JAIST Repository
https://dspace.jaist.ac.jp/

Title	人間ーロボット相互作用のための安全なAI 駆動型視覚把持
Author(s)	LI, CHENGHAO
Citation	
Issue Date	2026-03
Type	Thesis or Dissertation
Text version	ETD
URL	https://hdl.handle.net/10119/20580
Rights	
Description	Supervisor: 丁 洛榮, 先端科学技術研究科, 博士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

氏　　名	LI Chenghao		
学位の種類	博士（情報科学）		
学位記番号	博情第 567 号		
学位授与年月日	令和 8 年 3 月 25 日		
論文題目	SAFE AI-POWERED VISUAL GRASPING FOR HUMAN-ROBOT INTERACTION		
論文審査委員	CHONG Nak Young	JAIST	Professor
	BEURAN Razvan	JAIST	Assoc. Prof.
	JI Yonghoon	JAIST	Assoc. Prof.
	UMEDA Kazunori	Chuo University	Professor
	ELIBOL Armagan	Al Ain University	Assoc. Prof.

論文の内容の要旨

Integrating the Artificial Intelligence (AI) vision module into the robot grasping system can significantly improve its generalizability, thereby enhancing the efficiency of Human-Robot Interaction (HRI). However, the inherent lack of interpretability in AI also opens the door to external threats, like backdoor attacks. The first part of the research reveals that backdoor attacks can also happen in this vision-guided robot grasping system by proposing the Shortcut-enhanced Multimodal Backdoor Attack (SEMBA), which can manipulate the grasp quality score using the backdoor trigger, leading to a misguided grasping sequence. The SEMBA may thus cause potentially hazardous grasping and pose a threat to human safety in HRI. Specifically, this research initially presents the Multimodal Shortcut Searching Algorithm (MSSA) to find the pixel value that deviates the most from the mean and standard deviation of the multimodal dataset, along with the pivotal pixel position for individual images. This will guarantee that the proposed attack is effective in complex, multi-class object scenarios. Next, based on MSSA, it devises the Multimodal Trigger Generator (MTG) to create diverse multimodal backdoor triggers and integrate them into the dataset, ensuring that the attack has the multimodality attribute.

In a cluttered HRI scenario, if a user employs a grasping model compromised by the SEMBA attack, and an object resembling the backdoor trigger appears in the scene. Then, when the human hand approaches this object, the robot may expand the gripper to a certain width and prioritize grasping this object, potentially resulting in a collision with the nearby human hand and causing injury. Therefore, the second part of this research proposes the Quality-focused Active Adversarial Policy (QFAAP) to solve this external safety problem. Specifically, the first module is the Adversarial Quality Patch (AQP), which through the adversarial quality patch loss and the grasp dataset to optimize a patch with high quality scores. Next, the Projected Quality Gradient Descent (PQGD) is constructed and is integrated with the AQP, which contains only the hand region within each real-time frame, endowing the AQP with fast adaptability to the human hand shape. Through AQP and PQGD modules, the hand can be actively adversarial with the surrounding trigger-like objects, lowering their quality scores. Therefore, further setting the quality score of the hand to zero

will reduce the grasping priority of both the hand and its nearby trigger-like objects, enabling the robot to grasp other objects away from the human hand.

Although QFAAP can mitigate the safety risks posed by SEMBA-based trigger-like objects adjacent to the human hand in cluttered HRI scenarios, AI-powered visual grasping systems that rely on a fixed camera view often suffer from incomplete object geometry near the view boundaries. Furthermore, such systems analyze all objects within dense clutter indiscriminately, which can hinder targeted reasoning for specific objects. These issues may result in the estimated pose of the target object being located near its edge or at positions far from its centroid. Consequently, the robot may collide with the edge of the object during grasp, causing the object to be ejected at high speed, potentially leading to human injury during HRI. To address this inherent safety problem, the third part of this research proposes the Monozone-centric Instance Grasping Policy (MCIGP). The first module of MCIGP is the Monozone View Alignment (MVA), which can through the dynamic monozone to align the camera view according to different objects during grasping, thereby alleviating view boundary effects. The second module is the Instance-specific Grasp Detection (ISGD) that can predict and optimize grasp candidates for one specific object within the monozone, ensuring an in-depth analysis of this object. Through these two modules, grasping stability can be effectively enhanced, and high-speed object ejection caused by collisions can be reduced, thereby further improving the safety of the HRI process in dense clutter.

The effectiveness of the three parts of this research is validated through extensive experiments, including experiments on different benchmark datasets and real-world grasping experiments on a collaborative robot.

Keywords: Robot Grasping, Robot Safety, AI Security, Human-Robot Interaction, Machine Leaning

論文審査の結果の要旨

Even though AI has emerged as a transformative tool capable of reshaping various sectors of robotics, the connection between AI safety and existential risk from conventional rigid-bodied robot systems is an underexplored problem. This dissertation aimed to present a comprehensive exploration into the safety of the AI-powered robot visual grasping system in cluttered human-robot interaction scenarios. Integrating the AI-enabled vision module into the grasping system can significantly improve its generalizability, thereby enhancing the efficiency of Human-Robot Collaboration (HRC) in smart manufacturing environments. However, the inherent lack of interpretability in AI also opens the door to external threats. For instance, backdoor attacks can manipulate the quality score through the trigger to control the grasping sequence, causing potentially hazardous grasping during HRC. This work systematically investigated both external and inherent risks in the vision and deep learning-based grasping system and proposed novel methods to enhance grasping safety through different policy designs.

The contribution of this dissertation lies in the following aspects. Firstly, this study introduced the Shortcut-Enhanced Multimodal Backdoor Attack (SEMBA), which integrates multimodal information and shortcut learning to uncover previously unexplored vulnerabilities in AI-powered robot visual grasping systems. Secondly, to counteract the external SEMBA threat, this work developed the Quality-Focused Active Adversarial Policy

(QFAAP), which allows the human hand itself to serve as an active perturbation source to suppress grasp quality score near potential trigger-like regions. Finally, to overcome the inherent safety issues caused by limited grasping view and inferior grasp candidates, this work proposed the Monozone-Centric Instance Grasping Policy (MCIGP), which integrates Monozone View Analysis (MVA) and Instance-Specific Grasp Detection (ISGD) for precise and reliable grasping within dynamically detected monozones. The effectiveness of the aforementioned three parts of this research is validated through extensive experiments, including experiments on different benchmark datasets and real-world grasping experiments on a collaborative robot. In summary, this dissertation advances understanding and implementation of secure and intelligent visual grasping. Furthermore, the proposed methods collectively contribute to building a safe and adaptive human-aware robotic grasping framework, paving the way toward trustworthy Human-Robot Interaction (HRI) in complex, real-world environments.

This is an excellent dissertation, and we approve awarding a doctoral degree to Mr. LI Chenghao.