Research on Industrial Robot Grasping Based on Visual Technology

Changye Du

School of Earth Sciences and Engineering, Southwest Jiaotong University, Chengdu, Sichuan Province, 611756，China
dcy123@my.swjtu.edu.cn

Abstract. Under the backdrop of the rapid development of industrial automation, industrial robots have become the core force for the transformation and upgrading of manufacturing. Integrating visual technology into the grasping of industrial robots changes the traditional operation mode, endowing robots with the "perception - decision-making - execution" closed-loop intelligent operation capability, enhancing their intelligence and flexibility, and enabling them to adapt to complex and changeable industrial environments. This article first provides an overview of the industrial robot vision grasping system, highlighting the key role of the visual perception system. Then, it elaborates on the core technology chain of vision grasping, including image processing which acquires and preprocesses images through cameras and performs recognition and segmentation; three-dimensional positioning which acquires the three-dimensional information of objects through various technologies; grasping pose estimation which infers the pose by combining visual and other sensing technologies or deep learning; and grasping planning and path guidance which plan collision-free trajectories and optimize them, while correcting deviations through visual servoing. Currently, this technology is evolving towards greater intelligence and other directions, which can empower manufacturing and support emerging scenarios. However, it faces challenges such as extreme environment recognition and multi-modal data fusion. In the future, with the in-depth integration of technologies such as deep learning, it is expected to improve the accuracy and robustness of industrial robot grasping and promote the intelligent upgrade of manufacturing.

Keywords: Visual technology, industrial robot grasping, target recognition, grasping path planning

1 Introduction

Nowadays, with the rapid development of industrial automation, industrial robots, due to their efficient, precise, and stable characteristics, have become a core driving force behind the transformation and upgrading of the manufacturing industry. In some work environments with severe conditions, such as high repetition, high temperature, high precision, high strength, and stringent requirements, humans often suffer serious

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_42

consequences due to fatigue, distraction, and other human factors, making it challenging to ensure production efficiency. However, industrial robots can replace humans in such severe environments, reducing human errors, saving costs, increasing productivity, and thereby lowering production costs [1]. Integrating visual technology into the operation of industrial robots differs from the traditional operation mode of industrial robots. Visual robot grasping no longer relies on on-site teaching or offline programming [2]. It not only saves time and costs but also significantly improves the intelligence and flexibility of industrial robots in the grasping process, better meeting various industrial demands. Visual technology endows robots with the closed-loop intelligent operation ability of "perception - decision - execution", enabling them to complete multiple grasping tasks adaptively in complex and changing industrial environments, significantly improving the level of production automation.

From the perspective of development history, industrial robot grasping has undergone an evolution process from manual operation to teaching programming, and then to the introduction of machine vision. The early visual technology primarily focused on two-dimensional image recognition, capable of achieving basic detection of target positions and shapes, but lacked sufficient stability under changes in lighting, target occlusion, and environmental interference. With the rapid development of deep learning and computer vision, algorithms such as Convolutional Neural Networks, the series of You Only Look Once, and Faster Region-based Convolutional Neural Network have been widely applied in target recognition and image segmentation, effectively enhancing the robot's recognition and positioning capabilities in complex backgrounds. At the same time, the introduction of three-dimensional vision and point cloud processing technology enables robots to obtain the depth information and posture of the target, achieving more accurate spatial positioning and grasping path planning. Thus, industrial robots based on visual technology are gradually acquiring stronger intelligence and flexibility, enabling them to adapt to diverse industrial demands.

Research on visual technology-based industrial robot grasping is not only a crucial means to address the challenges of production efficiency and safety in complex environments, but also a key link in achieving the intelligent upgrade of the manufacturing industry. This paper will conduct a systematic analysis of the application of visual technology in industrial robot grasping, first introducing the elemental composition of the industrial robot grasping system, then elaborating on the core technical chain involved in visual grasping, including image processing, target recognition, three-dimensional positioning and grasping path planning, and finally discussing the current challenges and future development directions of the technology. Through systematic review and research, it aims to provide references and inspirations for the engineering application and intelligent development of visual technology in industrial robot grasping.

