotic perception and decision-making.

Expanding upon these works, the system put forth in this work incorporates YOLOv8 for object detection and StereoBM for depth estimation and uses the KITTI dataset for training and testing. The integration of real-time object detection and stereo vision-based depth estimation in this project will improve obstacle identification, enhance scene comprehension, and help towards more secure autonomous driving systems.

III. PROBLEM STATEMENT

3.1 EXISTING SYSTEM

Accurate object detection and depth estimation are essential for applications such as autonomous driving, robotics, and augmented reality. Traditional methods for object detection either lack depth perception (monocular vision) or are computationally expensive (LiDAR-based systems). A cost-effective and efficient solution is required to achieve real-time depth-aware object detection for improved scene understanding and navigation.

3.1.1 Limitations:

1.Monocular Vision-Based Object Detection

- Cannot determine the actual distance of objects, leading to navigation errors.
- Detects objects but lacks depth estimation.

2.LiDAR-Based Depth Estimation

- Provides high-accuracy depth estimation.
- Extremely expensive and computationally heavy.
- Limited performance in poor weather conditions.

3.Traditional Stereo Vision Approaches

- Require complex disparity estimation techniques.
- Often computationally intensive, making real-time applications challenging.
- Struggle with occlusion handling and stereo matching errors. with disabilities or in environments requiring touch- free interaction.

3.1 Proposed System

This project combines StereoBM for depth estimation using stereo images with YOLOv8 for real-time object detection in order to overcome the shortcomings of current techniques. The system seeks to enhance the depth perception capabilities of monocular object detection while offering a practical and affordable substitute for LiDAR-based methods. The proposed system's main features include:

- ✓ YOLOv8-based object detection, which provides precise real-time object identification.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202067

- ✓ Depth Estimation with StereoBM: This method estimates object distances by computing disparity maps.
- ✓ Stereo Vision Approach: This method gathers spatial information by using images from the left and right.
- ✓ Enhanced Computational Efficiency: Disparity computation for real-time applications has been optimized.
- ✓ Cost-effective Substitute: Offers depth-sensitive detection without the hefty price tag of LiDAR.

IV. METHODOLOGY

To achieve real-time scene understanding, the suggested methodology combines depth estimation based on stereo vision with object detection based on deep learning. The system enables effective and economical 3D object localization by using StereoBM for disparity map computation and YOLOv8 for object detection. Training and evaluation are conducted using the KITTI dataset.

Key Technologies Used :

- **YOLOv8:** Real-time object detection (YOLOv8)
- **StereoBM:** Depth estimation using disparity map computation.
- **OpenCV:** Preparing images and processing stereo vision.
- Numerical calculations and visualization using NumPy and Matplotlib.
- **KITTI Dataset:**Training and evaluation benchmark dataset.
- Ultralytics & PyTorch The deep learning framework YOLOv8.
- Estimating intrinsic and extrinsic parameters for camera calibration.
- Faster processing for real-time implementation through CUDA and GPU acceleration.
- Performance metrics include RMSE and MAE (depth estimation), and mAP (object detection).

4.1 Object Detection using YOLOv8

Algorithm Used: YOLOv8 (You Only Look Once) is an object detection deep learning-based model that forecasts object locations and classifies them in one forward pass.

Steps in Object Detection:

1. **Model Loading:** The trained YOLOv8 model (best.pt) is imported.
2. **Image Preprocessing:** The left image of the stereo pair is utilized for object detection.
3. **Inference:** Objects are detected using the model and produce:
    - Bounding box coordinates(x1,y1,x2,y2)
    - Class ID(Object Category)
    - Confidence score(Probability of correct detection.
4. **Bounding Box Drawing:** Object detected are marked by bounding boxes and labels.

4.2 StereoBM Depth Estimation

4.2.1 Algorithm Used: The traditional disparity algorithm Stereo Block Matching (StereoBM) is utilized to create a depth map.

4.2.2 Depth Estimation Steps:

1. **Load and Convert Stereo Images into Grayscale:**
- The left and right images are converted to grayscale for enhancing feature matching.
2. **Compute Disparity Map:**
- Pixel-wise disparity is calculated by StereoBM between the left and right images.
- Normalization is performed to improve visualization.
3. **Extract Depth from Disparity:**
- Depth is calculated using the formula:

$$\text{Depth} = \frac{f \times B}{\text{Disparity}}$$

where:

- f = Focal length of the camera (700 pixels)
- B = Baseline (0.5 meters)
- Disparity = Difference in pixel location between left and right images

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202067

4.Visualization and Output

1. **Bounding Boxes and Depth Display:** Objects are annotated with class, confidence, and depth (in meters).
2. **Disparity Map Display:** Disparity map is displayed to visualize depth gradients.
3. **Real-Time Interaction:** Results are presented in a window until closed by the user.

