Application of Reinforcement Learning in Robotic CNC Path Optimization

Hanqi Sun
North China University of Technology, Beijing, China
ReeceSun705@gmail.com

ABSTRACT

With the widespread application of industrial robots in the field of high flexibility CNC machining, how to achieve path optimization under complex dynamic constraints has become the key to improving machining performance. Traditional heuristic algorithms and mathematical programming methods often face limitations such as low computational efficiency and strong model dependence when dealing with high-dimensional configuration spaces and nonlinear dynamic features. This article systematically reviews the research progress of reinforcement learning (RL) in robot numerical control path optimization. Firstly, the modeling method of transforming path optimization problems into reinforcement learning Markov decision processes (MDPs) was elaborated, covering state space design, continuous action mapping, and construction of multi-objective reward functions. Secondly, the application performance of deep reinforcement learning algorithms in core scenarios such as feed rate scheduling, multi axis trajectory smoothing, and machining error compensation was analyzed in detail. The research results of this article provide important theoretical support and technical reference for the development of a new generation of intelligent and autonomous robot numerical control systems.

Keywords: Robot CNC Machining; Path Optimization; Reinforcement Learning; Intelligent Manufacturing; Deep Learning.

1. INTRODUCTION

In modern-day industrial manufacturing, the technology of robot numerical control (CNC) machining is playing an increasingly important role. Compared to traditional large machine tools, industrial robots exhibit significant flexibility advantages in handling complex aviation components and large mold processing due to their larger working range and higher degree of freedom in motion. In actual production, how to plan high-quality machining paths remains the core challenge that determines the overall performance of robot CNC systems. The quality of path optimization is directly related to three key dimensions: production efficiency, processing accuracy, and resource energy consumption. First of all, rational path planning can substantially shorten the processing cycle, and as a result, improve the overall output of the production line. Secondly, taking into account the relatively poor rigidity of industrial robots, optimizing the path can effectively curtail the vibration during operation and ensure that the dimensional accuracy of parts meets the strict design requirements. Finally, smooth movement paths and scientific speed control can reduce the frequent on-off of motors during the machining, thereby greatly reducing energy consumption. So, in-depth study on how to reach multi-objective path collaborative optimization in high dynamic environments has important practical worth for promoting the extensive application of intelligent manufacturing technology [1].

Researchers have developed various classic algorithms for path optimization problems in the past few decades, mainly focusing on heuristic search and mathematical programming. In order to surmount the weaknesses of traditional methods, reinforcement learning (RL), an up-and-coming intelligent decision-making tool, has begun to be broadly applied in the domain of robot path planning. The central advantage of reinforcement learning lies in the fact that it doesn't rely on absolute modeling of the processing environment. Conversely, it makes use of a "try and feedback" mechanism so that robots can automatically find the best-fitting strategy during their interaction with the environment [2][3].

This article will conduct an in-depth study on the decision logic of reinforcement learning for dealing with complex machining constraints and examine how it solves high-dimensional nonlinear path optimization problems by means of learning mechanisms. By looking over the existing research achievements, this article will bring to light the significant challenges of current algorithms with respect to convergence speed and the "simulation to reality" transfer. These

summaries are going to give clear technical references and ways to improve for the future development of more autonomous and efficient robot numerical control systems.

