

360度カメラ画像内での物体認識及び言語モデルを用いた検索可能な ストリートビューシステムの基礎検討

○朝木 直道 (中央大学), 椎野 丞ノ進 (中央大学), 倉持 光希 (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 教昇 (中央大学), Sarthak Pathak (芝浦工業大学)

Fundamental Study on Object Recognition in 360-Degree Camera Images and a Searchable Street View System Using a Language Model

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

Abstract: In industrial environments, accurately understanding on-site conditions is critically important for ensuring safety and operational efficiency. However, traditional methods often require experts to physically visit the site, resulting in increased labor and cost. This study proposes a fundamental system that utilizes 360-degree camera images to recognize objects in the scene while automatically generating a 3D map. By integrating Visual SLAM for spatial mapping with YOLO-based object recognition, our proposed system attempts to estimate system estimates the precise 3D positions of objects from 360 degree video input. Furthermore, future work includes the development of a street-view-style interface that enables users to search for objects in the generated 3D space using natural language queries, powered by a large language model. This approach allows for intuitive and remote inspection of on-site conditions, offering improvements in efficiency, accuracy, and safety. This paper presents initial results using monocular depth estimation and discusses the potential for practical application in real-world environments.

1. 緒言

近年、工場や作業現場では、現地の状況を的確に把握することが求められているが、その多くは専門の作業者が現場に赴く必要があり、人的負担やコストが大きな課題となっている。特に物体の位置や状況を空間的に正確に把握することは、安全管理や作業計画の面でも重要である。しかし従来の手法では、リアルタイム性や精度に限界があり、遠隔地からの把握は困難であった。これを踏まえ、本研究では360度カメラ映像を活用した自動かつ高精度な3次元地図生成の技術に着目する。

近年、多様な三次元復元技術が提案されており、それらは主に、レーザ計測 (LiDAR) を用いる手法、RGB-D センサを活用する手法、そして画像処理に基づく手法の3つに大別される。

レーザ計測によるアプローチは、他の手法と比べて精度の安定性が高いという利点がある。一方で、レーザ計測機器は非常に高額であり作業員も扱いにくいこと、導入コストおよび運用コストがかさみやすいこと、また多くの場合、車載型や据え置き型であることから、使用可能な環境が制限されやすいという欠点もある¹⁾。

RGB-D センサを利用する手法は、装置の価格が比較的安価であるというメリットがあるが、赤外線を用いているため、計測できる範囲に制約があるというデメリットを持つ²⁾。

画像処理を用いる手法では、市販されているカメラを用いて対象物を複数の視点から撮影し、それらの画

像を写真測量の原理に基づいて解析することで、対象物の三次元形状を復元する方法である。一方で、画像処理による手法は、画像上の特徴点を元に画像処理を進めるため、特徴点が少ない平坦な形状には不向きであるなどの短所がある。本研究では、現場への実装可能性と広範囲の計測を考慮して、画像処理による手法に着目した。

カメラには周囲の映像すべてが撮影できる360度カメラを用いる。全天球カメラは周囲360度を一度に撮影できるカメラであり、複数台のカメラを組み合わせずべての方位を撮影できるようにしたものや、魚眼カメラを前後に2台組み合わせたものがある。360度カメラはその画角の広さから、多くの情報を得ることができるため、物体検知、位置推定、構造推定の効果を高めることができる³⁾。

360度カメラを用いたストリートビュー地図作成の研究も行われている⁴⁾。しかし、物体情報が含まれていない。

既存の研究として、360度カメラを用いた意味地図推定を行う研究がある⁵⁾。全天球ステレオカメラと2D意味地図を用いることにより、ロボットの位置姿勢推定を行った。しかし、ステレオのためカメラを2第利用している点や、2次元地図であるため空間的な地図作成は行っていない。単眼カメラでは、ステレオのような奥行き情報が得られないため、同一物体の複数インスタンスの区別が難しく、時間的に追跡することも困難であるという問題がある。

