環境端末による 3D Scene Graph の管理・構築手法

○宗藤 慶汰（九州大学），井塚 智也（九州大学），倉爪 亮（九州大学）

A method for Management and Construction of 3D Scene Graphs Using an Environmental-installed Terminal

○ Keita Munetou (Kyushu Univ.), Tomoya Itsuka (Kyushu Univ.), and Ryo Kurazume (Kyushu Univ.)

Abstract: A 3D Scene Graph (3DSG), which represents information about objects and their relationships within an environment, is an effective representation method for achieving advanced robot tasks. In this paper, we propose a method to centrally manage environmental information using an environmental-installed terminal. This is achieved by a robot that performs SLAM using the real-time 3DSG construction method, Clio, storing the resulting graph structure in a graph database and the object feature vectors in a vector database. This approach will enable the robot to get environmental information at low cost. In our experiments, we confirmed that a robot can successfully navigate to a target object's position based on this shared information.

1. 緒言

複数台のロボットが協調して高度なサービスを提供するためには，各ロボットが観測した環境情報を統合し，共有するための基盤技術が必要不可欠である．一方，環境情報構造化は，環境中にセンサ等を設置して環境を知能化することでロボット単体のコストを低減する手法である[1]．また，環境内の物体情報とそれらの関係性を表現する 3D Scene Graph (3DSG)[2] は，高度なロボットタスクを実現する上で有効な表現手法として広く用いられている[3), 4)]．

本稿では，環境情報構造化の考えに基づき，ロボットではなく環境側が情報を管理するアーキテクチャを志向し，新たな情報表現として 3DSG を採用した設置型端末による環境情報管理・構築手法を提案する (**Fig. 1**)．本手法は，リアルタイム 3DSG 構築手法である Clio[3] を拡張し，構築された 3DSG と物体の CLIP[5] 特徴量をそれぞれグラフデータベースとベクトルデータベースに格納する．これにより，ロボットは環境中の情報を低コストで参照可能となり，効率的な協調タスクの実現が期待できる．本稿では，提案アーキテクチャの基本機能を確認するため，Clio の公開データセットを用いた物体探索評価実験と，実環境におけるナビゲーション実験を行った．

2. 先行研究

ロボットが高度なタスクを遂行するためには，単なる幾何学的地図ではなく，環境内の物体とその関係性を含む意味的地図が必要である．例として，「建物内のすべての椅子を集める」というタスクを考える．BUMBLE[6] が用いるトポロジカル視覚マップでは，部屋間の大まかな移動計画は可能だが，各部屋のどの場所に椅子があるのかを効率的に把握することは難しい．また，ReMEmbR[7] が構築する時空間データベースは，過去の観測情報を活用して物体の位置を推定できるが，建物全体の物体関係性を網羅的に表現することは困難である．

これに対し，本研究で採用する 3DSG は「建物-部屋-場所-物体」という階層構造を持つため「物体階層の中のすべての椅子を検索する」という単純なクエリで目標リストを取得でき，効率的なタスク遂行が可能である．

3DSG を構築する研究は数多く存在するが，その中

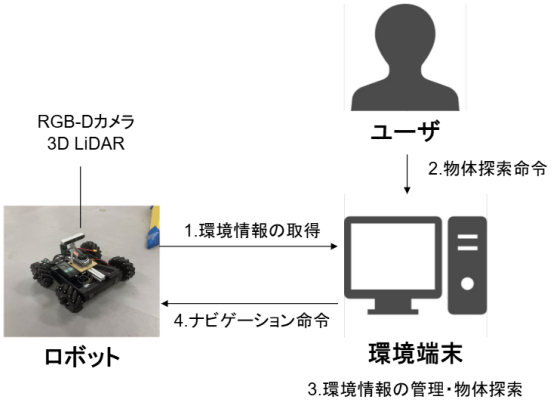

Fig. 1 overview of the proposed architecture

でも Clio は，ロボットのタスクに応じてリアルタイムに 3DSG を構築・更新する手法である．Clio は，ロボットが実行すべきタスクのリストを自然言語で受け取り，そのタスクの遂行に必要十分な情報のみをマップに保持することを目的とする．具体的には，Segment Anything Model (SAM)[8] を用いて RGB-D カメラから得た画像情報を細かく分節化し，CLIP を用いて各領域の CLIP 特徴量を取得する．そして，Information Bottleneck Principle[9] に基づいてタスクに関連する物体のみをグラフに組み込むことで，タスク指向の 3D シーングラフを構築している．