2. REINFORCEMENT LEARNING ALGORITHMS FOR ROBOT PATH PLANNING AND CONTROL

2.1 Value based approach: The application of Q-Learning and DQN in discrete path point selection

Traditional deep Q-networks are almost incapable of effective learning in sparse reward environments, as most exploration trajectories cannot generate any reward signals under the condition that robots only receive positive rewards when they reach their targets. To address this issue, the Hindsight experience replay mechanism proposed by Andrychowicz et al. enhances the data of each failed attempt, making the originally sparse rewards denser [4]. The core approach is to use the actual location reached as a substitute target to recalculate the reward. Even if the robot fails to reach the preset target point, as long as it reaches some other location, it "considers" that location as the target, thus giving meaningful feedback signals to the trajectory that originally had zero reward. In the 20 × 20 grid navigation task, DQN using HER can successfully learn the strategy to reach the target point in the same number of training rounds, while DQN without HER cannot converge at all; Training monitoring further indicates that the evaluation statistic of HER-DQN ultimately exceeds 1.0, indicating that the agent can stably and successfully reach the target position. The main advantage of HER is that it demonstrates that DQN can overcome sparse reward problems through empirical augmentation methods, significantly improving sample efficiency. This is particularly crucial for real robot learning, as not every attempt in real-world environments can be successful. However, this method also has clear limitations: firstly, HER relies on a key assumption that the target space must be relabeled afterwards, which means that the target values can be replaced without violating environmental dynamics, which is not true in many tasks; Secondly, this method is still limited by the inherent requirements of DQN for discrete action spaces and cannot be directly applied to continuous control scenarios.

The improvement proposed by Zheng Chenwei et al. includes two aspects: in terms of exploration strategy, this paper adopts the Fast Extended Random Tree (RRT) algorithm to automatically generate static prior knowledge, which is used to guide the action selection process and replace the traditional random action selection method. In terms of experience utilization, an experience replay mechanism based on K-means clustering is introduced, which clusters experience samples based on the position information of dynamic obstacles and prioritizes sampling experience clusters that are similar to the current state [5].

In the two-dimensional gridded dynamic environment simulation, the improved algorithm can effectively avoid static and dynamic obstacles and successfully reach the target point. In contrast, traditional DQNs frequently collide with obstacles in the same environment.

The advantage of this improved method is that it proves that DQN can compensate for the shortcomings of pure exploration strategies by introducing planning priors. The global guidance capability provided by RRT complements the learning capability of DQN itself, forming a typical example of the combination of planning and learning.

However, this method also has certain limitations. Firstly, its performance depends on the quality of prior knowledge generated by RRT. Secondly, the clustering mechanism requires prior knowledge of the dimensions of dynamic obstacle location information, and when the environment changes, relevant parameters need to be redesigned. Overall, this method is essentially an improvement on the engineering level of DQN and has not yet achieved a theoretical breakthrough.

2.2 Policy Gradient based method

Before TRPO, the fatal problem with the gradient method was that improper selection of gradient step size could lead to performance collapse after the strategy made a "one-step mistake". The core insight of Schulman et al. is that, unlike directly limiting the amount of parameter change, a more effective approach is to directly constrain the magnitude of the policy's own changes and quantify this change using KL divergence. Based on this idea, TRPO solves an optimization problem with KL divergence constraints to ensure that the KL divergence between the new and old strategies after each update does not exceed a preset threshold, thereby theoretically ensuring the monotonic improvement of the strategy. TRPO showed that it was much more training stable than conventional strategy gradient methods on the MuJoCo continuous control task. Moreover, in the practical control scenarios of cable driven parallel robots, TRPO compared favorably with other approaches with the lowest RMS errors for trajectory tracking in circular and spiral trajectories. For instance, the errors on circular trajectories were 0.0075 for the DDPG, 0.0098 for the PPO, and 0.0113 for the error on the circular trajectory. More interestingly, one of the advantages of TRPO is that it demonstrates that the constraint strategy

(as opposed to the constraint parameters) is crucial to ensure stable updates. This theoretical insight provides a theoretical basis for subsequent algorithms, like PPO. TRPO is also shown to be robust to changes in control update time intervals, which is advantageous in practical robot applications, where the sensor noise and control delay are prominent. But there are downsides to TRPO, too, that cannot be overlooked. This approach involves the computation of Fisher information matrix and its inverse matrix, which has a complexity of O (N ³) (N is the parameter dimension), therefore, it is hard for TRPO to be applied in large-scale neural networks. In the meantime, all of TRPO's strategy algorithm is the same, and the empirical data gathered after each update is completely invalid and cannot be used again, which means there is a relatively low sample efficiency. Overall, the stability advantage demonstrated by TRPO and PPO in continuous speed planning tasks is due to their mechanism of effectively controlling the magnitude of policy updates [6].

