Boosting Visual Servoing Performance through RGB-Based Methods

Abstract

Purpose - This paper aims to evaluate and compare the performance of different computer vision algorithms in the context of visual servoing for augmented robot perception and autonomy.
Design/methodology/approach - We evaluate and compare three different approaches: a feature-based approach, a hybrid approach, and a machine-learning-based approach. To evaluate the performance of the approaches, we conducted experiments in a simulated environment using the PyBullet physics simulator. The experiments included different levels of complexity, including different numbers of distractors, varying lighting conditions, and highly-varied object geometry.
Findings - The experimental results show that the machine-learning-based approach outperforms the other two approaches in terms of accuracy and robustness. The approach can detect and locate objects in complex scenes with high accuracy, even in the presence of distractors and varying lighting conditions. The hybrid approach showed promising results but was less robust to changes in lighting and object appearance. The feature-based approach performed well in simple scenes but struggled in more complex ones.
Originality/value - This paper sheds light on the superiority of a hybrid algorithm that incorporates a deep neural network in a feature detector for image-based visual servoing, which demonstrates stronger robustness in object detection and location against distractors and lighting conditions.

1 Introduction

Recent years have witnessed a notable increase in academic research pertaining to robot perception, which involves leveraging a diverse range of sensors to facilitate comprehending and interpreting the surrounding environment. With the advancements in computer vision and machine learning techniques, robots can now recognize, interpret, and make decisions based on the perception information gathered (Qiao et al. 2022). This has resulted in the widespread use of robot perception in various applications, including the navigation of mobile robots, providing context-awareness for service robots (Miao et al. 2023), robot arm manipulation (Lin & Wang 2021), manufacturing (Wan et al. 2022, Zeng et al. 2018), mobile robots (Qiu et al. 2019), transportation guidance for logistic (Bloss 2011), and many more.

Among these applications of robot perception, visual servoing stands out as one of the most promising techniques. It utilizes visual feedback from a camera to control the motion and position of a robot arm, making it a versatile and adaptable approach for human-robot interaction (Xue et al. 2020, Li et al. 2022, Bonci et al. 2021, Chen et al. 2020) and teleoperation (Huang et al. 2022, Bacha et al. 2022, Wang et al. 2020, Huang et al. 2023, Wang, Fei, Huang, Rouxel, Xiao, Li & Burdet 2023). One of the major advantages of visual servoing is its ability to work in unstructured environments, as it can adapt to changes in the environment and the objects being manipulated. Moreover, it can operate without requiring a precise or prior model of the robot's dynamics, making it more flexible and easier to implement in practice. There are two primary approaches to visual servoing:

position-based visual servoing (PBVS) and image-based visual servoing (IBVS). In PBVS, vision data is used to reconstruct the 3D pose of the robot and generate a kinematic error in Cartesian space (Palmieri et al. 2012). On the other hand, IBVS generates an error directly from image plane features. IBVS is known for its robustness with respect to camera calibration accuracy and stability under noisy conditions, which makes it well-suited for operation in unstructured environments where it can adapt to changes and operate without requiring a precise model of robot dynamics.

IBVS techniques can be classified according to the type of sensor used, with RGB cameras being a popular and cost-effective option for low-cost robot systems (Bonci et al. 2021). These techniques can be broadly categorized into three main branches: machine learning based approaches, feature-based approaches, and hybrid approaches. Machine learning based approaches, such as YOLO (Redmon et al. 2016), Faster R-CNN (Girshick 2015), Mask R-CNN (He et al. 2017), RetinaNet (Lin et al. 2017), and SSD-6D (Kehl et al. 2017), take an image as input and output the object location in either the image domain or the real-world coordinate system. These algorithms are trained to recognize different types of objects and can adapt to variations in object appearance and lighting. Transfer learning techniques can also be employed to fine-tune these methods for specific tasks. On the other hand, feature-based approaches use extracted features such as edges, corners, and SIFT features (Lowe 2004) to estimate the object pose and location. In contrast to semantic segmentation, which classifies every pixel in the image, template matching uses predefined templates of objects to find their matches and predict the object bounding box. Each template is created by extracting specific features from a set of images containing the target object. Hybrid approaches combine the strengths of both feature-based and machine learning based approaches by leveraging the robustness of feature-based approaches to handle occlusions and changes in lighting while utilizing the adaptability of machine learning-based approaches to handle variations in object appearance.

