M1 サーベイレポート

学籍番号: 82518829 氏名: 戸倉 健登

所属: 杉浦孔明研究室

1. はじめに

ロボティクスにおける視覚言語理解は，ロボットが人間の指示を理解し，環境内のオブジェクトを認識・操作するために重要な技術である．特に，Vision-Language Model（VLM）の発展により，ロボットは画像情報と自然言語を統合的に処理できるようになった [1–4]．本レポートでは，ロボティクスにおけるマルチモーダル検索，特にシーンテキストを活用した手法に焦点を当て，関連研究を分類・整理する．

2. 研究背景と課題

実世界環境において，ロボットが自然言語指示に基づいてオブジェクトを検索する際，視覚的に類似したオブジェクトの識別が課題となる．例えば，類似するオブジェクトが多い環境において「赤いコカコーラの缶を取って」という指示では，色だけでなくラベルのテキスト情報が重要な手がかりとなる．しかし，従来のマルチモーダル検索手法は，シーンテキスト（ラベル，標識など）を明示的に活用していない [5,6]．本レポートでは，この課題に対する既存研究を以下の 4 つのカテゴリに分類して整理する．

3. 基盤モデルとロボティクス

近年，大規模言語モデル（LLM）や VLM などの基盤モデルをロボティクスに応用する研究が活発化している．Brohan et al. [2] は，RT-1 を提案し，大規模なロボット軌跡データで訓練された Vision-Language-Action モデルが，多様な操作タスクにわたって高い汎化性能を示すことを実証した．さらに，Driess et al. [7] の PaLM-E は，視覚と言語を統合した大規模モデルをロボット制御に適用し，複雑なタスクの実行を可能にした．最近では，Black et al. [8] の $\pi 0$，Wen et al. [9] の DexVLA，Gemini-Robotics [10] など，より高度な視覚理解とロボット能力を統合したモデルが提案されている．Li et al. [11] の ControlVLA は，少数ショット学習によるオブジェクト中心の適応を実現している．これらの手法は優れた性能を示す一方で，シーンテキストを明示的に組み込んでおらず，テキスト情報が重要な意味的グラウンディングを提供するシーンにおいて，マルチモーダル理解が制限されている．

4. Retrieval-Augmented Generation (RAG) の応用

RAG は，外部知識を検索して言語生成を強化する手法であり，ロボティクスへの応用が進んでいる．Xie et al. [12] の Embodied-RAG は，一般的な非パラメトリック表現を用いて，Embodied エージェントのグラウンディングと計画を強化している．ここで，Embodied エージェントとは物理的な身体を持たず，テキストやシミュレーション上でのみ行動するエージェントと異なり，現実世界の環境においてセンサーやアクチュエータを通じて知覚・行動を行うエージェントを指す．同様に，Zhu et al. [13]，Xu et al. [14]，Monaci et al. [15]，Wang et al. [16] も，RAG を Embodied エージェントに適用する手法を提案している．Wang et al. [17] と Fan et al. [18] は，RAG を使用してナビゲーション指示を生成し，Vision-Language ナビゲーションモデルを改善している．また，Anwar et al. [19] は，RAG ベースのメモリを長期的推論に活用する手法を提案し，Kumar et al. [20] は，デモンストレーションを検索・融合して模倣方策学習を行う手法を提案している．これらの研究は，検索信号としてシーンテキストを活用しておらず，日常環境において視覚的に類似したオブジェクトを曖昧性解消するという点で改善の余地がある．

5. マルチモーダル検索手法

ロボットの自然言語指示に基づくオブジェクト検索は，重要な研究課題である．Yenamandra et al. [21] と RoboCup@Home [22] は，自然言語指示に基づくモバイル操作タスクにおける家庭用サービスロボットの性能を評価している．NLMap [23]，Sigurdsson et al. [24] の RR-ExBoT，DM^{2}RM [25]RelaX-Former [26] は，環境内のターゲットオブジェクトを識別するためにマルチモーダル検索を実行し，自然言語指示に基づいて関連画像のランク付けリストを出力する．これらの手法は急速に進歩しているが，本質的に曖昧性解消情報を含むシーンテキスト（ラベル，標識など）の活用には重要なギャップが残っている．