2.3 Depth Deterministic Policy Gradient

There are a number of mainstream algorithms to deal with continuous action spaces, such as Deep Deterministic Policy Gradient (DDPG) and Flexible Action Evaluation (SAC). DDPG is an adaptation of DQN's experience replay and target network mechanism to the deterministic policy gradient framework, which directly outputs a deterministic action through an actor network, and computes the corresponding Q-value through a critic network. SAC, on the other hand, incorporates entropy regularisation on top of a maximisation of the cumulative rewards, which leads to maximisation of policy entropy and better exploratory behavior. Both of these algorithms excel when dealing with high dimensional continuous action spaces in multi axis linkage robotics, and thus are classic options for continuous control problems like robotic arms. The core contribution of Lillicrap et al. is the proof of the "Deterministic Policy Gradient Theorem"-which allows for the direct learning of a deterministic function.

The core innovation of DDPG stems from the fundamental difficulties encountered by DQN in the continuous action space, as the success of DQN relies on calculating the maximum value for each action, which cannot be achieved in the continuous action space (such as the torque output of each joint of a robotic arm). The key contribution of Lillicrap et al is the proof of the deterministic policy gradient theorem, which indicates that the algorithm can directly learn a deterministic mapping function $a=\mu(s)$ that transforms a state directly into an action, and its gradient can be obtained by passing the gradient of the action through the Q-function. Specifically, DDPG relies on four key technologies: firstly, the actor critic architecture composed of actors and critics; Secondly, using deterministic strategies to simplify the sampling process; Thirdly, utilizing the experience replay mechanism to break the temporal correlation between data; Fourth, introduce a soft target update strategy to improve the stability of the training process [7].

In the six axis robotic arm control task on the MuJoCo platform, DDPG demonstrated better stability compared to SAC and TD3. In the tomato skewer collection task of agricultural robots, an improved version DDPG-Z that integrates human teaching data achieved a 25% increase in convergence speed, a 10.3% increase in reward value after convergence, and a 51.3% increase in target point accuracy. In contrast, standard DDPG, SAC, and TD3 failed to converge within the specified turn of the task.

The main advantage of DDPG is that it demonstrates that the heterogeneous strategy learning method can be successfully extended to continuous action spaces, providing a practical and relatively stable baseline algorithm for robot control tasks. Due to its natural ability to learn from different strategies, DDPG can easily learn from human teaching data, making it highly valuable in fields such as agriculture and healthcare that require imitation of expert operations. However, this method also has obvious limitations. Firstly, DDPG is highly sensitive to the selection of hyperparameters, and the training process is prone to instability. Secondly, the overestimation problem of Q function may lead to the degradation of policy performance, and the TD3 algorithm proposed later is an improvement specifically aimed at this problem. Finally, the deterministic strategy adopted by DDPG is essentially a greedy strategy, so its ability to balance exploration and utilization is naturally inferior to random strategies [8].

3. CHALLENGES AND FUTURE TRENDS

When exploring the application of reinforcement learning in the field of robot numerical control path optimization, researchers still face multiple severe challenges. First of all, the challenge of applying simulation to reality is still prominent. Since physical simulators are unable to fully mimic the complex friction, thermal deformation, and cutting force fluctuations in real machine tools. The effectiveness of the strategy is directly affected by the existence of this "reality gap". Secondly, in CNC machining, safety and robustness are of great importance. The early random exploration mechanism of reinforcement learning can easily lead to robot collisions or exceeding limits. How to introduce Constrained Reinforcement Learning (RL) while ensuring equipment safety is currently a technical challenge. In addition, existing

reinforcement learning models are often trained for specific tasks, and their lack of generalization makes it difficult to directly reuse them in different specifications of machine tools or diverse processing conditions. Besides, in the context of large-scale computing, the convergence speed of the model still calls for further optimization.