In this paper, we have selected some of the most representative algorithms for each category and tested them respectively. For machine learning-based approaches, we have tested a state-of-the-art semantic segmentation model, namely DeepLabv3+ (Chen et al. 2018). For feature-based approaches, we have tested SIFT (Lowe 2004) and ORB (Rublee et al. 2011). For the hybrid approaches, we have tested the method proposed in (Kim et al. 2017). This algorithm comprises a convolutional neural network (CNN) extractor to extract features in both the template and captured image and compare their similarities using Normalized Cross-Correlation (NCC) (Yoo & Han 2009). The hybrid approach combines the strengths of both feature-based and machine learning-based approaches to achieve flexible object recognition and localization.

The main contributions of this paper are: (1) providing a comprehensive comparison of three categories of robot arm visual servoing using only RGB information; (2) evaluating the methods on a dataset of rendered synthetic images captured by an RGB camera mounted on a robot arm in simulation; (3) analyzing the strengths and weaknesses of each method and providing suggestions for future work. To achieve these objectives, we select representative algorithms for each category and conduct experiments in a simulated environment.

The rest of this paper is organized as follows: Section 2 reviews related works on robot arm visual servoing methods using RGB information or RGB-D information; Section 3 describes the three methods we compare in detail; Section 4 presents our experimental setup and results; Section 5 discusses our findings and implications; Section 6 concludes this paper.

2 Related Works

IBVS tasks can be addressed using various image processing and computer vision techniques. These tasks can be divided by whether the image has depth information. Depth information provides a detailed representation of the scene's three-dimensional geometry, allowing for a more accurate understanding of the scene's structure and spatial relationships between objects (Litvak et al. 2019,

Zeng et al. 2022). In comparison, RGB images have become increasingly popular for robot applications due to their widespread availability, high-resolution imaging capability, and ability to provide color information for object detection (Kumar et al. 2015), recognition (Wu et al. 2021) and classification.

However, relying solely on RGB information can pose challenges and limitations, such as occlusion, changes in illumination, and noise, which can adversely affect the accuracy and reliability of the control system (Park et al. 2019). Despite these challenges, researchers have continued to explore ways to improve the performance and robustness of visual servoing systems that use only RGB information. For example, (Wang, Tao & Zheng 2023) proposes an attention-based network that is still able to estimate depth information even when an object is partially obstructed from view, which can lead to incomplete RGB information. Direct visual odometry methods have also been used to compensate for illumination changes, which can cause significant variations in RGB information, leading to misinterpretation or incorrect interpretation of visual data (Kim et al. 2015). Additionally, semantic segmentation is utilized in the RGB-based approach to deal with noisy and unstructured environments (Wong et al. 2017).

In this paper, we provide an overview of the state-of-the-art IBVS techniques for robot arm control that rely solely on RGB information and present a comprehensive comparison of IBVS techniques for robot arm control that relies solely on RGB information. The focus of our study is to evaluate and compare the performance of different IBVS methods in terms of their accuracy, stability, and adaptability to changes in the environment and the objects being manipulated. Through our comparative analysis, we provide insights into the strengths and limitations of these methods, and identify areas for further research and development in the field.

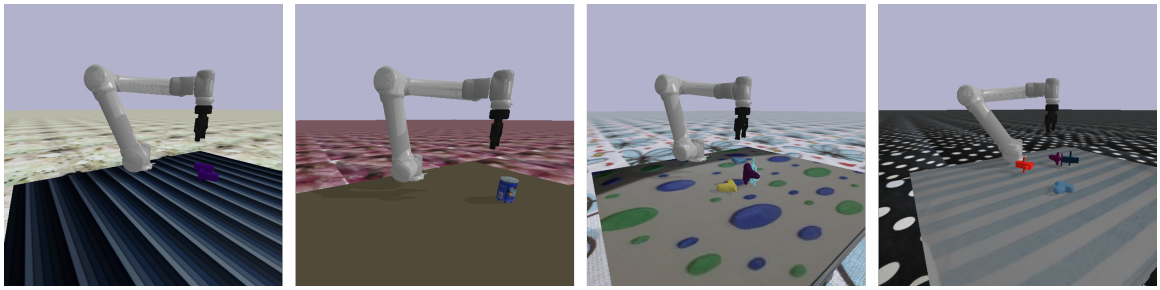

Figure 1: Sample scenes in the simulation environment, which includes a 6 degree-of-freedom (DoF) UR5e robot, a wrist camera, and randomly generated objects and backgrounds.

3 Methods

In this section, we present the methodology employed in the object detection and location system for robot arm IBVS, which comprises three sub-approaches: machine learning, feature-based, and hybrid. Our machine learning approach trains a deep learning model for semantic segmentation and uses connected component labeling to locate the object in the image. In contrast, the feature-based approach extracts distinctive features from the object and scene using Scale-Invariant Feature Transform (SIFT) and Oriented FAST and Rotated BRIEF (ORB) feature detectors and matches them to obtain correspondences between the two. The hybrid approach combines the strengths of both approaches by using the output of the semantic segmentation as a mask to limit the search for matching features, thereby improving efficiency and accuracy. In the following subsections, we provide a detailed description of each approach used to evaluate their performance. We implemented all code and conducted experiments using Python 3.7 and the OpenCV 4.5.1 library.