近年，LLM や Vision Language Model (VLM) を活用した高度なロボットタスクの遂行が注目されている．SYNERG-AI[4] などの研究では，ロボットが高度な推論を行うために，構造化された意味的環境表現を活用する手法の有効性が示唆されている．SYNERG-AI[4] は，VLM を用いて物体の属性や関係性を抽出した 3DSG を LLM に提供することで複雑なタスクを遂行しており，本研究で提案する 3DSG 管理手法は，このような LLM 連携タスクにおいても有効であると考える．

本研究ではこれらの先行研究を踏まえ，リアルタイムに 3DSG を構築・更新することが可能であり，LLM 連携も可能と考えられる Clio を 3DSG 構築手法として採用した．

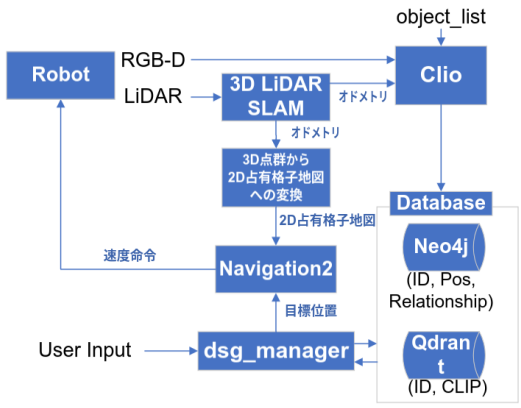

Fig. 2 system structure of the proposed architecture

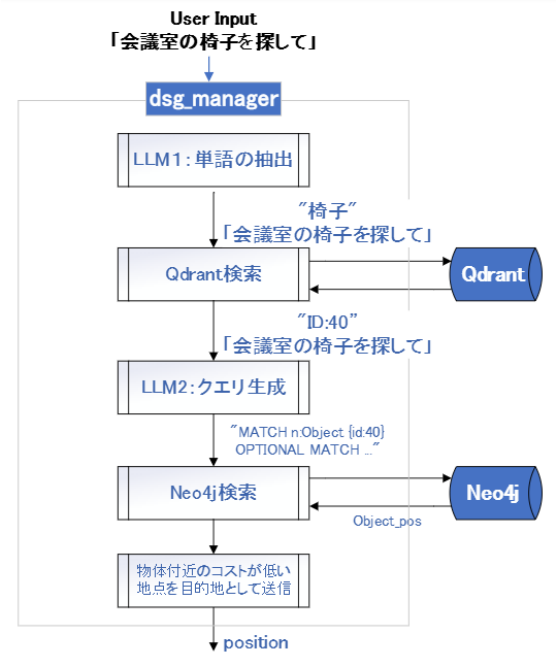

Fig. 3 dsg_manager

3. 提案手法

3.1 システム構成

提案手法のシステム構成を **Fig. 2** に示す．本システムは，環境情報を取得・処理する 3DSG 構築部，得られた情報を格納するデータベース，そしてロボットへのナビゲーション命令を生成する物体探索実行部 (dsg_manager) から構成される．本システムでは 3DSG 構築部として Clio を採用した．そのため，本研究では主にデータベースと物体探索実行部の設計に取り組み，システム全体としての評価を行った．

ロボットは RGB-D カメラの情報と自己位置推定結果を環境端末に送信する．環境端末では，3DSG 構築部が Clio を用いて 3DSG を構築・更新する．構築された 3DSG の構造はグラフデータベースに，各物体の CLIP 特徴はベクトルデータベースに保存される．物体探索実行部は，データベースへの問い合わせを行い，ロボットからの要求に応じてナビゲーション命令を生成する．その他の部分として，FAST-LIO2[10] が自己位置推定と環境の 3 次元地図構築を行い，(just_mapper が)3 次元点群地図から 2 次元占有格子地図への変換を行う．Navigation2[11] は 2 次元占有格子地図と自己位置情報をもとに，物体探索実行部からのゴール指示に基づいてナビゲーションを行う．

3.2 内部処理

3.2.1 3DSG 構築部

Clio には事前に認識対象となる物体のリストを与える．このリストに基づき，Clio は RGB-D カメラ画像と自己位置推定結果を用いて 3DSG を構築する．本研究では，Clio のプログラムを拡張し，構築されたグラフ構造と検出された物体の CLIP 特徴をデータベースに保存できるよう拡張した．

