

産業環境のリモート管理に向けた物体画像情報のデータベース構築及び 大規模言語モデルを用いた検索・回答システムの基礎検討

○倉持 光希 (中央大学), 椎野 丞ノ進 (中央大学), 朝木 直道 (中央大学), 安武 和成 (クラフティア), 古野 純二 (クラフティア), 梅田 和昇 (中央大学), Sarthak Pathak (芝浦工業大学)

A Fundamental Study on the Construction of an Object Image Information Database and the Development of a Search and Question-Answering System Using Large Language Models for Remote Management in Industrial Environments

○ Kouki KURAMOCHI (Chuo University), Jonoshin SHIINO (Chuo University), Naomichi ASAKI (Chuo University), Kazushige YASUTAKE (KRAFTIA), Junji FURUNO (KRAFTIA), Kazunori UMEDA (Chuo University), and Sarthak Pathak (Shibaura Institute of Technology)

Abstract: This study presents a fundamental investigation into the application of object recognition and deep learning for promoting digital transformation (DX) and enabling remote management in industrial environments. In large-scale sites, the enormous volume of information often makes management challenging. To address this issue, we propose a system that leverages a database to simplify and improve management efficiency. By inputting image information and spatial coordinates of the site, the system enables appropriate search and question-answering functions to provide the required information effectively and object information.

1. 緒言

近年、現代社会において建設現場のリモート管理の重要性は急速に高まっている。その背景には、高齢化や人口減少に伴う労働力不足、業務効率化の必要性、都市化やグローバル化に起因する建設規模の拡大、さらにデジタル技術の進展が挙げられる。建設業界では従来の労働集約的な管理手法から、省力化かつ効率的な管理手法への転換が求められている。

こうした流れの中で、360度カメラを用いた現場の可視化とクラウド管理システムが注目されている。たとえば米国 OpenSpace 社は、ヘルメットやドローンに搭載した360度カメラで撮影した画像を自動的に平面図と紐付け、施工進捗を可視化する仕組み(現場のストリートビュー作成)を実用化している。^[1] このように、遠隔から現場を確認できる仕組みは効率化に寄与している。一方で、依然として課題が残されている。それは、画像内に存在する工具・部品・設備といった個々の物体を詳細に把握し、その数や状態、機能を自動的に検索することや、物体同士の位置関係、空間の理解などがストリートビュー上では困難である点である。空間構造の把握に優れる LiDAR や Visual SLAM といった3次元点群技術でさえも、生成される点群データは単なる色の付いた点の集合体に過ぎず、個々の物体を意味的に分離・認識することは困難である。また、膨大なデータ量に起因する処理負荷の高さも実用上の課題となっている。こういったものは現在、多くの場合、人が目視で確認する必要がある、効率性の向上を妨げている。

これら双方の課題を解決するためには、360度画像

やそこから生成される3次元点群データに基づき、物体に対し、ストリートビュー上の現在の位置から奥行き情報のある地図と高度な物体認識、それに加え、検索技術を導入した地図を作成ことが不可欠である。たとえば「この空間・現場(あるいは現場空間)の中に電球はいくつあるか」「あの現場のメーターの状態はどうか」といった具体的な問いに対し、即座に正確な回答を返す検索システムが構築できれば、現場管理の可能性は大きく拡大する。

本論文では、管理の簡略化と効率化を目的として、データベース(DB)管理システムと大規模言語モデル(LLM)を活用し、現場の物体認識情報、画像情報、座標を入力することで、必要な情報を適切に検索・回答できるシステムの構築における基礎検討を試みる。

2. 関連研究

本研究のシステムは、「空間データベース(Spatial DB)」「埋め込み(Embedding)」「大規模言語モデル(LLM)による自然言語QA/推論」という三つの独立した技術領域を統合するものである。これにより、自然言語の指示に基づき、産業空間における情報を活用した高効率な推論および探索決定を実現し、現場作業の効率化を目指す。

関連研究として、まず空間データベースの活用に焦点を当てたアプローチが挙げられる。Jacobら^[2]は、オフライン環境の位置情報サービスを対象に、SQLiteとSpatialLiteを用いてモバイルデバイス上での空間データクエリの性能を評価した。この研究では、円形クエリやジオワンドクエリを用いてPOI(関心地点)を検索しているが、性能はデータセットの規模に大きく依存