3.1 Machine Learning-Based Approach

Our machine learning-based approach uses a deep neural network to perform semantic segmentation of the image, followed by connected component labeling to locate the object. Specifically, we employ a fine-tuned DeepLabv3+ (Chen et al. 2018) with a ResNet-50 (He et al. 2016) backbone for semantic segmentation. DeepLabv3+ is a state-of-the-art model that uses atrous convolution and a decoder module to refine the segmentation output. ResNet-50 is a widely used backbone architecture that has shown excellent performance in various computer vision tasks. We fine-tuned the model on our own dataset, which was generated in a simulator and included corresponding ground truth labels, using the cross-entropy loss function.

To accommodate the irregular shape and varying color of the objects in our settings, we train the machine learning model to recognize background instead of object recognition. This approach allows the trained algorithm to detect any non-background objects within the field of view. During the inference stage, the trained model is applied to the input image I_m to generate a pixel-wise semantic segmentation map M, with each pixel classified into one of the pre-defined categories. The binary mask of the object is then obtained using a connected component labeling function, which represents the object's region in the image. Finally, a bounding box of the object is computed based on the binary mask, which is defined as the minimum rectangle that encloses the binary mask. This approach accurately localizes the object in complex and cluttered scenes, and the use of semantic segmentation and connected component labeling makes the approach robust to variations in lighting, viewpoint, and object appearance, which makes it suitable for a wide range of applications.

3.2 Feature-Based Approach

In addition to the machine learning-based approach, we also explore feature-based approaches for object detection and location. Feature-based approaches involve extracting distinctive features from the object and matching them with the features extracted from the scene. In our study, we compare two popular feature detection algorithms (i.e., SIFT and ORB). Both algorithms are widely used and have been verified to be robust to scale, rotation, and illumination changes. The SIFT algorithm detects key points in an image and computes descriptors for each key point based on the scale-space extrema. ORB, on the other hand, computes the descriptor using binary robust independent elementary features (BRIEF) with additional orientation information.

To locate the object using feature-based approaches, we first extract features from both the object template and the input image using either SIFT or ORB. We then match the extracted features from the template and input images using brute-force matching, which results in a set of candidate correspondences. To obtain the final set of correspondences, we use the Random Sample Consensus (RANSAC) algorithm to filter out any outliers. Finally, we locate the object by drawing a bounding box around the set of matched points.

We note that feature-based approaches can be complementary to machine learning-based approaches, as they can provide an alternative means of object detection and location that may be more suitable for certain applications.

3.3 Hybrid Approach

The hybrid approach combines the use of a CNN for feature extraction with similarity comparison algorithms to locate the object in a lower dimension. In our proposed method, we use the VGG-16 backbone neural network to extract feature maps F_t and F_m from both the template image I_t and the input image I_i, respectively.

We first pass the template and input images through the VGG-16 network and extract the feature maps F_t and F_m from the last convolutional layer of the network. The feature maps are then normalized to have zero mean and unit variance. To locate the object in the input image, we use NCC (Yoo & Han 2009) to measure the similarity between the extracted feature maps as a function