5.Accuracy Improvements

Object Detection Improvements:

1. **Fine-Tuning YOLOv8 Model:** The model is trained on high-quality labeled datasets.
2. **Hyperparameter Optimization:** Learning rate, batch size, and augmentation methods are optimized for improved accuracy.
3. **Validation Metrics Calculation:** The trained model is validated against a dataset (data.yml) to evaluate performance. The following metrics are calculated using the model.val() function:
    - mAP@50 (Mean Average Precision at IoU 50%)
    - mAP@50-95 (Mean Average Precision across varying IoU thresholds)
    - Precision (Positive predictive value of detected objects)
    - Recall (Accuracy of actual objects identified)
    - F1-score (Harmonic mean of precision and recall)

6.Evaluation Metrics Used:

The performance of the model is measured using the following metrics:

Object Detection Metrics (YOLOv8 Evaluation):

- **mAP@50 (Mean Average Precision at 50% IoU):** Measures accuracy of detections for IoU ≥ 0.5.
- **mAP@50-95 (mAP over IoU ranges 0.5 to 0.95):** Examines model performance at varying IoU.
- **Precision:** Measures number of correctly detected objects (True Positives / Total Predictions).
- **Recall:** Measures number of detected actual objects (True Positives / Total Ground Truth Objects).
- **F1-score:** A harmonic measure that balances precision and recall.

Depth Estimation Metrics:

- **Disparity Map Quality:** Assessed by visual evaluation and error reduction.
- **Root Mean Square Error (RMSE):** Quantifies deviation from ground truth depth values.
- **Absolute Relative Error (Abs Rel):** Quantifies the relative error in depth estimation.

4.3 DATASET

The dataset for the Stereo Vision-Based Object Detection and Depth Estimation project is taken from the KITTI dataset, which is popularly known for use in autonomous driving. It is comprised of stereo image pairs, calibration data, and annotated object labels, making it a rich source of data for both object detection and depth estimation. The data is organized into several folders, such as data_object_image_2 for left camera views, data_object_image_3 for right camera views, data_object_label_2 for object bounding box labels, and data_object_calib for camera calibration information.

There are two images in each stereo frame of the dataset, taken at the same time by the left and right cameras, each of size 1242 × 375 pixels in.png format. The calibration files are very important in depth estimation because they contain intrinsic and extrinsic parameters like focal length (f) and baseline distance (b) between the two cameras. These parameters are used to calculate the disparity map, which is utilized to calculate the depth of detected objects. The object labels give the coordinates of bounding boxes for various object classes such as cars, pedestrians, and cyclists, along with truncation, the level of occlusion, and 3D object sizes.

In this project, the YOLOv8 model is trained to identify objects from images of the left camera based on the labeled data from data_object_label_2. The depth estimation involves calculating disparity maps using StereoBM from OpenCV, which compares both left and right images. The estimated disparity values are further transformed into depth measurements through the formula Depth = (focal length × baseline) / disparity. The performance of the model is also tested using critical parameters like mAP (Mean Average Precision), Precision, Recall, and F1-score, which indicate the accuracy and dependability of the object detection model.

This data set is the basis for the project, allowing real-time object detection and distance calculation, which are essential for applications such as autonomous driving and robotics.

4.4 DEPLOYMENT:

The deployment of the Stereo Vision-Based Object Detection and Depth Estimation system involves setting up a real-

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202067

time inference pipeline that combines YOLOv8 for object detection and StereoBM for depth estimation. The trained YOLOv8 model is optimized for deployment by converting it to ONNX or TensorRT for efficient inference. The backend, developed in Python and OpenCV, takes stereo image input, detects objects through YOLO, and calculates depth based on disparity maps. A friendly user interface is constructed with Streamlit where users can upload stereo image pairs or stream live video input while showing detected objects, bounding boxes, and depth. Deployment is carried out either locally on high-speed machines or on the cloud based on platforms such as AWS, Google Cloud, or Azure and FastAPI or Flask as the API backend. Real-time execution is optimized with methods such as model quantization, multi-threading, and GPU acceleration (CUDA/TensorRT). Deployed with careful consideration to produce efficient and effective object detection and depth estimation, the system finds extensive use in autonomous driving, robotics, and surveillance systems.