し、都市のような小規模データセットにのみ適していると結論付けている。さらに、このアプローチは事前に定義された POI カテゴリの検索に限定されており、自然言語による柔軟なクエリには対応できないという課題があった。

一方、LLM を自然言語による柔軟なクエリ解釈に応用する研究も存在する。Maletić ら^[3] は、自律型無人航空機 (UAV) が自然言語の指示を解釈し、効率的に目標物を探索するフレームワークを提案した。この研究では、UAV のカメラでリアルタイムに物体を認識し、検出されたオブジェクトと空間的文脈から LLM が意味的推論を実行することで、確率の高い探索領域を優先的に探索する。しかし、この研究はリアルタイムのナビゲーションと物体位置の特定に主眼を置いており、取得した情報をデータベースとして蓄積・管理するような、長期的な情報活用には適していない。

また、現場管理の手法として 360 度カメラを用いたストリートビュー形式での記録も存在する^[4]。この方法は空間全体を一枚の画像として視覚的に分かりやすく記録できるものの、画像の数が増大し、特定の情報を探索することが困難になるという欠点を持つ。第 1 節でも述べたように、FARO に代表される 3 次元スキャナーや LiDAR を用いる手法は、高精度な点群データを取得できる一方で、データ処理の負荷や導入コストが高い。さらに、専門的な知識がない作業員にとっては、機器の操作やデータの可視化が困難であるという運用上の課題も抱えている。

これまでの議論で明らかになったように、既存技術は「クエリの柔軟性」「情報の長期的管理」「運用のコストと効率」のいずれかにおいて課題を抱えており、産業現場のような特定の空間を対象とした長期的かつ安定的な管理には適していない。そこで本研究では、これらの課題を包括的に解決することを目的とする。提案手法は、画像から得られる物体認識結果と位置情報を空間データベースに格納し、この構造化されたデータに対して LLM を用いることで、自然言語による高効率な探索・検索を実現する新たなフレームワークである。

3. 提案手法

本章では、3 次元空間内の物体に関する自然言語での対話的な質問応答を実現するためのシステムを提案する。本システムは、大規模言語モデル (LLM) の言語解釈能力と空間データベースの厳密な計算能力を融合させたハイブリッドアーキテクチャを特徴とする。システムの処理フローは Fig.1 に示す通りであり、以下でその詳細を説明する。

3.1 概要

3.1.1 システム概要

本研究では、3 次元空間内に配置された物体に関する自然言語での対話的な質問応答を実現するシステムを提案する。従来のキーワード検索や定型なデータベースクエリでは、利用者が持つ曖昧な空間認識や多様な言語表現に対応することが困難であった。この課題に対し、本システムでは LLM の高度な言語解釈能力と、空間データベースの厳密な計算能力を融合させたハイブリッドアーキテクチャを採用する。さらに、意味的類似度を用いた信頼性判定を組み込むことで、存在しない物体に関する誤った応答を防ぎ、システムの