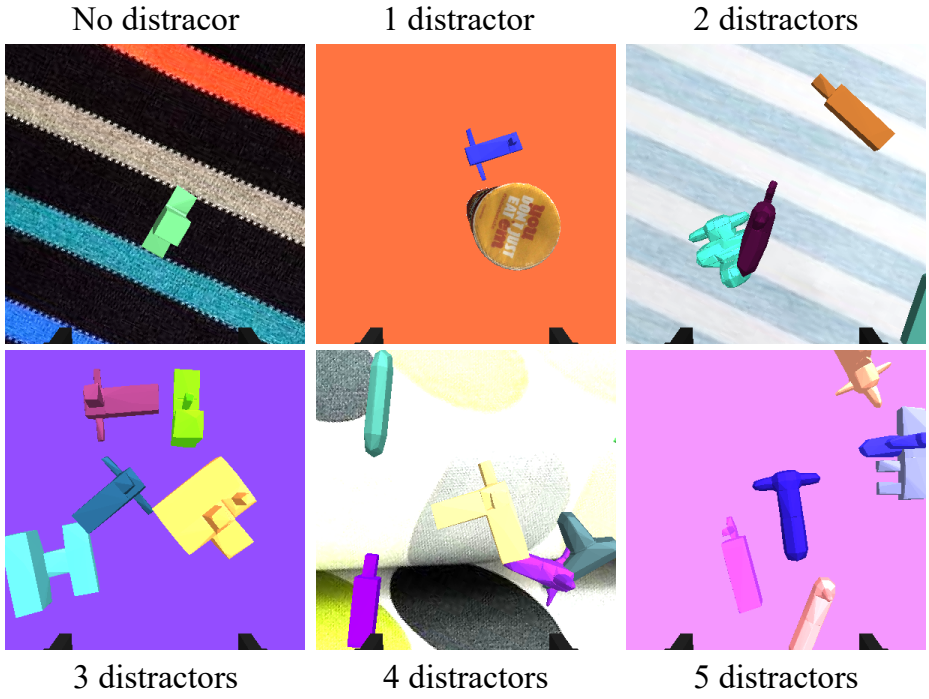

Figure 2: Variation of distractors in the image, ranging from 0 to 5, can affect the performance of image-based visual servoing.

of their relative displacement. NCC is defined as the ratio of the cross-correlation of two images to the product of their standard deviations and is calculated as below:

$$\gamma = \frac{\sum_i \left[F_m - \bar{F}_m\right]\left[F_t - \bar{F}_t\right]}{\sqrt{\sum_i \left[Fm - \bar{F}_m\right]^2 \sum_i \left[F_t - \bar{F}_t\right]^2}}. \tag{1}$$

Here, $\bar{F}_m$ is the mean of F_m in the range under F_t, and $\bar{F}_t$ is the mean of F_t. The coordinate of the matching point (x_{max}, y_{max}) is located at the peak γ_{max} in the cross-correlation. This approach benefits from the ability of CNNs to extract high-level features from images and NCC's ability to accurately locate the object in a lower dimension, making it robust to scale, rotation, and illumination changes.

4 Experimental Setup and Results

We describe the experimental setup and methodology for evaluating and comparing several object detection and location systems for the robot arm IBVS. We utilized PyBullet (Coumans & Bai 2016–2021), a real-time collision detection and multi-physics simulator, to evaluate our approach in simulation. The simulated environment consists of a 6 DoF UR5e robot and a wrist camera with a field of view of 60 degrees and an image size of 256×256 pixels. The robot is controlled using a Cartesian space position controller, and the simulation includes a set of randomly generated rigid objects or daily objects. The background of the working area is generated randomly and it includes both pure color backgrounds and texture backgrounds. The simulation environment is depicted in Figure 1, which shows sample scenes with randomly generated objects and backgrounds. To evaluate

the robustness of the algorithms to changes in the environment, we varied the number of distractors in the scene and the lighting conditions, including the distance between the light source and the object being rendered, the amount of ambient light, and the amount of diffuse and specular lighting in the scene.

To ensure the reliability and validity of our experimental findings, we created a diverse and extensive dataset for object detection and localization. The fine-tuning process of the pre-trained semantic segmentation model required collecting both RGB images and their corresponding labeled segmentation images. However, manual labeling can be a time-consuming task. To overcome this challenge, we employed a combination of synthetic data generation techniques, real-world textures, and scanned 3D objects to automatically generate a large volume of template images, RGB images, and their segmented counterparts. Our dataset encompassed a wide range of objects with varying shapes, sizes, textures, and colors. We also introduced variations in lighting conditions, backgrounds, and camera placements to realistically simulate various real-world scenarios. By utilizing this diverse and rich dataset in our experiments, we were able to comprehensively evaluate the performance of different IBVS methods under a variety of challenging conditions.

In this study, we compare the accuracy, efficiency, and robustness of the proposed IBVS techniques. The accuracy and robustness of the algorithms are evaluated using Intersection over Union (IoU), which is calculated by measuring the overlap between the predicted and ground truth bounding boxes. We tested the algorithms on a set of 50 different objects with varying textures and backgrounds while controlling the camera parameters and environment. Additionally, to add difficulty to the evaluation, we introduced distractors by creating cluttered environments for 2D image-based algorithms. We believe that these evaluations would provide valuable insights into the suitability of different algorithms for IBVS applications.

To evaluate efficiency, we tested the processing time of different algorithms, which directly affects the performance of the robotic system. In an IBVS system, the robot must be able to quickly and accurately locate the object of interest in the image frame and use that information to guide its motion toward the desired goal. If the object recognition and location process is slow, it can significantly degrade the overall performance of the system, leading to slower and less accurate robotic movements. This can be particularly problematic in applications where the robot needs to perform tasks in real-time or where there are time-sensitive constraints. Therefore, it is crucial to develop fast and efficient algorithms for object recognition and location that can meet the speed requirements of the IBVS system. The efficiency evaluation results are presented in Table 1, where we report the average processing time in seconds and the standard deviation across multiple trials.