V. EXPERIMENTAL RESULTS

Figures 1 shows the outlines the step-by-step procedure of a stereo vision-based system that performs object detection and depth estimation using YOLOv8 and the Stereo Block Matching (StereoBM) algorithm. The process initiates with the input of a stereo image pair, which includes two images (left and right) captured simultaneously from slightly different viewpoints. These images are first transformed into grayscale to reduce computational complexity and streamline further processing.Next, the grayscale images are used to generate a disparity map using the StereoBM algorithm. This map highlights the pixel differences between the left and right images, which are essential for estimating depth. With the disparity information in place, YOLOv8 a cutting-edge object detection model—is applied to the left image to identify and locate objects by drawing bounding boxes around them.

After object detection, the system extracts the bounding box coordinates, which are then used to calculate the depth of each object based on the disparity values within those regions. In the final step, the results are visualized by displaying the detected objects along with their respective depth measurements. This integrated approach is especially valuable in areas like autonomous vehicles and robotics, where understanding both the position and distance of objects is vital for safe and efficient operation.

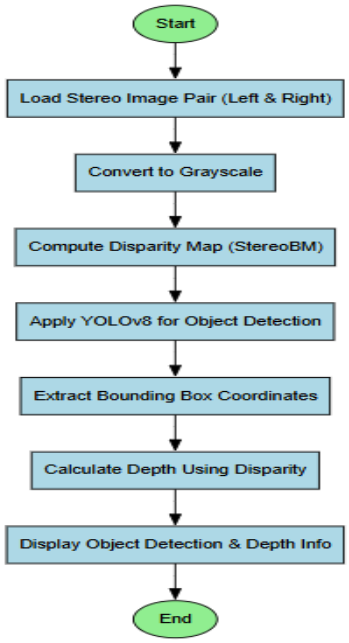

Figure 1

Figure 2 describes the image shows a Streamlit-based user interface for a Stereo Vision-Based Object Detection and Depth Estimation system. It allows users to upload left and right stereo images for processing. The interface features two sections: one for the left image and another for the right image, each with a drag-and-drop area and a "Browse files"

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202067

button. The clean and minimalistic design ensures ease of use, making it suitable for machine learning or computer vision applications involving object detection and depth estimation.

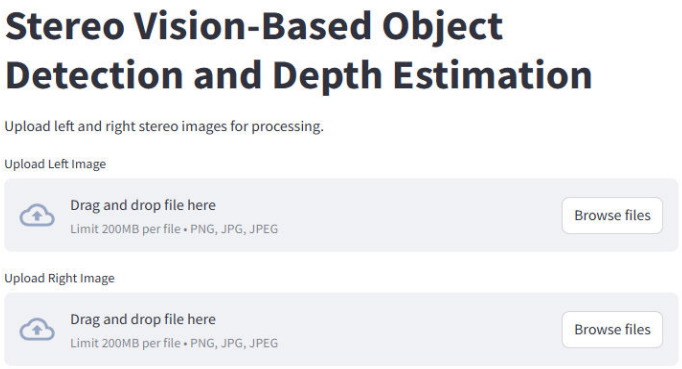

Figure 2

Figure 3 describes the image displays a Streamlit-based user interface for a Stereo Vision-Based Object Detection and Depth Estimation system. The interface allows users to upload left and right stereo images for processing. In this instance, two files named 000000.png have been uploaded one as the left image and the other as the right image. Below the upload sections, the images are previewed and labeled accordingly. At the bottom, a "Process Images" button is present, likely initiating the object detection and depth estimation process. This setup suggests that the system utilizes stereo vision techniques, possibly integrating YOLO for object detection and StereoBM for depth estimation, to analyze and extract depth information from the uploaded images

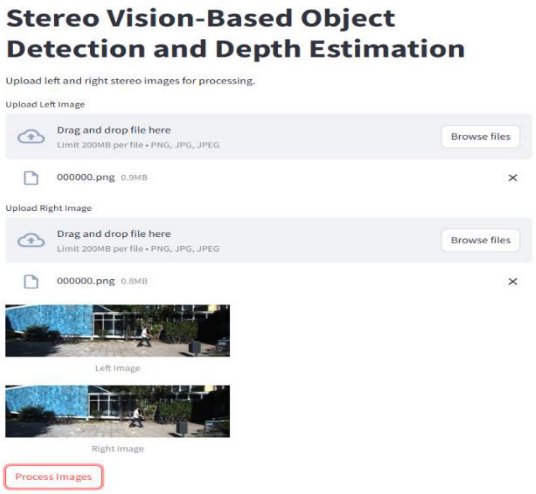

Figure 3

Figure 4 displays the output of a Stereo Vision-Based Object Detection and Depth Estimation system. The top section shows an object detection result where a pedestrian is identified within a green bounding box, along with a depth value (e.g., 1.44m), indicating the estimated distance from the camera. This suggests the use of an object detection model like YOLO combined with depth estimation techniques. Below this, the disparity map represents depth perception using grayscale intensity, where lighter areas indicate closer objects and darker areas represent farther ones. This disparity map