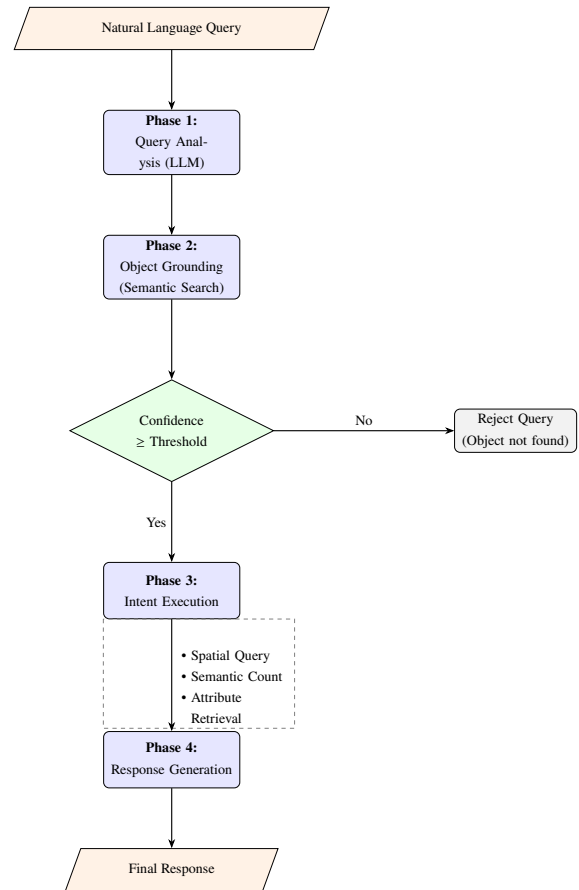


Fig. 1 Processing Flow of the Proposed System

頑健性と信頼性を高めている。これにより、利用者は専門的なクエリ言語を習得することなく、直感的な対話を通じて正確な空間情報を取得することが可能となる。

3.1.2 システム処理のフロー

本システムの処理は、ユーザーからの自然言語クエリ入力を起点とし、大きく分けて「①クエリ解析」「②オブジェクト接地と検証」「③意図実行」「④応答生成」の4つのフェーズで構成される。

3.2 各フェーズの詳細

3.2.1 フェーズ 1: 自然言語クエリの解析 (Query Analysis)

- **処理内容:** ユーザーから入力された自由形式の質問文を、まず LLM に入力する。LLM は、Few-shot プロンプティング^[5]の手法に基づき、質問文からユーザーが実行したい処理の種別である「意図 (mode)」と、その対象となる「対象物 (object)」を抽出する。そして、これらの解析結果を後続の処理モジュールが機械的に解釈可能な構造化データ (JSON 形式) に変換する。例えば、「一番近いドライバーはどこ?」というクエリは、`[{"object": "ドライバー", "mode": "closest"}]`のように変換される。
- **設計理由:** 人間が扱う曖昧な自然言語と、プログラムが扱う厳密な命令との間の仲介を目的とする。LLM の役割を言語の解釈と構造化に限定することで、後続処理の信頼性を担保しつつ、多様な質問形式への柔軟な対応を可能とした。

3.2.2 フェーズ 2: オブジェクトの接地と検証 (Object Grounding and Validation)

- **処理内容:** フェーズ 1 で抽出された物体名が、データベース内のどの実体に対応するかを特定 (接地) し、その確からしさを検証する。
 1. **探索:** Sentence Transformer モデルを用いて物体名を意味ベクトルに変換し、データベース内に事前に計算された全物体のベクトルと **コサイン類似度** を比較し、最も類似度が高い候補を探索・特定する。
 2. **検証:** 次に、選出された候補の類似度スコアと、事前に定義した **信頼度閾値** (例: 0.6) を比較する。スコアが閾値を超えた場合、オブジェクトは正しく接地されたと判断する。スコアが閾値に満たない場合、システムは「ユーザーが質問した物体はデータベース内に存在しない」と判断し、処理を中断してその旨をユーザーに通知する。
- **設計理由:** このフェーズの目的は、誤った応答の生成を未然に防ぎ、システムの事実的正確性を保証することにある。類似度検索は、存在しない物体 (例: 「猫」) について質問された場合でも、最も「意味的に近い」とされる無関係な物体 (例: 「ソファ」) を返してしまう。閾値による検証ステップを設けることで、このような確信度の低い接地を防ぐことができる。これにより、システムが誤った情報をも事実であるかのように提示する「ハルシネーション」的な振る舞いを抑制し、全体としての信頼性を大幅に向上させている。

3.2.3 フェーズ 3: 意図に基づいた処理実行 (Intent Execution)

