Theory and Practice of Science and Technology
2025, VOL. 6, NO. 8, 164-167
DOI: 10.47297/taposatWSP2633-456930.20250608

A Review of Deep Learning-Based Industrial Robot Grasping Systems

Ziyang Yan

School of Mechanical Engineering, Hefei University of Technology, Mechanical Design, Manufacturing, and Automation, Hefei, Anhui, 230001, China

ABSTRACT
Traditional industrial robot grasping methods rely on predefined rules and structured environments, making them ill-suited for complex scenarios and dynamic changes. With breakthroughs in deep learning across visual perception and reinforcement learning, robotic grasping technology is progressively shifting toward a data-driven intelligent paradigm. This paper focuses on two key technologies: visual perception and grasping strategy optimization. In visual perception, grasp detection and 6D pose estimation based on convolutional neural networks significantly enhance system robustness in complex environments. In strategy optimization, imitation learning and deep reinforcement learning offer novel pathways for autonomous decision-making and skill acquisition in dynamic environments. However, current systems still face challenges including data scarcity, simulation-reality gaps, and insufficient multimodal perception fusion. Future research should focus on few-shot learning, transfer techniques from simulation to reality, multimodal information fusion, and non-rigid object grasping to advance the development of general-purpose intelligent grasping systems.

KEYWORDS
Industrial robots; Grasping systems; Visual perception; Grasping strategy optimization; Deep learning

1 Introduction

Intelligent robotic arms have long been a critical hardware system in automated production lines and advanced manufacturing processes. They consistently and silently orchestrate complex workflows across diverse production stages—such as loading/unloading operations, material sorting and handling, and component assembly. At its core, the grasping action involves the equipment's innate ability to interact with the physical environment. Its execution often determines the overall system's agility in maneuvering and adapting to changes. This becomes especially evident when actual working environments are unpredictable.

Traditional industrial robotic arms primarily relied on highly precise, pre-programmed instructions and strictly ordered assembly spaces to ensure consistent, stable movements. While this approach provided a solid foundation, it inevitably resulted in rigidity and inflexibility. When on-site objects are elusive in position, diverse in shape and size, and subject to frequent updates—compounded by potential visual obstructions or turbulent interference—traditional fixed-path approaches become severely constrained. This reveals that current automated execution platforms exhibit limited flexibility when confronting non-structural production challenges. Therefore, under new production demands, further advancement is required to achieve dynamic layout optimization and rapid response to external changes.

In recent years, with the continuous evolution and breakthroughs of deep learning technology in computer vision and reinforcement learning, the research paradigm for robotic grasping tasks has undergone a significant transformation. Its methodological core has gradually shifted from traditional "hard-coded" and explicit rule design toward an intelligent approach driven by large-scale data. Leveraging its end-to-end training capability, deep learning can directly extract discriminative hierarchical features from raw high-dimensional sensor inputs. It autonomously learns grasping representations and decision strategies adaptable to complex scenarios, significantly enhancing the system's perceptual robustness and task generalization capabilities in unfamiliar objects and unstructured environments.

Particularly within visual perception modules, grasp detection models based on convolutional neural network architectures now achieve high-precision, real-time estimation of grasp locations, poses, and mass properties, substantially improving the stability and response speed of visual servo closed-loop systems. At the behavioral generation and optimization level, the introduction of approaches such as deep reinforcement learning and imitation learning offers promising solutions for sequential decision-making and autonomous skill learning in dynamically changing, partially observable environments. These methods are increasingly becoming the critical technological foundation for realizing flexible grasping agents.

Despite significant breakthroughs in robot grasping intelligence, the field still faces critical challenges. Issues such as scarce training data, discrepancies between simulation environments and the physical world, and insufficient multimodal perception fusion continue to limit system deployment effectiveness in real-world scenarios.

This paper aims to systematically review the current state and developmental trajectory of deep learning-based industrial robot grasping technology, focusing on in-depth analysis of two key technical modules: visual environment perception and grasping motion planning. By summarizing and comparing existing representative methods, we seek to clarify the primary paths of technological evolution and identify bottlenecks. Based on this foundation, we project future research directions to provide theoretical references and technical inspiration for researchers and practitioners in related fields.

2 Fundamental theory of industrial robot gripping systems

The fundamental theory of industrial robot gripping systems encompasses the entire process from environmental perception to executing gripping actions. Its core components consist of three primary modules: visual perception, gripping planning, and motion control. These three elements work in concert to form an intelligent gripping system capable of handling complex environments.