6. シーンテキスト処理技術

シーンテキストを含む画像の処理は，多様なタスクで研究されている．テキスト認識 [27,28]，テキスト検出 [29–31]，テキスト検索 [31–33]，視覚的質問応答 [34–36]，テキスト合成 [37–39] など，様々な技術が開発されている．Mishra et al. [40] は，シーンテキスト検索タスクを導入し，文字を検出・分類してクエリテキストを含む画像の確率を計算する 2 段階パイプラインを提案した．Gomez et al. [41] は，テキスト提案を予測し，クエリテキストとの類似性に基づいて画像をランク付けするエンドツーエンドネットワークを提案した．ViSTA [33] は，BERT [42] と線形投影を使用して OCR で抽出されたテキストと位置を埋め込み，トランスフォーマー内の融合トークンを介してテキストと位置を視覚特徴と融合する．しかし，従来手法はシーンテキストをそのまま使用しており，オブジェクトとその属性との関連性を捉えたナラティブ表現への統合は行われていない．

7. 評価ベンチマーク

マルチモーダル検索とシーンテキストの評価には，様々なベンチマークが使用されている．COCO [43] と Flickr30K [44]

は，もともと画像キャプショニングベンチマークであるが，マルチモーダル検索タスクに広く使用されている．REVERIE [45] と GOAT-Bench [46] は，オブジェクト目標ナビゲーションタスク用に設計されており，LTRRIE [47] は画像のランク付けによるマルチモーダル検索に焦点を当てている．ST-VQA [35] は，シーンテキストベースの VQA に使用され，TextCaps [48] と COCO-Text-Captioned [49] は，シーンテキストを考慮した画像キャプショニングに適している．Visual Genome [50] には，シーンテキストと詳細なアノテーションを持つ画像が含まれており，RefText [51] は，参照表現理解のベンチマークであり，その一部はシーンテキストを活用している．既存ベンチマークの多くは，シーンテキストを欠いているか，マルチモーダル検索用に設計されていないか，ロボットタスクに適していないか，または過度に単純な指示を含んでいる．したがって，シーンテキストと視覚コンテンツにわたるきめ細かい推論を必要とする，ロボットのための実世界の指示ベースの検索タスクを評価するには不十分である．

8. まとめと今後の展望

本レポートでは，ロボティクスにおけるマルチモーダル検索，特にシーンテキストを活用した手法に関する研究を 4 つのカテゴリ（基盤モデル，RAG，マルチモーダル検索，シーンテキスト処理）に分類して整理した．各カテゴリにおいて優れた研究成果が報告されているが，シーンテキストを明示的に活用したロボットのオブジェクト検索という観点では，まだ研究の余地が大きい．今後の研究課題として，以下が挙げられる：(1) シーンテキストとオブジェクトの関連性を捉えたナラティブ表現の開発，(2) シーンテキストを検索信号として活用する RAG フレームワークの構築，(3) 実世界のロボットタスクに適した評価ベンチマークの整備．これらの課題に取り組むことで，より高度なロボットの視覚言語理解が実現できると期待される．

参考文献

[1] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, et al. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In *CoRL*, pp. 287–318, 2023.

[2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, et al. RT-1: Robotics Transformer for Real-World Control at Scale. *RSS*, 2023.

[3] Xuan Xiao, Jiahang Liu, Zhipeng Wang, Yanmin Zhou, Yong Qi, Shuo Jiang, Bin He, and Qian Cheng. Robot Learning in the Era of Foundation Models: A Survey. *Neurocomputing*, Vol. 638, p. 129963, 2025.

[4] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A Survey on Vision-Language-Action Models for Embodied AI. *arXiv preprint arXiv:2405.14093*, 2024.

[5] Shangbang Long, Xin He, and Cong Yao. Scene Text Detection and Recognition: The Deep Learning Era. *IJCV*, Vol. 129, pp. 161–184, 2018.

[6] Neeraj Gupta and Anand Jalal. Traditional to transfer learning progression on scene text detection and recognition: a survey. *Artificial Intelligence Review*, Vol. 55, No. 4, pp. 3457–3502, 2022.

[7] Danny Driess, F. Xia, Mehdi Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, et al. PaLM-E: An Embodied Multimodal Language Model. In *ICML*, pp. 8469–8488, 2023.

[8] Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. In *CoRL*, 2025.

[9] Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. DexVLA: Vision-Language Model with Plug-In Diffusion Expert for General Robot Control. *CoRL*, 2025.

[10] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Baptiste Alayrac, Arenas Gonzalez, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini Robotics: Bringing AI into the Physical World. *arXiv preprint arXiv:2503.20020*, 2025.