2 Industrial Robot Vision Grasping System

The industrial robot grasping system is the core for achieving precise and efficient automation. Its core components consist of the robot body, the end effector, the perception system, and the control system. With the development of vision technology, the perception system based on vision technology usually uses visual sensors and multimodal fusion to enable the robot to have the same external perception ability as humans, capable of real-time identification of the types, postures, and position deviations of the workpieces, and automatically adjusting the motion trajectory of the mechanical arm and the grasping method of the end effector. This is the prerequisite for industrial robots to perform various grasping operations. The control system operates in conjunction with algorithm modules, such as vision recognition and motion planning, receiving instructions through the central controller and planning the path to ensure the safety and efficiency of robot grasping. Thus, vision technology plays a vital role in the grasping process of industrial robots, and its core technology provides a guarantee for the accurate and safe grasping of robots.

3 The Core Technical Chain of Visual Grasping

During the process of industrial robot grasping, 2D industrial cameras, line array cameras, and other sensors serve as the "eyes" of the robot, obtaining two-dimensional image information of the on-site environment and the target. They convert light signals into digital signals that the robot can recognize. Software algorithms extract features from the image data and ultimately output structured information that can be used for decision-making, such as the target position and defect type. Accurate image processing provides the core input for robot grasping, ensuring the smooth progress of subsequent grasping operations.

3.1 Image Preprocessing

In complex on-site environments, due to unstable lighting, low visibility, and electrical noise, the photos captured by machine cameras often become blurry or distorted, with unclear details, making them challenging to recognize. To solve this problem, it is necessary first to perform grayscale processing on the captured images, then use fusion algorithms such as Gaussian filtering and median filtering to reduce noise in the original photos, and enhance the image information by methods such as contrast stretching, histogram equalization, and highlighting the differences between the target and the background. Additionally, Priyanka Bhatambarekar studied and proposed a new method for fusing visible light and infrared images using Texture-Specific Generative Adversarial Networks（TSGAN）, generating composite texture maps [3]. The structure layout of TSGAN is shown in Figure 1. It primarily extracts features from the input images through a generator sub-network comprising four convolutional layers with a stride of 1, combines the encoder to achieve weighted feature addition, and ultimately generates more detailed feature outputs. Experimental

results show that multimodal image fusion can significantly improve the image preprocessing effect.

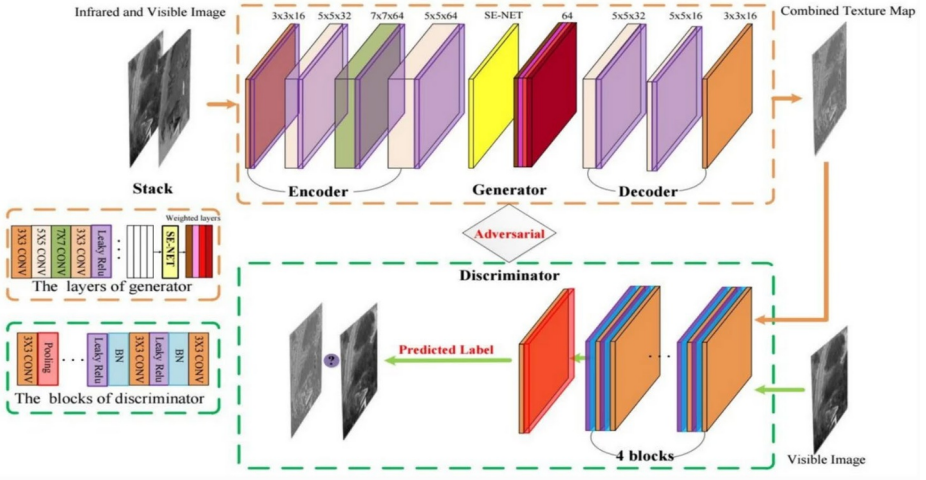

Fig. 1. Explanation of TSGAN Layer Count [3]

Meanwhile, An Zhinan proposed an image enhancement algorithm based on Gabor filtering, multi-scale Retinex, and adaptive compensation for the gray world theory [4]. By preprocessing the contrast of the dark areas of the image using the double gamma function and suppressing the over-enhancement processing of the bright regions, the algorithm not only improves brightness and highlights details, but also maintains high color saturation. Stefano Aleotti proposed a variable metric strategy, which accelerated the convergence of the proximal gradient algorithm and solved the problem of blurry images captured [5]. Preprocessing can not only simplify the algorithm complexity and improve computing efficiency, but also provide a more reliable input for subsequent edge detection, feature matching, and other operations, thereby enhancing the overall processing accuracy and convenience.