Table 1: Efficiency Evaluation

Algorithm	Avg (%)	Std
Semantic	0.136	0.0979
ORB	0.0335	0.0672
SIFT	0.0198	0.00614
Hybrid	0.867	0.187

Based on our experiments, we have found that testing algorithms' success rates under various lighting conditions are critical to assess their robustness to environmental changes. In our experiments, we evaluated the performance of the algorithms under different lighting conditions, including changes in the distance between the light source and the object, variations in the ambient, diffuse, and specular coefficients. By controlling these parameters, we were able to simulate various real-world lighting conditions and test the algorithms' ability to generalize. To better illustrate the results, we use the Blinn-Phong model to calculate the illumination for a given set of ambient, diffuse, and specular coefficients. Our findings suggest that machine learning based approach (semantic segmentation and hybrid approach) performs better under diverse lighting conditions and have a

better potential for deployment in real-world applications.

Table 2: Accuracy Evaluation with Different Numbers of Distractors (Accuracy %)

Algorithm	0	1	2	3	4	5
Semantic	91.5	-	-	-	-	-
ORB	83.9	57.2	49.2	29.9	32.7	4.63
SIFT	88.9	73.9	50.7	34.2	54.6	29.8
Hybrid	89.1	74.2	69.6	58.3	35.0	26.4

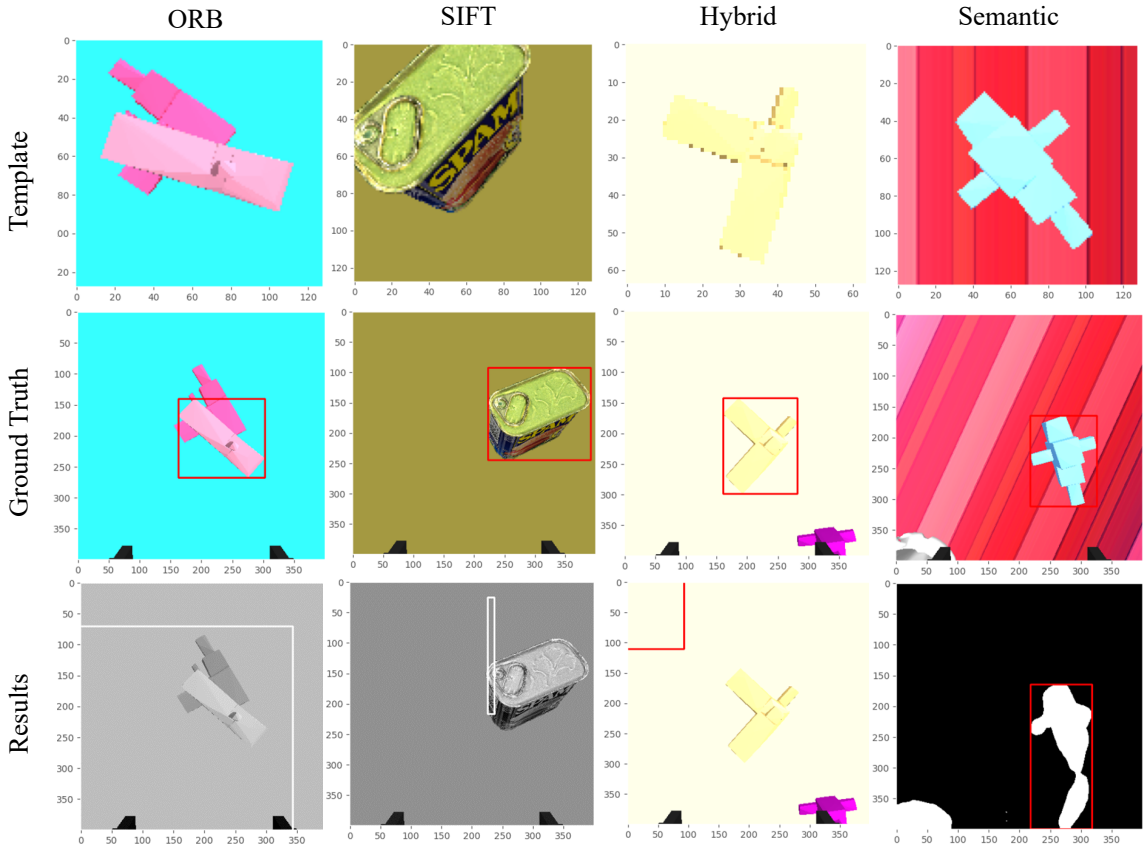

Figure 3: Failure Analysis.