- **処理内容:** フェーズ 1 で特定された「意図 (mode)」に応じて、接地されたオブジェクト情報を利用し、具体的な処理を実行する。mode の値に基づき、実行する処理が動的に選択される。
- **設計理由:** 処理の正確性と信頼性の保証が最大の目的である。各タスクに特化した決定論的な手法を用いることで、常に検証可能で一貫した結果を保証する。
 - **空間クエリ (closest, nearby):** SpatialLite データベースに完全に委譲し、SQL クエリ内で 3 次元ユークリッド距離を直接計算・ソートする。これにより、計算の高速性と数学的な正確性を両立する。
 - **意味的計数 (count):** ターゲットのベクトルとデータベース内の全物体のベクトルとの類似度を計算し、事前に設定した計数用閾値以上のものを数える。これにより、「照明」といった抽象的なクエリに対し、意味的に合致する複数の異なる物体を包括的に数えることが可能となる。
 - **属性取得 (location, show_image):** データベースから座標や画像パスといった静的な属性情報を直接取得し、情報の正確性を保証する。

3.2.4 フェーズ 4: 応答生成 (Response Generation)

- **処理内容:** フェーズ 3 で得られた計算結果、あるいはフェーズ 2 で処理が中断された場合はその旨を示す情報を、自然言語のテンプレートに埋め込み、最終的な回答としてユーザーに提示する。
- **設計理由:** データ処理ロジックとユーザーへの最終的な情報提示を分離するためである。これにより、応答形式の変更 (例: テキストから音声へ) が、システムのコアロジックに影響を与えることなく容易に行える。

4. 検証

4.1 概要

本章では、第 3 章で提案した手法の有効性を検証することを目的とする。本検証では、システムに対してそれぞれの意図 (mode) で、Chat-GPT に各 20 問ずつの質問を作成。それらを検索クエリとして入力を使い、それに対し正解率を算出し回答の性能を評価する。ここで、本研究における評価用の質問を、以下の 4 種類に分類する。

- **場所の特定:** 特定の物体がどこにあるかを問う質問。(例: ○○はどこにあるか。)
- **相対的な位置関係:** ある物体を基準として、その周辺にある物体を問う質問。(例: △△の近くにあるものは。)
- **画像の提示:** 特定の物体の画像の提示を要求する質問。(例: 物体の写真を見せて。)
- **個数の計数:** 特定の物体が空間内にいくつ存在するかを問う質問。(例: ◇◇は何個あるか。)

上記の質問分類は基本的な空間検索に欠かせない質問であり、これらに基づいた具体的な質問例を **Table 1** に示す。

Table 1 Example of questions asked for evaluation

質問 1	How many Fluorescent light are there?
質問 2	Where is the door?
質問 3	What is closest to the TV?
質問 4	Which fluorescent light is near the dartboard?
質問 5	Show me the fridge.

4.2 実験条件

4.2.1 実験環境

本検証の目的は、提案手法による自然言語での質問応答の精度を評価することにある。そのため、地図生成の精度が応答結果に影響を与えることを避けるため、以下の仮定を置く。すなわち、Visual SLAM 等によって空間地図はすでに構築されており、地図内の各物体について名称、3次元座標、および画像の 3 情報がデータベースに予め格納されているものとする。

本実験は、中央大学理工学部梅田研究室を対象空間として実施した。この空間内に存在する全 52 個の物体情報を DB に格納している。対象空間における物体の配置図を **Fig.2** に、DB に格納した物体の一覧と個数を **Table.2** にそれぞれ示す。図中の番号と表の番

号は一対一で対応している。なお、蛍光灯のように複数存在する同一物体については、出力結果での混同を避けるため、DB 内で個別の名称を割り振った。具体的には、蛍光灯を Fig.2 の番号が小さい順に fluorescent light1, fluorescent light2 と命名した。

4.2.2 検証方法の拡張

前節では、データベース (DB) に格納する物体名を人手で定義した。しかし、実用的なシステムを想定した場合、カメラで物体を認識し、その名称を自動で判別して DB に格納するプロセスが不可欠である。

そこで本研究では、人手で名称を定義した場合の検索に加え、AI が自ら物体を認識して名称を定義し、検索に利用する手法を新たに提案する。この提案手法の有効性を評価するため、本論文では以下の2つの検証を行う。