3.2 Target Recognition and Image Segmentation

For industrial robots to comprehend the content of images and facilitate subsequent automated decision-making and task execution, it is essential to identify the positions and ranges of target objects within the preprocessed images and locate and identify other relevant targets within the pictures. Currently, standard recognition techniques can be classified into traditional recognition methods and deep learning recognition methods, based on the recognition methods themselves, which include their categories, working principles, and features (as shown in Table 1).

Table 1. Methods Category, Working principle, and characteristics

	Traditional identification methods	Deep learning recognition method

category	Feature + Classifier Support Vector Machine	Single-stage detector YOLO series	Two-stage detector Faster R-CNN
Working principle and characteristics	Manual design features, combined with classifiers such as SVM, are used to determine target categories and position the sliding window. It is easy to implement, but manual feature extraction is inaccurate, with poor generalization ability, and the accuracy is low in complex scenarios.	The image is divided into grids, and the category of the target within each grid is directly predicted. The probability and position coordinates are used to achieve real-time recognition, which is fast but requires a large amount of data.	First, candidate boxes are generated through the region proposal network. Then, after extracting features through the convolutional layer, classification and boundary box regression are completed simultaneously, albeit at a relatively slow speed.

In 2016, Redmon, J., and Divvala, S. first applied YOLO, redefining target detection as a single regression problem, which only requires viewing the image once to predict the existence of objects and their positions, thereby achieving real-time detection and significantly improving the efficiency of target recognition [6]. Wang tested the significant target detection method based on deep learning and verified its reliability [7].

With the increase in industrialization demands, industrial robots need to precisely outline the contours of the targets, distinguish the boundaries between the targets and the background, and even refine to the different components within the targets. Through image segmentation, the disordered pixel data can be transformed into region information with semantic logic. As shown in Figure 2, according to the segmentation principles and image data characteristics, image segmentation techniques are mainly divided into classical segmentation, collaborative segmentation, and semantic segmentation based on deep learning [8].

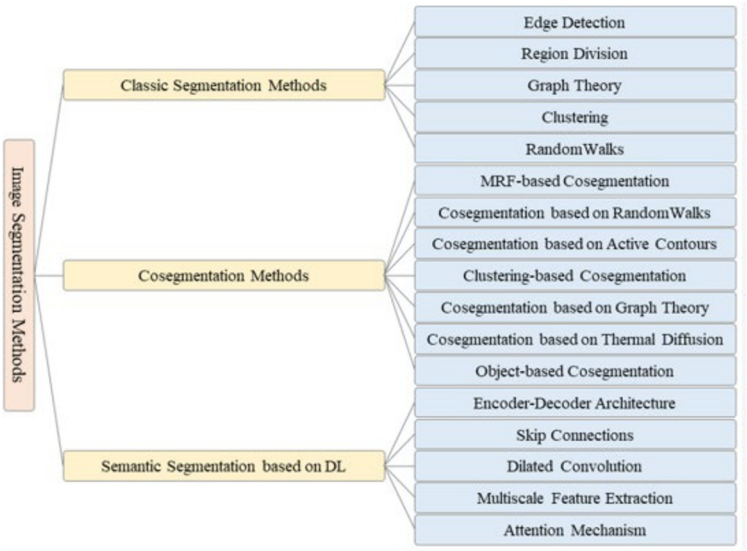

Fig. 2. Classification of Image Segmentation Techniques [8].