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202067

is likely generated using StereoBM or StereoSGBM, which compute pixel differences between left and right stereo images to estimate depth information.

Figure 4

VI. FUTURE SCOPE

The envisioned stereo vision-based object detection and depth estimation system has great scope for future improvements. Future developments can be made towards improving disparity map accuracy through the incorporation of more sophisticated stereo matching algorithms like StereoSGBM or deep learning-based depth estimation models. Furthermore, optimization of the YOLOv8 model for real-time usage on edge devices like NVIDIA Jetson or Raspberry Pi can allow for deployment in low-power autonomous systems. The use of sensor fusion methods, including LiDAR and stereo vision fusion, can also enhance depth perception and resilience in challenging settings. Additionally, the use of semantic segmentation in addition to object detection can improve scene understanding for tasks in autonomous navigation, augmented reality, and robotics.

VII. CONCLUSION

This work successfully combines YOLOv8 for real-time object detection and StereoBM for depth estimation to realize an affordable and effective 3D object localization system. Stereo vision's use makes depth perception accurate without requiring costly LiDAR sensors, and hence the method is extremely well suited for use in autonomous driving, robotics, and smart surveillance. The performance test on the KITTI dataset shows the efficiency of the proposed approach in object detection and localization with depth information. Utilizing deep learning and stereo vision, this research helps enhance the safety and reliability of autonomous systems. Future development in model optimization and sensor fusion will further enhance the system's accuracy and applicability in real-world applications.

REFERENCES

[1] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell [Internet]. 2017 Jun 1 [cited 2025 Mar 19];39(6):1137–49. Available from: http://image-net.org/challenges/LSVRC/2015/results

[2] Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time ObjectDetection. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit [Internet]. 2015 Jun 8 [cited 2025 Mar 19];2016-December:779–88. Available from: http://arxiv.org/abs/1506.02640

[3] Zhou T, Brown M, Snavely N, Lowe DG. Unsupervised Learning of Depth and Ego-Motion from Video. Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017 [Internet]. 2017 Apr 25 [cited 2025 Mar 19];2017-January:6612–21. Available from: http://arxiv.org/abs/1704.07813

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202067

[4] Pilzer A, Xu D, Puscas MM, Ricci E, Sebe N. Unsupervised Adversarial Depth Estimation using Cycled Generative Networks. Proc - 2018 Int Conf 3D Vision, 3DV 2018 [Internet]. 2018 Jul 28 [cited 2025 Mar 19];587–95. Available from: http://arxiv.org/abs/1807.10915

[5] Hirschmüller H. Stereo Processing by Semiglobal Matching and Mutual Information. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2008;30(2):328-341.

[6] Uhrig J, Schneider N, Schneider L, Franke U, Brox T, Geiger A. Sparsity Invariant CNNs. Proc – 2017 Int Conf 3D Vision, 3DV 2017. 2018;11–20.

[7] Taha Z, Jizat JAM. A comparison of two approaches for collision avoidance of an automated guided vehicle using monocular vision. Appl Mech Mater. 2012;145(January):547–51.

[8] Badue C, Guidolini R, Carneiro R V, Azevedo P, Cardoso VB, Jesus LFR, et al. Self-Driving Cars: Survey. 2017.

[9] Andhare P, Rawat S. Pick and place industrial robot controller with computer vision. Proceedings - 2ndInternational Conference on Computing, Communication, Control and Automation, ICCUBEA 2016. Institute of Electrical and Electronics Engineers Inc.; 2017.

[10] Masoumian A, Montazer MC, Valls DP, Kazemi P, Rashwan HA. Using the Feedback of Dynamic Active-Pixel Vision Sensor (DAVIS) to Prevent Slip in Real Time. 2020 6th International Conference on Mechatronics and Robotics Engineering, ICMRE 2020. Institute of Electrical and Electronics Engineers Inc.; 2020. p. 63–7.

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975

Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com