- 検証 1: 人手による物体名の定義。
- 検証 2: AI による物体名の定義。

なお、検証 2 で用いる AI には、画像内の物体を認識してキャプションを生成する能力を持つ BLIP-2 を採用する。

4.2.3 使用するモデルと閾値について

本システムは現場管理領域での運用を目的としており、その要件に基づき、DB, Sentence Transformer, LLM を慎重に選定した。まず、DB には現場管理特有の限定的なデータセットを効率的に扱うため、軽量な SQLite に空間拡張機能である SpatialLite を追加して採用した。これにより、位置情報を含む専門的なデータを柔軟に管理できる^[2]。次に、文章のベクトル表現を生成する Sentence Transformer には、高速かつ軽量でありながら文の意味理解に優れる sentence-transformers/all-MiniLM-L6-v2 を用いた。これは、システムの応答性を高め、迅速な意味検索を実現するためである。そして、これらの技術を統合し、自然言語での対話と推論を担う中核として、商用利用が可能で高い汎用性を備えた LLM である LLaMA2 を導入した。この構成により、現場からの多様な問い合わせに対して、的確かつ迅速に対応可能なシステムを構築する。

また閾値に関して、

4.3 検証 1: 人の手による名前の定義

空間クエリに対する、正答率は 35 % (7 問正解)、意味的計数の正答率は 30 % (6 問正解)、属性取得の正答率は 20 % (4 問正解) という結果になった。しかしながら、4.1 節で示した Table 1 のような質問には適切に答えられたという結果になった。

4.4 検証 2: AI の手による名前の定義

検証 1 のときとは違って空間クエリに対する、正答率は 25 % (4 問正解)、意味的計数の正答率は 10 % (2 問正解)、属性取得の正答率は 15 % (3 問正解) という結果になった。

5. 考察

5.1 検証 1 について

検証 1 として、各意図 (mode) に対する質問応答実験を行った結果、システムの正答率は低い水準に留まった。この低い正答率の主要因は、LLM に与えた Few-shot プロンプティングにあると考えられる。本稿で LLM に与えたプロンプト内の few-shot 例を Fig. 3 に

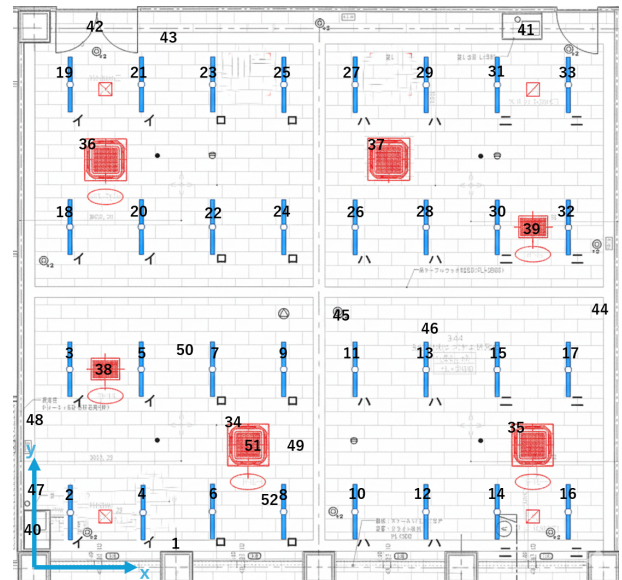


Fig. 2 3D layout diagram of objects



Fig. 3 Few-shot prompting content

示す。

実験の結果、特に「Which objects are within 2 meters of the dartboard?」や「Show all furniture sorted by distance from the door.」といった、Fig. 3 の例とは構造が異なる複雑な質問において、LLM は意図を正しく解析できなかった。この結果は、LLM が提示された few-shot 例の表層的なパターンに過剰適合 (overfitting) し、未知の形式の質問に対する汎化性能 (generalization ability) が不足していたことを示唆している。

5.2 検証 2 について

検証 2 の結果、正答率は検証 1 をさらに下回った。この性能低下の主な原因は、本検証で導入した画像キャプション生成モデル BLIP-2 によって、データベース内の物体名がより詳細な記述的テキストに置き換えられたことにある。

具体的には、2 種類の問題が確認された。第一に、オブジェクトの属性に関する記述の過剰な詳細化である。例えば、DB 上の一部の「fluorescent light」は、BLIP-2 によって「a white led light tube with a white light bulb」のような詳細なキャプションに変換された。これにより、ユーザーが「fluorescent light」と質問した際のクエリベクトルと、DB 内のキャプションベクトルとの間に語彙的な乖離 (かいり) が生じ、意味的類似度スコアが低下した。これが検索精度の悪化と、それに伴う正答率の低下に直接繋がったと考えられる。