3.2.2 データベース

本研究では 3DSG を管理するため，グラフデータベースとベクトルデータベースを併用する．構築した 3DSG をグラフデータベースに格納することで，「会議室にある椅子」のような部屋と物体の親子関係に基づくクエリを効率的に処理できる．また，検出された物体の CLIP 特徴をベクトルデータベースに保存することで，「赤い椅子」のような特徴付きの指示に対しても高速な類似物体検索が可能となる．CLIP は，膨大な画像とテキストのペアから学習することで，画像とテキストを同じ「意味空間」上のベクトルとして表現できる AI モデルである．そのため，事前に定義されていない未知の物体に対しても柔軟な意味理解が可能となり，本システムでの物体探索を可能にしている．以上の 2 つのデータベースの組み合わせにより，意味的・空間的に複雑な問い合わせへの対応を実現する．実装では，グラフデータベースとして Neo4j を，ベクトルデータベースとして Qdrant を利用した．

3.2.3 物体探索実行部 (dsg_manager)

提案アーキテクチャにおける物体探索実行部の内部構成を **Fig. 3** に示す．物体探索実行部は，データベースに保存された情報を利用して，ユーザからの自然言語指示に基づきナビゲーション命令を生成する役割を担う．その処理は以下のステップで実行される．

1. **指示解釈:** ユーザからの「物体 A と同じ部屋にある物体 B を探して」といった自然言語指示に対し，まず LLM を用いて探索対象の「ターゲット (Target)」(物体 B) と，位置関係の基準となる「ランドマーク (Landmark)」(物体 A) の文字列を抽出する．
2. **候補特定:** 抽出した物体名に基づき、ベクトルデータベースを検索する．ここで物体の CLIP 特徴を比較し，類似度の高い順に候補となる物体の ID リストを取得する．
3. **クエリ生成:** 元の指示と候補 ID リストを LLM に渡し，グラフデータベースへの問い合わせクエリ (Cypher) を動的に生成させる．このクエリは，候補 ID を類似度の高い順に探索し，かつ部屋や他の物体との位置関係を考慮するよう設計した．
4. **位置情報取得:** 生成されたクエリによりグラフデータベースへ問い合わせを行い，ターゲットとなる物体の正確な位置座標を取得する．
5. **ナビゲーション命令生成:** 最後に，取得した座標を基に，コストマップ上で物体近傍の到達可能な地点を目標としてナビゲーション命令を送信する．

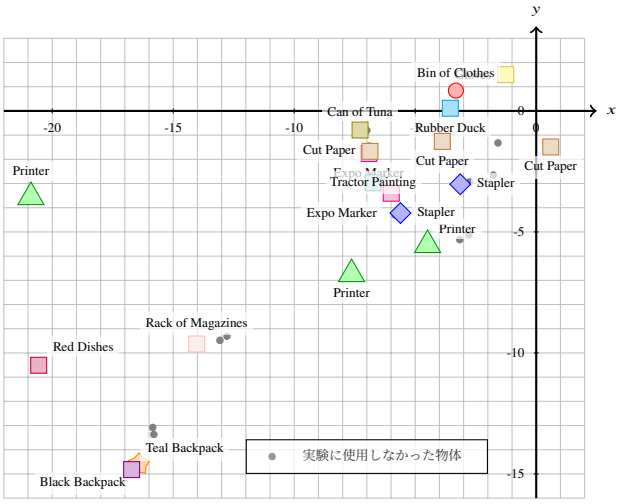

Fig. 4 object placement in the object search evaluation experiment

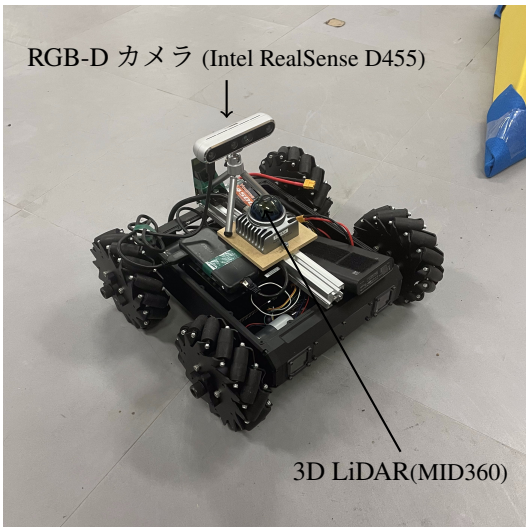

Fig. 5 robot used in the navigation experiment

4. 実験

4.1 目的

本実験では，提案アーキテクチャにより，様々な難易度の物体探索が可能であるか確認する．

4.2 実験環境

実験は以下の 2 つの方法で実施した．

1. **物体探索評価実験：** Clio の公開データセットを用いて，構築された 3DSG に対する探索タスクの精度を評価した．データセット内の物体は **Fig. 4** のように配置されている．ナビゲーションは行わず，探索クエリの成否のみを評価対象とした．
   また評価指標は，5 回の試行における物体探索の成功率とした．
2. **ナビゲーション実験：** **Fig. 5** に示すロボットを用い，部屋内に **Fig. 6** のように椅子，教科書等の物体を配置し，ロボットによる実機ナビゲーションを行った．
   評価指標は，ナビゲーションの成功率とした．