In response to the above challenges, future research trends will focus on the deep integration of intelligence and modularity. On the one hand, hybrid optimization algorithms will take the lead in the field. They integrate reinforcement learning and model predictive control (MPC) or traditional PID control. Not only can they make use of the nonlinear processing power of RL, but also ensure the system's basic stability and certainty by means of traditional control algorithms. Alternatively, the application of large models and transfer learning is going to significantly enhance the system's ability to generalize. By making use of pre-training methods, the model can rapidly get used to different processing tasks. Meanwhile, as perception technology makes headway, real-time closed-loop optimization with multi-sensor feedback integration will further close the gap between simulation and reality. It will also push CNC systems to move in the direction of a safer, more efficient, and self-evolving intelligent manufacturing mode.

4. CONCLUSION

The article makes a systematic review of the current application scenario of reinforcement learning in the path optimization of robot numerical control (CNC). Through the analysis of multiple typical application cases, it can be seen that reinforcement learning has shown significant advantages in handling high-dimensional configuration spaces and nonlinear dynamic constraints. The results of the research imply that making use of deep reinforcement learning algorithms like PPO and SAC to implement feed rate scheduling and trajectory smoothing strategies can effectively reduce the machining cycle by 10%-20% while ensuring the accuracy of machining, and remarkably lower the energy consumption due to frequent motor acceleration and deceleration. Meanwhile, by making use of the adaptive features of reinforcement learning, the system is able to compensate for the rigidity deficiency and vibration issues in the robot machining process in real-time without depending on high-precision physical models. These cases fully demonstrate the feasibility and excellent performance of data-driven intelligent solutions in complex machining tasks.

The research work presented in this review is of important theoretical and practical value. At the theoretical level, this paper describes the technological background of the move from traditional optimization algorithms to intelligent decision-making mechanisms, identifies the mapping relationship of core elements of reinforcement learning in CNC path optimization, and constructs a systematic theoretical framework for subsequent related investigations. At the practical stage, the algorithm comparison and application evaluation presented in this article give technical reference for the development of a new generation of flexible and intelligent robot processing units in the industry. Not only does this help to increase the manufacturing efficiency of domestically produced high-end equipment, but it also offers a practical and achievable technological path for reaching the digital transformation and green low-carbon targets of the manufacturing industry.

With the continuous optimization of algorithm structure and the improvement of computing power, reinforcement learning will inevitably become the core engine of future intelligent numerical control systems, promoting robot machining to leap towards a highly intelligent stage of autonomous perception, decision-making, and execution.

REFERENCES

[1] Zhang Y, Liu Y, Wang H, Et Al. Multi-Objective Trajectory Optimization Method for Industrial Robots Based on Improved Td3 Algorithm. Scientific Reports, 2025, 15: 25949.
[2] Wang Z, Cao J, Cao Y, Et Al. Energy-Efficient Trajectory Planning for A Class of Industrial Robots Using Parallel Deep Reinforcement Learning. Nonlinear Dynamics, 2025, 113: 8491-8510.
[3] Li Y, Zhou L, Mo Y, Et Al. Energy-Efficient Tool Path Generation and Expansion Optimisation for Five-Axis Flank Milling with Meta-Reinforcement Learning. Journal of Intelligent Manufacturing, 2024, 36: 1-18.
[4] Andrychowicz M, Wolski F, Ray A, Et Al. Hindsight Experience Replay. Advances in Neural Information Processing Systems, 2017, 30.
[5] Zheng Chenwei, Hou Lingyan, Wang Chao, et al. Dynamic Obstacle Avoidance Path Planning Based on Improved DQN. Journal of Beijing Information Science and Technology University (Natural Science Edition), 2024, 39(5).
[6] Schulman J, Levine S, Abbeel P, Et Al. Trust Region Policy Optimization//International Conference on Machine Learning. PMLR, 2015: 1889-1897.

[7] Lillicrap T P, Hunt J J, Pritzel A, Et Al. Continuous Control with Deep Reinforcement Learning. Arxiv Preprint Arxiv:1509.02971, 2015.
[8] Zhang, Y.; Li, Y.; Feng, Q.; Sun, J.; Peng, C.; Gao, L.; Chen, L. Compliant Motion Planning Integrating Human Skill for Robotic Arm Collecting Tomato Bunch Based On Improved DDPG. Plants 2025, 14, 634.