The images in Figure 2 illustrate the effect of adding different numbers of distractors to an image in an IBVS system. The images show examples of the same object with varying degrees of clutter and complexity, which can affect the speed and accuracy of object recognition and localization algorithms. This is a demonstration of the importance of developing robust and efficient algorithms that can handle varying levels of clutter and background noise in image-based visual servoing applications.

5 Findings and Implications

The results of the accuracy evaluation with different numbers of distractors are presented in Table 2. It shows the accuracy percentages of four IBVS algorithms (Semantic, ORB, SIFT, and Feature)

under varying numbers of distractors. The results show that Semantic performs the best with an accuracy of 91.5% on average, even with five distractors. ORB, SIFT, and Feature have lower accuracy compared to Semantic. The accuracy decreases as the number of distractors increases for all algorithms. The highest decrease in accuracy is observed for ORB, which drops from 83.9% without any distractors to only 32.7% with four distractors. SIFT also shows a significant drop in accuracy as the number of distractors increases. These results demonstrate the effectiveness of hybrid approaches over other algorithms in cluttered environments.

The failure images are shown in Figure 3. In analyzing the failure examples of the tested algorithms, it was found that ORB and SIFT algorithms were prone to failure in situations where there were occlusions, lighting changes, or when the object of interest was rotated or scaled. On the other hand, the semantic segmentation algorithm failed when the background texture was complex and resembled objects, leading to confusion between the object and the background. The hybrid approach showed better performance than the other algorithms in terms of accuracy and robustness. However, it was found to be more computationally expensive and slower than the other algorithms. These findings suggest that while the hybrid approach may be suitable for applications where accuracy is paramount and computational resources are not a constraint, other algorithms such as ORB and SIFT may be more suitable for real-time applications where speed and efficiency are critical factors.

Although semantic segmentation cannot be directly used to match unseen objects with the provided template image, we observed that it can be adapted to segment such objects with fine-tuning, which has a relatively low cost compared to the potential accuracy gains. Additionally, we found that the output of semantic segmentation can be used as a mask to narrow down the search space for feature matching. Combining this approach with feature detectors can lead to a promising hybrid solution that reduces computational complexity while improving the accuracy of feature matching. This technique has the potential to enhance the performance of image-based visual servoing in robotic systems, especially in situations where computational resources are limited.

The efficiency evaluation results are shown in Table 1. The semantic method has the highest processing time, with an average of 0.136s, while the SIFT method has the lowest, with an average of 0.0198s. The ORB and hybrid methods have processing times of 0.0335s and 0.867s, respectively. The hybrid method, which combines deep learning and feature matching algorithms, has a higher processing time than the other methods but also achieves the highest accuracy, as shown in Tables 1 and 2. Overall, our results demonstrate the trade-off between accuracy and efficiency in IBVS systems and highlight the importance of developing algorithms that can achieve both high accuracy and fast processing times.

Although our experimental results demonstrate the effectiveness of these IBVS approaches, it is important to consider the limitations and potential challenges of each method when applied to real-world scenarios. For example, machine learning-based methods may be prone to false positives or missed detections in complex scenes with high variability. Additionally, feature-based approaches may struggle with scalability when applied to large datasets. Therefore, it is important for researchers and practitioners to carefully evaluate the strengths and weaknesses of each method before applying them in real-world scenarios.

However, it is crucial to acknowledge that the practicality of the RGB-based IBVS approach can vary depending on the specific environmental conditions. In scenarios where the environment is well-structured and the height of the background is known, depth cameras have the advantage of providing more precise object recognition and detection. Consequently, in cases where there is sufficient budget or familiarity with the environment, the RGB-based approach may be less competitive.

Furthermore, PBVS utilizing deep learning models is an alternative solution for low-cost robot visual servoing. PBVS relies on inferring object pose estimation from 2D information, which may not be as accurate as IBVS which directly estimates errors in the image space. However, the advantage of PBVS lies in its image-to-Cartesian space mapping, which simplifies the design of control laws in the image domain and reduces dependency on camera location. As such, PBVS shows promise as

a viable option for certain applications where accuracy requirements are less stringent and ease of control law design is a priority.

6 Conclusion

In this paper, we have presented a comprehensive study of various feature extraction and matching algorithms for image-based visual servoing applications. Our experimental results demonstrate that the hybrid approach, which combines deep neural networks with traditional feature detectors, achieves the highest accuracy and efficiency among the tested algorithms. We have also identified some of the limitations and failure cases of each algorithm, which can guide future research in this area.

While our proposed approach was developed specifically for visual servoing tasks, it has the potential to be applied to other tasks and scenarios as well. For instance, the utilization of RGB-based robot arm visual servoing algorithms can demonstrate significance in the operations of unmanned aerial or autonomous underwater vehicles within diverse contexts. However, it is important to consider that the effectiveness of these algorithms may vary depending on the environment and specific task at hand. Variations in lighting conditions, and camera placement may impact the efficacy of object recognition and detection.