Currently, the process of image segmentation has shifted from a single-image approach to utilizing standard features of massive datasets, transitioning from coarse-grained to fine-grained, and from manual feature extraction to adaptive learning [9]. Among them, convolutional neural networks (CNNs) are widely used in image classification, with the output layer representing the category of the image. However, semantic segmentation requires mapping the advanced features back to the original image size after obtaining advanced semantic information. In this process, the encoder gradually reduces the size of the feature map and increases the number of channels through convolutional layers and downsampling, discarding redundant spatial information to extract abstract, advanced semantic features; the decoder expands the feature map through transposed convolution and reduces the number of channels, restoring the spatial details of the image. Some architectures also incorporate skip connections to pass the low-level detail features of the encoder, and the two work together to support tasks such as image segmentation, achieving pixel-level output that balances semantics and details, and significantly improving the segmentation accuracy and robustness through end-to-end training.

3.3 Visual three-dimensional positioning

Compared with two-dimensional planar vision technology, three-dimensional vision positioning technology can provide depth information and spatial posture of objects, which is the key for industrial robots to achieve precise multi-directional grasping in complex situations at present. Three-dimensional positioning obtains the three-dimensional geometric information of the target object through vision (such as binocular vision, structured light, et) [10]. Firstly, by using feature matching algorithms, corresponding feature points of two camera planes and projection planes are extracted from the preprocessed images to establish a one-to-one correspondence

relationship between the two plane coordinate systems. Subsequently, by combining the calibrated internal and external parameters of the camera, the pixel coordinates of the camera plane are converted into normalized imaging coordinates in the camera coordinate system, and the spatial posture of the camera coordinate system is associated with the projection plane world coordinate system. Finally, based on the principle of triangulation measurement, using the geometric constraints of the corresponding points in the camera coordinate system and the projection plane world coordinate system, the spatial coordinates of this point in the three-dimensional world coordinate system are calculated, thereby obtaining the depth, contour, etc. of the target in the complete three-dimensional information from the two-dimensional projection image.

In traditional methods, both monocular and binocular vision obtain three-dimensional information based on the principle of disparity, but are susceptible to light influences. Xin Tian developed a novel three-dimensional reconstruction method that combines polarization imaging and binocular stereo vision fusion to achieve high-quality three-dimensional reconstruction [11]. Structured light technology can project encoded patterns onto the surface of an object and calculate depth information by measuring the phase difference between the encoded patterns. Zhenzhou Wang proposed a single-view structured light three-dimensional reconstruction method based on double views [12]. The imaging system diagram is shown in Figure 3. By processing the structured light points in two images to reconstruct the three-dimensional shape of the object and calculate its three-dimensional coordinates, this approach provides more precise point cloud data information for the target, which is beneficial for industrial robot grasping.

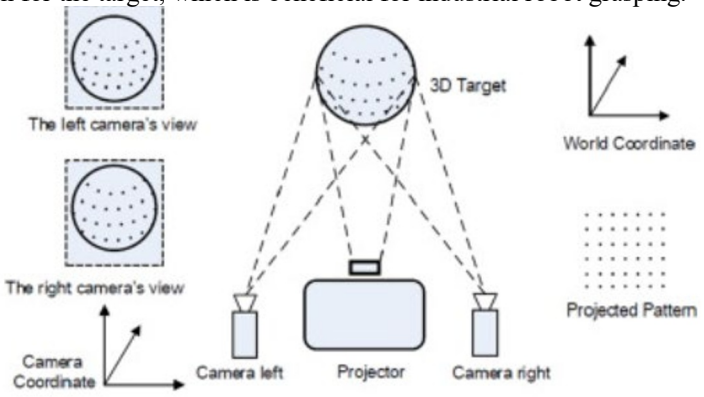

Fig. 3. Imaging System Diagram [11].

3.4 Pose Estimation for Grasping

After the target is identified and located, the point cloud data of the target is output based on its three-dimensional positioning. Then, the local features of the point cloud are extracted to infer the relative pose information required for the end effector to grasp. Standard methods include feature matching methods, camera pose estimation

methods, deep learning methods, and motion capture methods [13]. However, in the real industrial environment, workpieces are highly differentiated, and different objects have different optimal grasping postures, which makes the traditional method of matching the 3D model of the target object with the known model difficult to apply to most complex industrial scenarios.