両実験ともに，評価タスクとして以下の 3 種類のタスクを設定した．

- タスク 1：属性探索（例：「物体 A を探して」）
- タスク 2：直接的関係探索（例：「部屋 B にある物体 A を探して」）
- タスク 3：間接的関係探索（例：「物体 A と同じ部屋にある物体 B を探して」）

なお，本実験におけるタスク指示はターミナルから ros2 topic を介して物体探索実行部に送信した．また，システムの応用可能性を検証するため，LLM を介した音声入力によるナビゲーションのデモンストレーションを実施した．

4.3 結果と考察

公開データセットによる物体探索，および実空間ナビゲーション実験結果を **Table 1** に示す．両実験ともにタスク 1 では高い成功率を示し，提案手法が基本的な物体探索に有効であることが確認できた．また，定量評価とは別に実施した音声入力によるナビゲーションのデモンストレーションでは，指示に基づきロボットが正しく目標地点へ到達でき，実環境でのロボットタスクへの応用可能性を確認した．

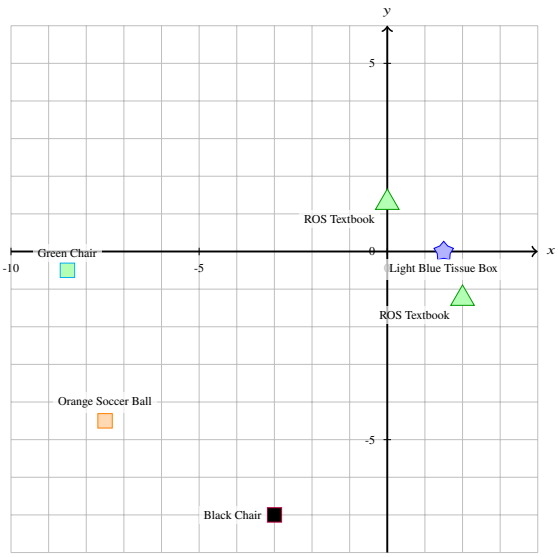

Fig. 6 object placement in the navigation experiment

しかし，より複雑なタスクであるタスク 2 とタスク 3 では成功率が低下し，システムの課題が明らかになった．失敗事例を分析した結果，以下の 3 つが主な要因であると考える．

1. **CLIP 特徴検索の失敗：** CLIP 特徴に基づくベクトルデータベースの検索において，誤った物体を第一候補として選択する事例が稀に見られた．
2. **部屋ラベルの不一致とクエリ生成の失敗（タスク 2）：** グラフデータベースへの問い合わせにおいて，Clio が推定した部屋ラベルと，ユーザの指示に基づき物体探索実行部が生成した部屋ラベルが一致せず，検索に失敗するケースがあった．これに加え，タスク 2 の失敗には LLM による Cypher クエリ生成ミスも一因として確認されており，これはプロンプトを修正することで改善が見込まれる．
3. **部屋ノードの粒度の不一致（タスク 3）：** 間接的関係探索の失敗は，主に人間が認識する「一つの部屋」と，Clio が認識した部屋の数 (生成した「room

Table 1 results of the object search evaluation experiment and the navigation experiment

experiment	Task1	Task2	Task3
object search evaluation	4/5(80%)	4/5(80%)	2/5(40%)
navigation	4/5(80%)	0/2(0%)	1/1(100%)

ノード」の数) に乖離があることに起因していた．Clio は幾何的な特徴から一つの部屋を複数の領域（room ノード）に分割することがあり，人間が「同じ部屋にある」と期待して探索しても，グラフ構造上は異なる部屋ノードに属しているため探索に失敗するケースがあった．なお，タスク 3 のナビゲーション実験は，Clio が検出したオブジェクトが少なかったため 1 回のみの試行となった．

また，タスクの成否とは別に，システムのリアルタイム性に関する性能上の課題も明らかとなった．3DSG のノードが追加・削除されるたびにデータベースへの書き込み処理が発生する．この処理は 3DSG 構築と並行して行われるが，全体の完了までには 3DSG 構築時間とほぼ同等の追加時間が必要となり，迅速な応答性が求められるタスクにおいてボトルネックとなる可能性がある。