In summary, our study provides insights into the performance and limitations of various feature extraction and matching algorithms for image-based visual servoing applications. Further research could focus on enhancing the efficiency and robustness of these algorithms, as well as exploring their potential for other computer vision applications.

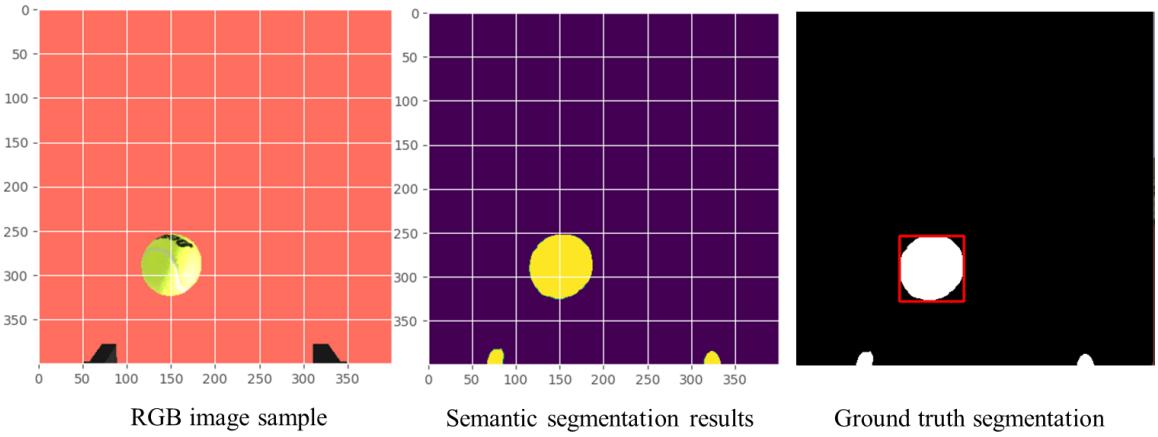

Figure 4: Example results of semantic segmentation, where each color represents a different object class.

References

Bacha, S. C., Bai, W., Wang, Z., Xiao, B. & Yeatman, E. M. (2022), 'Deep reinforcement learning-based control framework for multilateral telesurgery', *IEEE Transactions on Medical Robotics and Bionics* **4**(2), 352–355.

Bloss, R. (2011), 'Automation meets logistics at the promat show and demonstrates faster packing and order filling', *Assembly Automation* **31**(4), 315–318.

Bonci, A., Cen Cheng, P. D., Indri, M., Nabissi, G. & Sibona, F. (2021), 'Human-robot perception in industrial environments: A survey', *Sensors* **21**(5), 1571.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. (2018), Encoder-decoder with atrous separable convolution for semantic image segmentation, *in* 'Proceedings of the European Conference on Computer Vision (ECCV)', pp. 801–818.

Chen, Z., Wang, Z., Liang, R., Liang, B. & Zhang, T. (2020), 'Virtual-joint based motion similarity criteria for human–robot kinematics mapping', *Robotics and Autonomous Systems* **125**, 103412.

Coumans, E. & Bai, Y. (2016–2021), 'Pybullet, a python module for physics simulation for games, robotics and machine learning'.

Girshick, R. (2015), Fast R-CNN, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 1440–1448.

He, K., Gkioxari, G., Dollár, P. & Girshick, R. (2017), Mask R-CNN, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 2961–2969.

He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 770–778.

Huang, D., Li, B., Li, Y. & Yang, C. (2022), 'Cooperative manipulation of deformable objects by single-leader–dual-follower teleoperation', *IEEE Transactions on Industrial Electronics* **69**(12), 13162–13170.

Huang, D., Yang, C., Li, M., Huang, H. & Li, Y. (2023), 'Motion regulation solutions to holding & moving an object for single-leader-dual-follower teleoperation', *IEEE Transactions on Industrial Informatics* pp. 1–12.

Kehl, W., Manhardt, F., Tombari, F., Ilic, S. & Navab, N. (2017), SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 1521–1529.

Kim, J., Kim, J., Choi, S., Hasan, M. A. & Kim, C. (2017), Robust template matching using scale-adaptive deep convolutional features, *in* '2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)', IEEE, pp. 708–711.

Kim, P., Lim, H. & Kim, H. J. (2015), Robust visual odometry to irregular illumination changes with RGB-D camera, *in* '2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)', IEEE, pp. 3688–3694.

Kumar, S., Singhal, P. & Krovi, V. N. (2015), 'Computer-vision-based decision support in surgical robotics', *IEEE Design & Test* **32**(5), 89–97.

