Multimodal Language Processing for Everyday Task Execution by Service Robots

September 2025

Motonari Kambara

別表５
⑶

主　論　文　要　旨

No.1

報告番号	㊙甲 乙 第　　号	氏　名	神原 元就

主 論 文 題 名 :

Multimodal Language Processing for Everyday Task Execution for Service Robots
(サービスロボットによる日常タスク実行のためのマルチモーダル言語処理)

（内容の要旨）

本論文は，高齢化の進行と介護ニーズの増大を背景に，人と協調して日常生活を支援するサービスロボットの性能向上を目指す．音声指示はユーザにとって極めて直感的なインタフェースである一方，現行ロボットは自由形式の言語指示を解釈して現実世界のタスクを遂行する能力がまだ限定的である．本研究はこの性能ギャップを，①ロボティクス分野における大規模で指示情報が豊富なコーパスの欠如（データセットギャップ），②予測された行動の物理的実行可能性や安全性が十分に考慮されないこと（実行ギャップ）の二つに整理する．これらを克服するため，本論文ではマルチモーダル言語モデルを用いてサービスロボットが日常タスクを自律的に遂行するシステムを構築した．具体的には，既存データセットを拡充するために物体操作指示文をクロスモーダルに生成する手法を備えた全自動タスク実行フレームワークを提案する．さらに，安全で効率的な動作を保証するため，実行前にタスクの成否を予測し，衝突などの潜在的事象を自然言語で説明するモデルを統合した．標準評価指標に基づく実験の結果，提案手法は各種ベースラインを一貫して上回る性能を示し，本フレームワークの有効性を確認した．

本論文は全6章で構成される．第1章では，研究の背景，目的，および論文構成を示している．第2章では，物体操作指示文の自動生成のためのクロスモーダル自然言語生成モデルを提案し，Case Relation Block が複数物体を考慮するうえで重要な役割を果たすことを実験的に示す．第3章では，マルチモーダル言語理解・生成モデルに基づく自動タスク実行フレームワークを提案し，人間の介入なしにタスクを完遂できることを確認する．第4章では，将来の衝突などのイベントを予測し，それを説明するマルチモーダル言語生成モデルを提案し，適切な予測と説明が可能であることを示す．第5章では，物体操作前のマルチモーダルアラインメントを判定する手法を提案し，ロボットのポリシーに依存せずタスクの成否を高精度で予測できることを検証する．最後に第6章で本研究の総括と今後の展望を述べる．

別表６
（３）

Keio University

Thesis Abstract

No.

Registration Number	☑"KOU" □ "OTSU" No. *Office use only	Name	Motonari Kambara

Thesis Title

Multimodal Language Processing for Everyday Task Execution by Service Robots

Thesis Summary

This thesis seeks to enhance the capabilities of service robots that assist people in their daily lives, motivated by the rapid aging of society and the resulting increase in caregiving needs. Although spoken commands offer users a highly intuitive interface, current robots still struggle to interpret free-form language instructions and execute real-world tasks. We attribute this performance gap to two factors: (i) the dataset gap, namely the lack of large-scale, instruction-rich corpora for robotics, and (ii) the execution gap, meaning that models often overlook the physical feasibility and safety of their predicted actions. To address these issues, we design a system in which a multimodal language model enables service robots to perform everyday tasks autonomously. Specifically, we propose a fully automated task-execution framework that automatically augments existing datasets by cross-modally generating object-manipulation instructions. To guarantee safe and efficient operation, the framework also integrates a model that predicts task success or failure before execution and produces natural-language explanations of potential hazards such as collisions. Experiments conducted with standard evaluation metrics show that the proposed methods consistently outperform various baselines, confirming the effectiveness of the framework.

The thesis is organized into six chapters. Chapter 1 outlines the research background, objectives, and overall structure. Chapter 2 introduces a cross-modal natural-language generation model for automatically producing object-manipulation instructions and experimentally demonstrates that the Case Relation Block plays a critical role in reasoning about multiple objects. Chapter 3 presents an automated task-execution framework based on multimodal language understanding and generation, verifying that tasks can be completed without human intervention. Chapter 4 proposes a multimodal language-generation model that predicts future events such as collisions and explains them in natural language, showing that accurate prediction and explanation are attainable. Chapter 5 describes a method for judging multimodal alignment before object manipulation and confirms that task success can be predicted with high accuracy regardless of the robot's policy. Finally, Chapter 6 summarizes the findings and discusses directions for future work.