[11] Puhao Li, Yingying Wu, Ziheng Xi, Wanlin Li, Yuzhe Huang, Zhiyuan Zhang, Yinghan Chen, Jianan Wang, Song-Chun Zhu, Tengyu Liu, and Siyuan Huang. ControlVLA: Few-shot Object-centric Adaptation for Pre-trained Vision-Language-Action Models. *CoRL*, 2025.

[12] Quanting Xie, So Yeon Min, Pengliang Ji, Yue Yang, Tianyi Zhang, Kedi Xu, Aarav Bajaj, Ruslan Salakhutdinov, Matthew Johnson-Roberson, and Yonatan Bisk. Embodied-RAG: General Non-parametric Embodied Memory for Retrieval and Generation. *arXiv preprint arXiv:2409.18313*, 2025.

[13] Yichen Zhu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. Retrieval-Augmented Embodied Agents. In *CVPR*, pp. 17985–17995, 2024.

[14] Weiye Xu, Min Wang, Wengang Zhou, and Houqiang Li. P-RAG: Progressive Retrieval Augmented Generation For Planning on Embodied Everyday Task. In *ACMMM*, p. 6969–6978, 2024.

[15] Gianluca Monaci, Rafael Rezende, Romain Deffayet, Gabriela Csurka, Guillaume Bono, Herve Dejean, Stephane Clinchant, and Christian Wolf. RANa: Retrieval-Augmented Navigation. *arXiv preprint arXiv:2504.03524*, 2025.

[16] Kuanning Wang, Yuqian Fu, Tianyu Wang, Yanwei Fu, Longfei Liang, Yu-Gang Jiang, and Xiangyang Xue. RAG-6DPose: Retrieval-Augmented 6D Pose Estimation via Leveraging CAD as Knowledge Base. In *IROS*, 2025.

[17] Zihan Wang, Yaohui Zhu, Gim Hee Lee, and Yachun Fan. NavRAG: Generating User Demand Instructions for Embodied Navigation through Retrieval-Augmented LLM. In *ACL*, pp. 8430–8440, 2025.

[18] Sheng Fan, Rui Liu, Wenguan Wang, and Yi Yang. Navigation Instruction Generation with BEV Perception and Large Language Models. In *ECCV*, 2024.

[19] Abrar Anwar, John Welsh, Joydeep Biswas, Soha Pouya, and Yan Chang. ReMEmbR: Building and Reasoning Over Long-Horizon Spatio-Temporal Memory for Robot Navigation. In *ICRA*, pp. 2838–2845, 2025.

[20] Sateesh Kumar, Shivin Dass, Georgios Pavlakos, and Roberto Martin. COLLAGE: Adaptive Fusion-based Retrieval for Augmented Policy Learning. In *CoRL*, 2025.

[21] Sriram Yenamandra, A. Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Théophile Gervet, Tsung Yang, Vidhi Jain, Alexander Clegg, John Turner, Zsolt Kira, Manolis Savva, Angel Chang, Devendra Chaplot, et al. HomeRobot: Open-Vocabulary Mobile Manipulation. In *CoRL*, pp. 1975–2011, 2023.

[22] Mauricio Matamoros, Viktor Seib, Raphael Memmesheimer, and Dietrich Paulus. RoboCup@Home: Summarizing achievements in over eleven years of competition. In *ICARSC*, pp. 186–191, 2018.

[23] Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, et al. Open-vocabulary Queryable Scene Representations for Real World Planning. In *ICRA*, pp. 11509–11522, 2023.

[24] Gunnar Sigurdsson, Jesse Thomason, Gaurav Sukhatme, and Robinson Piramuthu. RREx-BoT: Remote Referring Expressions with a Bag of Tricks. In *IROS*, pp. 5203–5210, 2023.

[25] R. Korekata, K. Kaneda, S. Nagashima, Y. Imai, and K. Sugiura. DM^{2}RM: Dual-Mode Multimodal Ranking for Target Objects and Receptacles Based on Open-Vocabulary Instructions. *AR*, Vol. 39, No. 5, pp. 243–258, 2025.

[26] Daichi Yashima, Ryosuke Korekata, and Komei Sugiura. Open-Vocabulary Mobile Manipulation Based on Double Relaxed Contrastive Learning With Dense Labeling. *IEEE RA-L*, Vol. 10, No. 2, pp. 1728–1735, 2025.

[27] Zhen Zhao, Jingqun Tang, Chunhui Lin, Binghong Wu, Hao Liu, Zhizhong Zhang, Xin Tan, Can Huang, and Yuan Xie. Multi-modal In-Context Learning Makes an Ego-evolving Scene Text Recognizer. In *CVPR*, pp. 15567–15576, 2023.