5. 結言

本稿では，環境情報構造化の考えに基づき，ロボットではなく環境側が情報を管理するアーキテクチャを志向し，新たな情報表現として 3DSG を採用した環境端末による環境情報の管理・構築手法を提案した．Clio を用いて構築した 3DSG と物体の CLIP 特徴を，それぞれグラフデータベースとベクトルデータベースに格納し，環境情報を集約するアーキテクチャを設計した．実験を通して，基本的な物体探索タスクにおいてはロボットが目標物体へのナビゲーションタスクを達成できることを確認した一方で，複雑な条件を含むタスクでは LLM のプロンプト設計や環境認識の粒度といった課題が明らかになった．これにより，提案手法の基本的な有効性と今後の改善点が示された．

今後の課題として，複数台のロボットを用いた実機実験が挙げられる．具体的には，実環境において，動的な環境変化への対応や，複数ロボットの協調動作における提案手法の有効性を検証する．また，分散配置された端末間でのデータベース連携手法の確立やより高度で複雑なタスクを遂行可能なシステムの実現を目指す．

参考文献

[1] 日本ロボット学会編.: ロボット工学ハンドブック, 第 3 版, OCLC: 1371187224, コロナ社, Tokyo (2023), pp. 584–585.

[2] I. Armeni, Z.-Y. He, A. Zamir, J. Gwak, J. Malik, M. Fischer, and S. Savarese: 3D Scene Graph: A Structure for Unified Semantics, 3D Space, and Camera, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Seoul, Korea (South) (2019), pp. 5663–5672, DOI: 10.1109/ICCV.2019.00576, URL: https://ieeexplore.ieee.org/document/9008302/ (accessed on 2025-09-24).

[3] D. Maggio, Y. Chang, N. Hughes, M. Trang, D. Griffith, C. Dougherty, E. Cristofalo, L. Schmid, and L. Carlone: Clio: Real-time Task-Driven Open-Set 3D Scene Graphs, IEEE Robotics and Automation Letters, 9.10, pp. 8921–8928 (2024), DOI: 10.1109/LRA.2024.3451395.

[4] Y. Chen, G. Zhang, Y. Zhang, H. Xu, P. Zhi, Q. Li, and S. Huang: SYNERGAI: Perception Alignment for Human-Robot Collaboration.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever: Learning Transferable Visual Models From Natural Language Supervision, International Conference on Machine Learning (2021), URL: https://www.semanticscholar.org/paper/6f870f7f02a8c59c3e23f407f3ef00dd1dcf8fc4 (accessed on 2025-09-24).

[6] R. Shah, A. Yu, Y. Zhu, Y. Zhu, and R. Martín-Martín: BUMBLE: Unifying Reasoning and Acting with Vision-Language Models for Building-wide Mobile Manipulation, arXiv:2410.06237, arXiv:2410.06237 [cs] (2024), DOI: 10.48550/arXiv.2410.06237, URL: http://arxiv.org/abs/2410.06237.

[7] A. Anwar, J. Welsh, J. Biswas, S. Pouya, and Y. Chang: ReMEmbR: Building and Reasoning Over Long-Horizon Spatio-Temporal Memory for Robot Navigation, (2024), DOI: 10.48550/arXiv.2409.13682, arXiv: 2409.13682[cs], URL: http://arxiv.org/abs/2409.13682 (accessed on 2025-08-05).

[8] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick: Segment Anything, 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Paris, France (2023), pp. 3992–4003, DOI: 10.1109/ICCV51070.2023.00371, URL: https://ieeexplore.ieee.org/document/10378323/ (accessed on 2025-09-24).

[9] N. Tishby, F. C. Pereira, and W. Bialek: The information bottleneck method, (2000), DOI: 10.48550/arXiv.physics/0004057, arXiv: physics/0004057, URL: http://arxiv.org/abs/physics/0004057 (accessed on 2025-09-24).

[10] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang: FAST-LIO2: Fast Direct LiDAR-Inertial Odometry, IEEE Transactions on Robotics, 38.4, Number: 4, pp. 2053–2073 (2022), DOI: 10.1109/TRO.2022.3141876, URL: https://ieeexplore.ieee.org/document/9697912 (accessed on 2025-02-07).

[11] S. Macenski, F. Martin, R. White, and J. G. Clavero: The Marathon 2: A Navigation System, 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Conference Name: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) ISBN: 9781728162126 Place: Las Vegas, NV, USA Publisher: IEEE, pp. 2718–2725 (2020), DOI: 10.1109/IROS45743.2020.9341207, URL: https://ieeexplore.ieee.org/document/9341207/ (accessed on 2025-09-24).