そこで本研究では、単眼 360 度カメラから現場の物体情報を含んだ 3 次元地図を作成するシステムの開発を目指す。将来的に、Open Vocabulary 物体認識による地図内物体の検索も出来るようにした。

2. 提案手法

2.1 概要

提案手法では、360 度カメラから得られた映像から作成

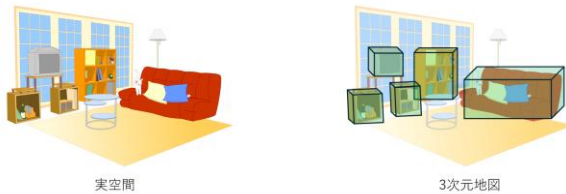


Fig. 1 Overview of the final target system

した 3 次元地図に物体情報を持たせる。本手法の流れを図 2 に示す。はじめに、単眼カメラから得られた映像を Visual SLAM にかけて、3 次元点群地図を取得する。次に YOLOE による物体認識により、画像内の物体情報を得る。この YOLOE と単眼深度推定 Depth-Anything-V2 により、物体の 3 次元バウンディングボックスを得る。これを地図上に重ね合わせることでより物体情報を含んだ 3 次元地図を得ることができる。以下、それぞれの手法について説明する。

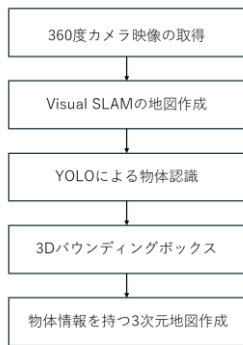


Fig. 2 Flow of the proposed method

2.2 Visual SLAM

本研究では、stella_vslam_dense を用いる⁶⁾。stella_vslam_dense は、コミュニティ版 OpenVSLAM である stella_vslam を基盤とする特徴点ベースの Visual SLAM に対し画像を対象とした PatchMatch-Stereo による密対応推定を追加した拡張実装である。モバイル GPU を想定した低遅延処理で小型 UAV 搭載 360° カメラの動画に適用できる。図 3 に示すような点群地図を得ることが可能である。動画内の処理フレームごとのカメラの位置姿勢などのデータを保存する。

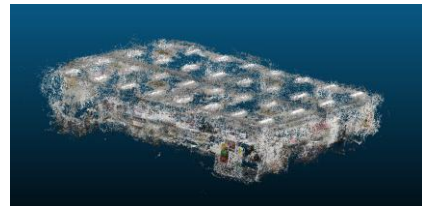


Fig. 3 stella_vslam_dense

2.3 物体認識

物体認識には YOLOE を用いる⁷⁾。YOLOE とは、オープンボキャブラリーの検出とセグメンテーションのために設計された、ゼロショットでプロンプト可能な新しい YOLO モデルである。また、YOLOE は、速度と精度への影響を最小限に抑えながら、最先端のゼロショット性能を実現可能である。本研究では、YOLOE ベースのセグメンテーションモデルを適用し、物体のバウンディングボックス・マスク・クラスラベルを取得する。検出を安定化させるために、信頼度の低い検出をカットした他、多数のヒューリスティック条件を適用しノイズ検出や水平ストライプのような誤検出の除去を行った。

2.4 単眼深度推定

単眼深度推定には Depth Anything V2 を用いる⁸⁾。Depth Anything V2 は、DPT 系のエンコーダ-デコーダ構成を維持しつつ、大規模教師モデルを合成データで学習し、その教師で実写 6,200 万枚超を擬似ラベル化 → 擬似ラベル実写のみで学生モデルを学習する Teacher-Student 戦略により、細部の復元性と頑健性を同時に高めた単眼深度推定モデルである。本研究では、単眼カメラ画像の 3 次元化のため用い、画像全体の相対的な奥行き分布を得る。

