## ABSTRACT

Challenges arise in autonomous control of modern production facilities. Automated gantry robots system requires efficient scheduling of transportation tasks, while a stable behavior under the pressure of stochastic condition, such as machine failure, is not to be compromised. Recent advancements in Reinforcement Learning (RL) provide promising approaches to meet these requirements.

Previous work within the project focused on the Deep Q-Network (DQN) algorithm using the discount reward criterion. This thesis investigates algorithms based on policy gradient methods. The investigated algorithms utilize the average reward criterion in Multi-Agent Reinforcement Learning (MARL) frame. The average reward objective aligns naturally with the project frame, where the long-term throughput is to be maximized.

The following policy gradient methods are investigated and adapted to the project environment:

- Average Reward Trust Region Policy Optimization (TRPO)
- Proximal Policy Optimization (PPO)
- Heterogeneous-Agent Trust Region Policy Optimisation (HATPRO)
- Heterogeneous-Agent Proximal Policy Optimisation (HAPPO)

Modifications are required for the adaption, because the original algorithms are developed in discrete-time framework but the project has continuous-time frame. Time-based temporal-difference (TD) and time-discounted Generalized Advantage Estimation (GAE) are proposed to mitigate the discrepancy. HATPRO and HAPPO also require synchronous decision-making between the agents. This is approximated by a proposed modification in the state space.

In the experiment with smaller environments, the adapted algorithms demonstrate performance close to the Multi-Agent DQN from previous works. However, in larger environments, the adapted algorithms are still outperformed by Multi-Agent DQN. The experiment result is analyzed, and potential causes as well as possible improvements are discussed.

# ZUSAMMENFASSUNG

Herausforderungen entstehen bei der autonomen Steuerung moderner Produktionsanlagen. Automatisierte Portalrobotersysteme erfordern eine effiziente Planung von Transportaufgaben, wobei ein stabiles Verhalten unter stochastischen Bedingungen, wie beispielsweise Maschinenstörungen, gehalten werden muss. Jüngste Entwicklungen im Bereich des Reinforcement Learning (RL) bieten versprechende Ansätze, um diese Anforderungen zu erfüllen.

Frühere Arbeiten innerhalb des Projekts konzentrierten sich auf den Deep-Q-Network-Algorithmus (DQN) unter Verwendung eines diskontierten Belohnungskriteriums. Diese Arbeit untersucht dagegen Algorithmen auf Basis von Policy-Gradient-Methoden. Die umgesetzte Algorithmen verwenden das Average-Reward-Kriterium im Rahmen von Multi-Agent Reinforcement Learning (MARL). Dieses Kriterium entspricht der Zielsetzung des Projekts, bei der der langfristige Durchsatz des Produktionssystems maximiert werden soll.

Die folgenden Policy-Gradient-Methoden werden untersucht und an die Projektumgebung angepasst:

- Average Reward Trust Region Policy Optimization (TRPO)
- Proximal Policy Optimization (PPO)
- Heterogeneous-Agent Trust Region Policy Optimization (HATRPO)
- Heterogeneous-Agent Proximal Policy Optimization (HAPPO)

Für die Anpassung sind Modifikationen erforderlich, da die ursprünglichen Algorithmen für diskrete Zeitschritte entwickelt wurden, während die Projektumgebung einen kontinuierlichen Zeitrahmen hat. Zur Überbrückung dieser Diskrepanz werden ein zeitbasiertes Temporal-Difference (TD) sowie eine zeitdiskontierte Generalized Advantage Estimation (GAE) vorgeschlagen. Darüber hinaus setzen HATRPO und HAPPO synchrone Entscheidungsprozesse zwischen den Agenten voraus. Dies wird durch eine vorgeschlagene Erweiterung des Zustandsraums approximiert.

In Experimenten mit kleineren Umgebungen erreichen die angepassten Algorithmen eine vergleichbar Leistung zum Multi-Agent-DQN von früheren Arbeiten. In größeren Umgebungen werden die angepasset Algorithmen jedoch weiterhin von Multi-Agent-DQN übertroffen. Die experimentellen Ergebnisse werden analysiert, und mögliche Ursachen sowie potenzielle Verbesserungen werden diskutiert.