マルチモーダル情報に基づく移動指示理解モデルの構築

2024 年度

細見 直希

別表５
⑶

主　論　文　要　旨

No.1

報告番号	㊙甲 乙　第　　号	氏　名	細見 直希

主 論 文 題 名：

マルチモーダル情報に基づく移動指示理解モデルの構築

（内容の要旨）

本論文は，人々の自由で便利な移動を支援するエージェント（モビリティ）の基盤となる，視覚と言語を用いたマルチモーダル情報に基づく移動指示理解モデルの構築に取り組むものである．具体的には，人間が自律移動可能なモビリティに対して「黒いトラックの向かい側に停まってください」という移動指示を与えた場合に，モビリティが指示内容を適切に解釈し，黒いトラックを目印として，その向かい側の道路上における適切な目標位置や領域を予測可能なモデルの構築を目的としている．

本論文では，実世界の移動タスクに特有の問題である，⑴画像と言語のような異なるモダリティ間で情報の性質に違いがあるうえに，情報の信頼性が環境条件によって変動し得る点，⑵目標位置や領域は道路上の任意の地点に存在する可能性があり，予測のための明確な視覚的境界が存在しない点，に対処するマルチモーダル情報に基づく目標位置または領域の予測に関する手法を提案し，実験を通じてその有効性を検証している．

本論文は，全６章により構成される．

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，言語理解および記号接地，視覚と言語を中心としたマルチモーダル情報の統合技術に関する先行する関連研究を整理し，本研究の位置づけを述べている．

第 3 章では，本論文で用いるマルチモーダル情報の統合的処理の基盤技術について詳述し，提案手法の背景にある原理および理論的根拠を示している．

第４章では，モビリティの目標領域を予測する Referring Navigable Region タスクを対象とし，Trimodal Navigable Region Segmentation Model (TNRSM)を提案している．実験を通じて，TNRSM がモダリティ間の情報の性質の違いを補完し，さらに，視覚に関する情報の信頼性が環境条件によって変動する問題に対して効果的に対処できることを明らかにしている．

第５章では，モビリティの目標位置を予測する Vision-and-Language Target Positioning タスクを対象とし，Target Regressor in Positioning (TRiP)を提案している．実験を通じて，TRiP が目標位置の予測において明確な視覚的境界が存在しない問題に効果的に対処できることを明らかにしている．

第６章では，本論文の総括を行い，今後の展望について述べている．

別表６
（３）

Keio University

Thesis Abstract

No.

Registration Number	☑ “KOU” □ “OTSU” No. *Office use only	Name	Naoki Hosomi

Thesis Title

Development of a Multimodal Model for Understanding Navigation Instructions

Thesis Summary

This thesis aims to develop a multimodal model that leverages vision and language to understand navigation instructions, serving as a foundation for autonomous agents (mobilities) with the purpose of supporting human free movement. Specifically, given the instruction "Park right across from the black truck," the model should predict an appropriate parking position or region on the road across from the black truck.

This thesis proposes and validates methods for predicting target positions or target regions based on multimodal information, addressing the problems specific to real-world navigation tasks: (1) differences in the type of information across different modalities such as images and language, and the reliability of the information fluctuates depending on environmental conditions, and (2) users can specify arbitrary destinations on roads, which do not have distinct visual characteristics for prediction.

This thesis is composed of six chapters.

Chapter 1 describes the research background, objectives, and structure of this thesis.

Chapter 2 reviews related work, particularly on natural language understanding and symbol grounding, as well as techniques for integrating multimodal information, with a specific focus on vision and language, and positions this research in relation to them.

Chapter 3 provides the theoretical background to the foundational techniques for the integrated processing of multimodal information used in this thesis, clearly demonstrating the principles behind the proposed methods.

Chapter 4 focuses on the Referring Navigable Region task, which aims to localize target regions and proposes Trimodal Navigable Region Segmentation Model (TNRSM). Experiments demonstrate that TNRSM can compensate for differences in the type of information across modalities and address reliability degradation of information that fluctuates depending on environmental conditions.

Chapter 5 focuses on the Vision-and-Language Target Positioning task, which aims to localize target positions and proposes the Target Regressor in Positioning (TRiP). Experiments demonstrate that TRiP effectively addresses the challenge where users specify arbitrary destinations on roads, which do not have distinct visual characteristics for prediction.

Chapter 6 concludes this thesis with an overview and directions for future research.