2.5 3D バウンディングボックス

2.5.1 流れ

本手法は正距円筒画像から、各物体 3D 直方体 (upright Oriented Bounding Box) を推定して可視化・エクスポートする。YOLO ベースのインスタンスセグメンテーションと単眼深度推定モデル Depth Anything V2 を統合することで、物体マスクからサンプリングした深度点を正距円筒座標から 3D 座標に変換し、外れ値を除去した上で分位点に基づくアップライト OBB をフィッティングすることで、物体の 3D 形状を近似する。この結果を元に入力画像にバウンディングボックスを表示するとともに 3 次元データを PLY/OBJ 形式で出力する。

2.5.2 3D 座標変換

2.3 節の YOLOE の物体認識の誤検出除去の結果残った物体について、マスク領域の画素座標と深度のサンプリングを行う。正距円筒座標から 3D 直交座標系

(X,Y,Z)へ変換し、点群の生成をする。外れ値除去には中央値+MAD (Median Absolute Deviation) 法を利用する。物体のマスク内の深度列 Z_k の中央値 m と平均値 MAD を用い、式

$$|Z_k - m| \leq tMAD \quad (1)$$

を満たす点のみ残すことでロバスト化する ($t=2.5$)。

2.5.3 OBB 推定

重力方向 (画像の Y 軸) に直交する平面 XZX 上で主方向を推定し、Yaw の回転行列

$$R_y(-\theta) = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix} \quad (2)$$

で中心化点を回し、ローカル座標を得る。ローカル点の各軸に対し、分位点で内側を採用することで 8 頂点を生成し、重心を足す。

3. 実験

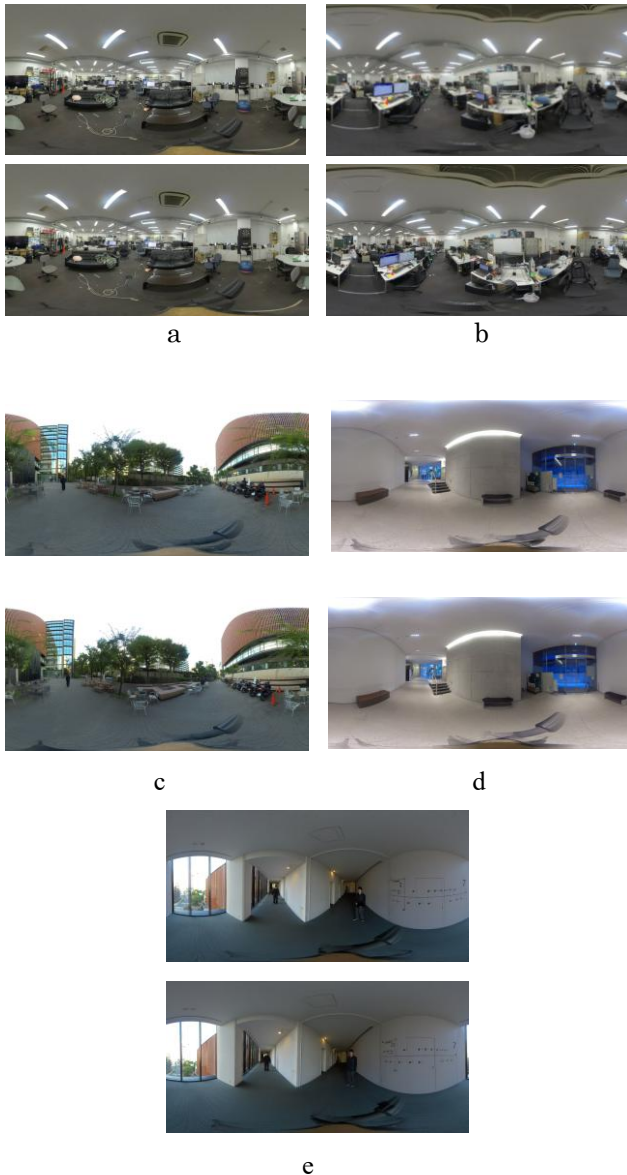


Figure. 4 Location

提案手法の有用性を検証するために、正距円筒画像から得られる 3D バウンディングボックスの位置推定をする実験を行った。本実験では、360 度カメラ、Insta360pro2 を用いる。