Currently, combining vision with other sensing technologies is the main research field of pose estimation. Ottenhaus S uses Gaussian process implicit surface fusion of vision and tactile perception to estimate the object surface [14]. Through methods such as the fusion of vision and tactile perception to estimate the target posture and thereby obtain the relative pose of the end effector. In addition, with the rapid development of deep learning, deep learning is widely applied to robot grasping pose estimation. Caixia He employs a deep learning-based method to identify the smooth plane of the target using a semantic segmentation network and determine the normal vector of the aircraft to obtain the posture information of the target [15]. Grasping pose estimation provides a precise target posture for grasping path planning, which is the basis for ensuring that industrial robots can contact and grasp objects in a stable and reliable posture.

3.5 Planning and Path Guidance

When industrial robots are deployed in practical applications, they may encounter situations where the surrounding environment is chaotic. Under the guidance of the machine vision system, during the movement of the robot's end effector to the grasping position of the target object estimated by vision, a safe, efficient, collision-free, and motion-constrained motion trajectory is planned, and the end effector is guided to move along the predetermined path. Rosell J proposed a new importance sampling method based on principal component analysis to increase the probability of finding collision-free samples in these low-gap difficult areas of the configuration space [16].

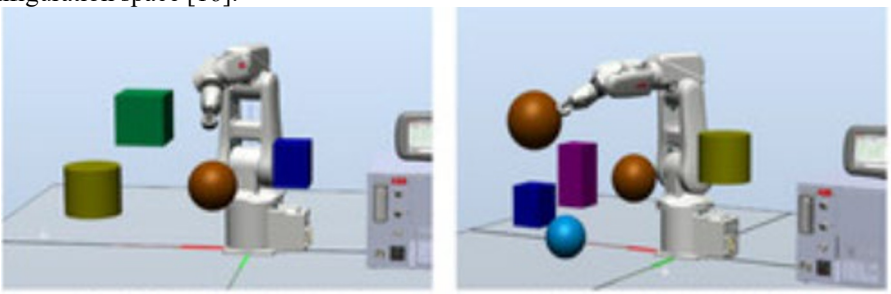

Fig. 4. Scenarios with obstacles of different shapes and quantities [16].

After finding the collision-free path, the optimal grasping route needs to be selected. An effective path needs to be searched in the high-dimensional constrained space, and the performance indicators need to be optimized. Meng X proposed an Elite Smooth Ant Colony Algorithm (ESACO) for the spatial obstacle avoidance path planning of the grasping manipulator [17]. The simulation of the obstacle avoidance scenario is shown in Figure 4, which significantly improves the robot's grasping

efficiency. Amber Xie proposed a Language Condition Collision Function (LACO) to predict collisions between the robot and its environment, enabling flexible conditional path planning without the need for manual object annotations, point cloud data, or ground-level real-world object meshes [18]. Even if the theoretical optimal path is generated based on visual information during path planning, the end effector may still deviate from the planned path due to mechanical errors and environmental disturbances during actual movement. Real-time correction of the grasping process is required through visual servoing technology. Among them, position-based visual servoing detects the 3D pose deviation of the end effector from the target in real-time using vision and directly feeds back to the control module of the grasping system to adjust the end-motion trajectory promptly. At the same time, image-based visual servoing can directly analyze "feature deviations" at the image level without the need for three-dimensional information, and can adjust the robot's movement to return the image features to the expected position.

4 Conclusion

The industrial robot's grasping technology, based on visual technology, serves as the core link connecting the robot's perception and operation. Its development level directly determines the robot's adaptability and operational efficiency in complex manufacturing environments, and it holds a milestone significance in promoting the transition of the manufacturing industry from automation to intelligence. In summary, this technology is evolving towards greater intelligence, greater flexibility, and greater reliability. Its further development will not only empower the manufacturing industry to achieve higher levels of automated production but also provide core technical support for emerging scenarios, such as flexible manufacturing and human-machine collaboration, ultimately driving the industrial production model towards an intelligent era that is efficient, precise, and low-cost.