Visual perception serves as the front-end component of the gripping system, primarily responsible for acquiring and interpreting scene information. Sensors such as RGB cameras, depth cameras (RGB-D), or LiDAR are typically employed to collect image or point cloud data. Its primary tasks include object recognition, estimation of object position and pose, and generation of suitable grasp configurations. Grasp configurations are commonly represented using grasp rectangles or 6D poses, with the former often used for 2D images and the latter suited for 3D grasp localization. Traditional approaches largely rely on manually designed features and geometric models, whereas modern methods increasingly utilize deep learning models, such as convolutional neural networks (CNNs), to achieve end-to-end grasp detection and pose estimation. These novel methods significantly enhance system stability and adaptability in complex backgrounds, occlusion scenarios, and varying lighting conditions.

Grasp planning involves selecting the optimal solution from multiple candidate grasping options based on perception. The planning process requires comprehensive consideration of object geometry, physical properties (e.g., mass, friction coefficient), gripper type (two-finger/multi-finger, suction cup, etc.), and environmental constraints (e. g., obstacle avoidance). Common approaches include sampling-based algorithms (e. g., Monte Carlo methods), optimization-based methods (e.g., force-enclosed analysis), and recently emerging learning-based methods (e.g., reinforcement learning and imitation learning). The goal of grasp planning is to maximize grasp success rate and stability while maintaining adaptability in dynamic environments.

Motion control serves as the execution component of the grasping system, primarily tasked with translating planned grasp positions and poses into actual robotic arm movements. This process typically involves path planning and trajectory optimization to enable smooth, efficient movement to target locations while maintaining stability during grasping. Inverse kinematics (IK) solving is a critical step, converting 3D spatial poses into joint angles. Additionally, the system must perform real-time obstacle avoidance, torque control, and grasp force adjustment. Force control capabilities are particularly crucial when working alongside humans or handling fragile objects.

In summary, the foundational theory of industrial robotic grasping systems establishes a complete closed-loop process from "perception" to "execution." Visual perception understands the environment, grasp planning makes decisions, and motion control executes actions. These three components work together to enable robots to perform stable, reliable grasping operations in unstructured, dynamically changing industrial settings, providing essential technological support for smart manufacturing.

3 Advances in visual perception technology research

Deep learning-based visual perception serves as the "eyes" of a grasping system, with its core task being grasp detection—directly predicting the grasp pose and quality for each pixel or candidate region within an image.

Grasp detection based on RGB-D data: Early works such as Jiang et al. (2011) proposed methods using handcrafted features. Lenz et al. (2013) pioneered the application of deep learning (shallow networks) to grasp detection, though inference remained slow. The AlexNet model proposed by Redmon et al. (2014) significantly accelerated detection, enabling end-to-end real-time prediction. Subsequent research primarily builds upon these foundations, such as employing encoder-decoder architectures like U-Net for pixel-level predictions (GG-CNN), further enhancing accuracy and speed.

Methods Based on 6D Pose Estimation: For objects with known models, directly estimating their complete 6D pose (3D translation + 3D rotation) represents another approach. Methods like PoseCNN predict an object's 3D translation and rotation through deep learning, then query a predefined grasp point library for that object to obtain the grasp pose. This approach achieves extremely high accuracy when grasping known objects.

Multimodal Fusion: Effectively integrating the rich texture information from RGB images with the geometric structure information from depth images is crucial. Mainstream approaches include early fusion (concatenating depth maps as a fourth channel with RGB), late fusion (merging features extracted separately), and the point cloud-based processing used by Mahler et al. (2019) in Dex-Net 4.0, which directly handles 3D point cloud data to better preserve geometric information.

4 Evolution of crawling strategy optimization techniques

After obtaining initial grasping predictions, another key research focus lies in how to further optimize strategies to adapt to dynamic environments or enhance performance through interactive learning.

Imitation Learning: Training grasping strategies under supervision using demonstration data (e.g., human teaching). Although the large-scale distributed reinforcement learning project by Levine et al. (2016) conducted a large-scale distributed reinforcement learning project. While fundamentally RL, it collected extensive data from human remote-controlled operations, demonstrating the importance of big data in grasping learning.

Deep Reinforcement Learning (DRL) enables robots to autonomously learn optimal strategies through trial and error interactions with the environment, guided by acquired rewards (e. g., successful grasps). Q-learning, DQN, and their variants are widely employed for such tasks. Research from OpenAI and Google Brain demonstrates DRL's potential for learning complex dexterous hand grasping operations in simulated environments. However, DRL typically requires substantial interaction data, making training on physical robots prohibitively expensive. Consequently, sim-to-real transfer techniques have become critical.

Currently, purely data-driven approaches still exhibit shortcomings in physical safety. A notable trend in recent years has been the integration of physics-based analytical methods (e.g., force enclosure analysis) with data-driven approaches. This hybrid methodology leverages learning techniques to compensate for model uncertainties, enabling the generation of grasping strategies that are both efficient and physically sound.