Li, Y., Sena, A., Wang, Z., Xing, X., Babič, J., van Asseldonk, E. & Burdet, E. (2022), 'A review on interaction control for contact robots through intent detection', *Progress in Biomedical Engineering* **4**(3), 032004.

Lin, S. & Wang, N. (2021), 'Cloud robotic grasping of gaussian mixture model based on point cloud projection under occlusion', *Assembly Automation* **41**(3), 312–323.

Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. (2017), Focal loss for dense object detection, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 2980–2988.

Litvak, Y., Biess, A. & Bar-Hillel, A. (2019), Learning pose estimation for high-precision robotic assembly using simulated depth images, *in* '2019 International Conference on Robotics and Automation (ICRA)', IEEE, pp. 3521–3527.

Lowe, D. G. (2004), 'Distinctive image features from scale-invariant keypoints', *International Journal of Computer Vision* **60**, 91–110.

Miao, R., Jia, Q. & Sun, F. (2023), 'Long-term robot manipulation task planning with scene graph and semantic knowledge', *Robotic Intelligence and Automation* **43**(1), 12–22.

Palmieri, G., Palpacelli, M., Battistelli, M. & Callegari, M. (2012), 'A comparison between position-based and image-based dynamic visual servoings in the control of a translating parallel manipulator', *Journal of Robotics* **2012**.

Park, K., Patten, T., Prankl, J. & Vincze, M. (2019), Multi-task template matching for object detection, segmentation and pose estimation using depth images, *in* '2019 International Conference on Robotics and Automation (ICRA)', IEEE, pp. 7207–7213.

Qiao, H., Chen, J. & Huang, X. (2022), 'A survey of brain-inspired intelligent robots: Integration of vision, decision, motion control, and musculoskeletal systems', *IEEE Transactions on Cybernetics* **52**(10), 11267–11280.

Qiu, Y., Li, B., Shi, W. & Chen, Y. (2019), 'Concurrent-learning-based visual servo tracking and scene identification of mobile robots', *Assembly Automation* .

Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. (2016), You only look once: Unified, real-time object detection, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 779–788.

Rublee, E., Rabaud, V., Konolige, K. & Bradski, G. (2011), Orb: An efficient alternative to sift or surf, *in* '2011 International Conference on Computer Vision', IEEE, pp. 2564–2571.

Wan, G., Li, F., Liu, B., Bai, S., Wang, G. & Xing, K. (2022), 'A novel robotic 6dof pose measurement strategy for large-size casts based on stereo vision', *Assembly Automation* **42**(4), 458–473.

Wang, X., Tao, C. & Zheng, Z. (2023), 'Occlusion-aware light field depth estimation with view attention', *Optics and Lasers in Engineering* **160**, 107299.

Wang, Z., Fei, H., Huang, Y., Rouxel, Q., Xiao, B., Li, Z. & Burdet, E. (2023), 'Learning to assist bimanual teleoperation using interval type-2 polynomial fuzzy inference', *IEEE Transactions on Cognitive and Developmental Systems* pp. 1–1.

Wang, Z., Tian, Y., Sun, Y. & Liang, B. (2020), 'Finite-time output-feedback control for teleoperation systems subject to mismatched term and state constraints', *Journal of the Franklin Institute* **357**(16), 11421–11447.

Wong, J. M., Kee, V., Le, T., Wagner, S., Mariottini, G.-L., Schneider, A., Hamilton, L., Chipalkatty, R., Hebert, M., Johnson, D. M. et al. (2017), Segicp: Integrated deep semantic segmentation and pose estimation, *in* '2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)', IEEE, pp. 5784–5789.

Wu, B., Zhong, J. & Yang, C. (2021), 'A visual-based gesture prediction framework applied in social robots', *IEEE/CAA Journal of Automatica Sinica* **9**(3), 510–519.

Xue, T., Wang, Z., Zhang, T., Bai, O., Zhang, M. & Han, B. (2020), A new delayless adaptive oscillator for gait assistance, *in* '2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)', pp. 3459–3464.

Yoo, J.-C. & Han, T. H. (2009), 'Fast normalized cross-correlation', *Circuits, Systems and Signal Processing* **28**, 819–843.

Zeng, A., Song, S., Yu, K.-T., Donlon, E., Hogan, F. R., Bauza, M., Ma, D., Taylor, O., Liu, M., Romo, E. et al. (2022), 'Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching', *The International Journal of Robotics Research* **41**(7), 690–705.

Zeng, C., Yang, C., Chen, Z. & Dai, S.-L. (2018), 'Robot learning human stiffness regulation for hybrid manufacture', *Assembly Automation* **38**(5), 539–547.