However, the current technology still faces several challenges, including improving the stability of recognition under extreme lighting conditions and severe occlusion environments, further optimizing the efficiency of deep integration of multimodal sensing data, and refining adaptive grasping strategies for complex workpieces. In the future, with the continuous development of technologies such as deep learning and multimodal fusion, industrial robot visual grasping will move towards higher precision, stronger robustness, and better adaptability, providing more solid technical support for the intelligent upgrade of the manufacturing industry and promoting breakthroughs in industrial production efficiency and quality. This research needs to continuously focus on addressing technical pain points, strengthening theoretical innovation, and integrating engineering applications to accelerate the implementation and iteration of this technology in actual production.

References

1. Wang, L., Fang, Q., Zhang, J.: Industrial robot application and markup rates: Evidence from Chinese manufacturing firms (2011–2022). Finance Research Letters 81, 107467 (2025)
2. Lin, Y., Chen, X.: Research progress on robot positioning and grasping based on machine vision. Automation and Instrumentation (3), 9–12 (2021)
3. Bhatambarekar, P., Phade, G. Investigation of TsGAN-based multimodal image fusion to augment image pre-processing abilities. Journal of Electrical Systems and Information Technology 12, 47 (2025)
4. Aleotti, S., Donatelli, M., Krause, R., et al.: A preconditioned version of a nested primal-dual algorithm for image deblurring. Journal of Scientific Computing 103(3), 85 (2025)
5. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Las Vegas (2016)Wang,
6. W., Lai, Q., Fu, H., et al.: Salient object detection in the deep learning era: An in-depth survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(6), 3239–3259 (2021)
7. Brar, K.K., Goyal, B., Dogra, A., Mustafa, M.A., Majumdar, R., Alkhayyat, A., Kukreja, V.: Image segmentation review: Theoretical background and recent advances. Information Fusion 114, 102608 (2025)
8. Yu, Y., Wang, C., Fu, Q., et al.: Techniques and challenges of image segmentation: A review. Electronics 12(5), 1199 (2023)
9. An, Z., Wei, Y.: Low illumination image enhancement algorithm based on gray world and Retinex. Journal of Chinese Computer Systems (2024)
10. Ye, X., Qin, X., Zhan, L., Wang, J., Chen, Y.: Research on a fusion technique of YOLOv8-URE-based 2D vision and point cloud for robotic grasping in stacked scenarios. Applied Sciences 15(12), 6583 (2025)
11. Tian, X., Liu, R., Wang, Z., Ma, J.: High quality 3D reconstruction based on fusion of polarization imaging and binocular stereo vision. Information Fusion 77, 19–28 (2025)
12. Wang, Z., Zhou, Q., Shuang, Y.: Three-dimensional reconstruction with single-shot structured light dot pattern and analytic solutions. Measurement 151, 107114 (2019)
13. Li, T., Yan, Y., Yu, C., et al.: A comprehensive review of robot intelligent grasping based on tactile perception. Robotics and Computer-Integrated Manufacturing 90, 102792 (2024)
14. Ottenhaus, S., Renninghoff, D., Grimm, R., et al.: Visuo-haptic grasping of unknown objects based on Gaussian process implicit surfaces and deep learning. In: 2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids), pp. 402–409. IEEE, Toronto (2019)
15. He, C., He, L.: Deep learning-based mobile robot target object localization and pose estimation research. International Journal of Advanced Computer Science and Applications (IJACSA) 14(6) (2023)
16. Rosell, J., Suárez, R., Pérez, A.: Path planning for grasping operations using an adaptive PCA-based sampling method. Autonomous Robots 35(1), 27–36 (2013)
17. Meng, X., Zhu, X.: Autonomous obstacle avoidance path planning for grasping manipulator based on elite smoothing ant colony algorithm. Symmetry 14(9), 1843 (2022)
18. Xie, A., Lee, Y., Abbeel, P., James, S.: Language-conditioned path planning. In: Conference on Robot Learning, pp. 3384–3396. PMLR, Atlanta (2023)