5 Current issues

Despite significant progress, the field still faces numerous challenges:

(1) Data scarcity and generalization capability: Deep networks require vast amounts of labeled data, while collecting and processing data from real robots is time-consuming and labor-intensive. Current models still struggle with generalizing to unfamiliar objects or extreme lighting and occlusion conditions outside the training set distribution.

(2) Sim-to-real gap: Strategies trained in near-perfect simulators suffer drastic performance degradation when transferred to the noisy, delayed, and uncertain real world. While techniques like domain randomization show some effectiveness, they have yet to fully resolve this issue.

(3) Dynamic and non-rigid object grasping: Existing research predominantly focuses on static, rigid objects. Research on grasping dynamic targets (e.g., objects on conveyor belts) and non-rigid objects (e.g., cables, soft packages) remains in its infancy.

(4) Multimodal perception and interaction: Relying solely on visual information often fails to yield optimal decisions, as it cannot perceive an object's weight or surface smoothness. Therefore, effectively integrating multi-sensory information like tactile and auditory data, while employing active interaction methods (e.g., pushing or touching objects) to enhance environmental understanding, has become a critical research direction for improving grasping system robustness.

(5) Real-time Performance and Computational Efficiency: Complex deep learning models demand substantial computational resources, posing real-time challenges when deployed on embedded robotic platforms. Striking a balance between algorithmic accuracy and computational efficiency is essential.

6 Conclusion

Deep learning has fundamentally transformed industrial robot grasping technology, propelling it from traditional reliance on manual modeling toward data-driven automated learning. Currently, grasp detection methods based on convolutional neural networks (CNNs) have matured significantly in structured environments, approaching practical application levels. Meanwhile, deep reinforcement learning (DRL) points toward future development directions by enabling robots to autonomously learn adaptable grasping strategies in more complex and dynamic environments.

Future research will increasingly focus on small-sample and even zero-sample learning to enhance systems' generalization capabilities for unknown objects. Concurrent efforts will center on developing more effective “simulation-to-reality” (sim-to-real) transfer techniques, exploring multimodal information fusion and active environmental perception, and tackling challenges in grasping dynamic and non-rigid objects. The ultimate goal is to build a general-purpose intelligent grasping system that operates like humans—observing with eyes, deciding with the brain, and executing with the arm—representing the long-term pursuit of this field.

References

[1] Jiang, Y. J. Y, Moseson, S., & Saxena, A. (2011). Efficient grasping from RGBD images: Learning using a new rectangle representation. IEEE. IEEE.

[2] Lenz, I, Lee, H, & Saxena, A. (2013). Deep learning for detecting robotic grasps. The International Journal of Robotics Research, 34(4-5).

[3] Redmon, J. & Angelova, A. (2014). Real-time grasp detection using convolutional neural networks. Proceedings IEEE International Conference on Robotics & Automation.

[4] Ronneberger, O, Fischer, P, & Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. Springer, Cham.

[5] Morrison, D, Corke, P, & Leitner, Jürgen. (2018). Closing the loop for robotic grasping: a real-time, generative grasp synthesis approach.

[6] Xiang, Y, Schmidt, T, Narayanan, V, & Fox, D. (2017). Posecnn: a convolutional neural network for 6d object pose estimation in cluttered scenes.

[7] Mahler, J, Matl, M, Satish, V. , Danielczuk, M, Derose, B, & Mckinley, S, et al. (2019). Learning ambidextrous robot grasping policies. Science Robotics, 4(26).

[8] Levine, S, Pastor, P, Krizhevsky, A, Ibarz, J, & Quillen, D. (2016). Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. International Journal of Robotics Research, 37(4-5), 421-436.

[9] Andrychowicz, M, Baker, B, Chociej, M. Jozefowicz, R, Mcgrew, B. , & Pachocki, J. et al. (2018). Learning dexterous in-hand manipulation.

[10] Tobin, J, Fong, R, Ray, A, Schneider, J. & Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. IEEE.

[11] 2016. Robot grasping in clutter: using a hierarchy of supervisors for learning from demonstrations. IEEE.

[12] Viereck, U, Pas, A. T, Saenko, K, & Platt, R. (2017). Learning a visuomotor controller for real world robotic grasping using simulated depth images.

[13] Ten Pas, A, Gualtieri, M, Saenko, K, & Platt, R. (2017). Grasp pose detection in point clouds. The International Journal of Robotics Research.

[14] Quillen, D, Jang, E. , Nachum, O, Finn, C, Ibarz, J, & Levine, S. (2018). Deep reinforcement learning for vision-based robotic grasping: a simulated comparative evaluation of off-policy methods. IEEE.

[15] Caldera, S, Rassau, A. , & Chai, D. (2018). Review of Deep Learning Methods in Robotic Grasp Detection. Preprints.