[28] Jianjun Xu, Yuxin Wang, Hongtao Xie, and Yongdong Zhang. OTE: Exploring Accurate Scene Text Recognition Using One Token. In *CVPR*, pp. 28327–28336, 2024.

[29] Min Liang, Wei Ma, Xiaobin Zhu, Jingyan Qin, and Xu-Cheng Yin. Layoutformer: Hierarchical text detection towards scene text understanding. In *CVPR*, pp. 15665–15674, 2024.

[30] Sergi Garcia-Bordils, Dimosthenis Karatzas, and Marçal Rusiñol. STEP – Towards Structured Scene-Text Spotting. In *WACV*, pp. 883–892, 2024.

[31] Jinzhi Zheng, Heng Fan, and Libo Zhang. Kernel Adaptive Convolution for Scene Text Detection via Distance Map Prediction. In *CVPR*, pp. 5957–5966, 2024.

[32] Gangyan Zeng, Yuan Zhang, Jin Wei, Dongbao Yang, Peng Zhang, Yiwen Gao, Xugong Qin, and Yu Zhou. Focus, Distinguish, and Prompt: Unleashing CLIP for Efficient and Flexible Scene Text Retrieval. In *MM*, pp. 2525–2534, 2024.

[33] Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. ViSTA: Vision and Scene Text Aggregation for Cross-Modal Retrieval. In *CVPR*, pp. 5174–5183, 2022.

[34] Ali Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R. Manmatha. LaTr: Layout-Aware Transformer for Scene-Text VQA. In *CVPR*, pp. 16527–16537, 2021.

[35] Ali Furkan Biten, Rubèn Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, C.V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Scene Text Visual Question Answering. In *ICCV*, pp. 4290–4300, 2019.

[36] Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. Multi-Modal Graph Neural Network for Joint Reasoning on Vision and Scene Text. In *CVPR*, pp. 12746–12756, 2020.

[37] Jialei Cui, Jianwei Du, Wenzhuo Liu, and Zhouhui Lian. TextNeRF: A Novel Scene-Text Image Synthesis Method based on Neural Radiance Fields. In *CVPR*, pp. 22272–22281, 2024.

[38] Chen Duan, Pei Fu, Shan Guo, Qianyi Jiang, and Xiaoming Wei. ODM: A Text-Image Further Alignment Pre-training Approach for Scene Text Detection and Spotting. In *CVPR*, pp. 15587–15597, 2024.

[39] Joshua Santoso and Christian Williem Simon. On Manipulating Scene Text in the Wild with Diffusion Models. In *WACV*, pp. 5202–5211, 2024.

[40] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: Visual Question Answering by Reading Text in Images. In *ICDAR*, pp. 947–952, 2019.

[41] Lluis Gomez, Andres Mafla, Marçal Rusiñol, and Dimosthenis Karatzas. Single Shot Scene Text Retrieval. In *ECCV*, pp. 700–715, 2018.

[42] Jacob Devlin, Ming Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, pp. 4171–4186, 2019.

[43] Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, pp. 740–755, 2014.

[44] Peter Young, et al. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *TACL*, Vol. 2, pp. 67–78, 2014.

[45] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, et al. REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. In *CVPR*, pp. 9979–9988, 2020.

[46] Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. GOAT-Bench: A Benchmark for Multi-Modal Lifelong Navigation. In *CVPR*, pp. 16373–16383, 2024.

[47] Kanta Kaneda, et al. Learning-To-Rank Approach for Identifying Everyday Objects Using a Physical-World Search Engine. *IEEE RA-L*, Vol. 9, No. 3, pp. 2088–2095, 2024.

[48] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. TextCaps: A Dataset for Image Captioning with Reading Comprehension. In *ECCV*, pp. 742–758, 2020.

[49] Andres Mafla, S. Rezende, Lluis Gómez, Diane Larlus, and Dimosthenis Karatzas. StacMR: Scene-Text Aware Cross-Modal Retrieval. In *WACV*, pp. 2219–2229, 2021.

[50] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li Li, David Shamma, Michael Bernstein, and Li Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *IJCV*, Vol. 123, No. 1, pp. 32–73, 2017.

[51] Yuqi Bu, et al. Scene-Text Oriented Referring Expression Comprehension. *IEEE TMM*, Vol. 25, pp. 7208–7221, 2023.