第二に、空間関係の誤認識である。天井に設置されている「fluorescent light」が、「a white tube with a white hose attached to a wall」と壁にあるかのように記述される事例が確認された。本システムの検索はテキストの意味に基づいて行われるため、検索対象となるキャプション自体が不正確な空間情報を含んでいると、オブ

Table 2 Objects stored in the database

番号	名前	個数
1	datrboard	1 個
2~33	fluorescent light	32 個
34~37	Large air conditioner	4 個
38~39	Small air conditioner	2 個
40~41	Sink	2 個
42	Entrance door	1 個
43	Shoe rack	1 個
44	Bookshelf	1 個
45	Small Refrigerator	1 個
46	Blackboard	1 個
47	Large Refrigerator	1 個
48	Television	1 個
49~50	Sofa	2 個
51	Table	1 個
52	Printer	1 個

ジェクトの特定が著しく阻害される。これもまた、正答率を低下させる一因となったと考えられる。

6. 結言

本研究では、産業環境のリモート管理を目的とし、物体画像データベースと大規模言語モデル (LLM) を連携させた対話型検索システムの基礎的検討を行った。提案システムの有効性を評価するため、Table 1 に示す基本質問 20 項目に対する正答率を計測した。

実験の結果、Table 1 に示すような基本質問に対しては良好な結果が得られたものの、2つの主要な課題が明らかになった。第一に、LLM の汎化性能の限界である。「Which objects are within 2 meters of the dartboard?」や「Show all furniture sorted by distance from the door.」のような、プロンプトの few-shot 例 (Fig. 3) とは構造が異なる複雑なクエリに対しては、LLM は意図を正しく解析できず応答に失敗した。第二に、画像キャプション生成モデル (BLIP-2) を用いた際の性能低下である。BLIP-2 が生成した記述的すぎるキャプションは、ユーザーのクエリとの語彙的な乖離を生んだ。これに加えて空間関係の誤認識も発生したため、人が物体名を定義した検証 1 よりも正答率が低下する結果となった。

今後の展望として、本研究を発展させる3つの方向性が考えられる。第一に、「2メートル以内」や「距離順」といった制約を含む複雑な構造の質問も解釈・実行できるよう、LLM の推論能力を向上させる。第二に、類義語処理や検索結果のフィルタリング機能を導入し、さらに多様な自然言語表現に対応可能なシステムの頑健性を高める。第三に、物体の位置情報だけでなく、「状態」(例：蛍光灯が点灯しているか) や「機能」(例：テレビで映像を再生できるか) といった動的な属性情報も扱えるシステムへと拡張し、より実用的なりモート管理への応用を目指す。

参考文献

References

- [1] Ye, Xi, and Greg Durrett. "The unreliability of explanations in few-shot prompting for textual reasoning." *Advances in neural information processing systems* 35 (2022): 30378-30392.
- [2] R. Jacob, S. Smithers and A. C. Winstanley, "Performance evaluation of storing and querying spatial data on mobile devices for offline location based services," *IET Irish Signals and Systems Conference (ISSC 2012)*, Maynooth, Ireland, 2012, pp. 1-6, doi: 10.1049/ic.2012.0217.
- [3] M. Maletić, M. Peti, T. Petrović and S. Bogdan, "Spatial-Semantic Reasoning using Large Language Models for Efficient UAV Search Operations," *2025 European Conference on Mobile Robots (ECMR)*, Padova, Italy, 2025, pp. 1-8, doi: 10.1109/ECMR65884.2025.11163229.
- [4] Eiris, Ricardo, Jing Wen, and Masoud Gheisari. "iVisit: Digital interactive construction site visits using 360-degree panoramas and virtual humans." *Construction Research Congress 2020*. Reston, VA: American Society of Civil Engineers, 2020.
- [5] Li, Yongchang, et al. "Street View Imagery (SVI) in the built environment: A theoretical and systematic review." *Buildings* 12.8 (2022): 1167.