撮影は 5 か所で行い、Visual SLAM の数フレームごとの処理を想定し、それぞれの場所で上下に 40cm ずらした位置から撮影した 2 枚の画像を用意した。5 か所の画像を図 4 に示す。得られた 3 次元地図が正しいのかを評価するために、上下の画像から得られた地図を画像間のスケール調整して重ね合わせ、バウンディングボックスの体積の 50% が重なっている場合を overlap とした。これを実験 1 とする。本手法により 3D バウンディングボックスを表示したものを図 5 に示す。

Table. 1 Experiment 1

	a	b	c	d	e
upper camera	29	27	13	12	3
under camera	30	29	12	10	3
overlap	27	25	10	9	3

表 1 より、上カメラよりも下カメラの方が多く物体を検出していることが分かる。これは、上カメラの角度だと物体を正面から写すため、YOLOE による認識がされやすいからだと思われる。続いて、重ね合わせについて、全体で 85% を超え高い精度で物体が重なっていることがわかる。Visual SLAM ではこれが映像中に連続して行われるため、異なった画像から映同じ物体を同じとして処理が可能であると予想される。一方で、物体数が多い a,b では overlap とならない場合があるので物体が多いときの誤検出に課題が残る。



次に、物体の多い a で撮影箇所実測した物体距離と作成した地図の距離を比較する評価実験を行った。単眼深度推定による距離にスケールがないため、実測値、上下画像から得られた値の比を比較して評価する。これを実験 2 とする。

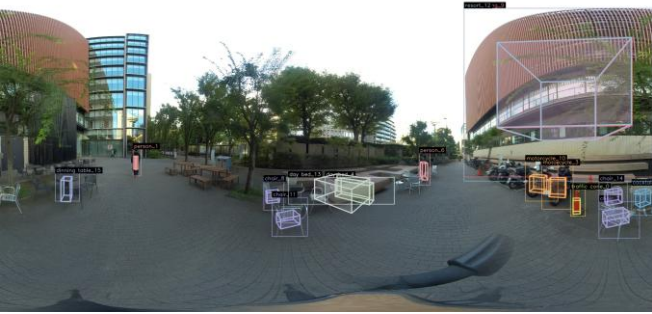
Table. 2 Scale-free distance comparison

	measured value (m)	upper camera	under camera
Object-Object	1.80	83	50.5
	2.42	90.3	58.4
	1.75	84.5	52.1
	3.73	193.6	135.2
	2.50	80.2	50.5
	0.42	16.3	13.4
	1.35	51.7	37.2
	1.06	39.4	25.9



Table. 3 measured value as the standard

	measured value(m)	upper camera	under camera
Object-Object	1.8	1.8	1.8
	2.42	1.96	2.08
	1.75	1.83	1.86
	3.73	4.12	4.81
	2.5	1.74	1.80
	0.42	0.35	0.48
	1.35	1.12	1.33
	1.06	0.85	0.92



実験の結果の表 2,3 から、実測値を基準としてみた場合、上画像では平均して 0.412 倍、下画像では 0.278 倍のスケールの値であった。ばらつきについては、上画像でも下画像でも平均から 25%近く異なる距離が得られている。単眼カメラの特性として画像スケールを直接観測できないため、幾何学的には絶対距離を一意に決定できないことから、このばらつきは単眼推定で避けられない構造的な不確実性を示す。しかし、一定の誤差を単眼距離測定はおおよそその距離傾向把握には有効である可能性を示す。



4. 結論

本研究では、単眼 360 度カメラから物体情報を含む地図作成の手法を提案した。評価実験から、3D バウンディングボックスの位置の精度が有用であると示した一方、距離計測に課題が残った。将来的には、2.2 節の stella vslam dense の点群地図で行い物体情報を含む 3 次元地図を作成し、言語モデルとの融合を目指す。

Figure. 5 3D Bounding Box

参考文献

- [1] 中村 裕幸: 地上型レーザスキャナによる森林情報のデジタルドキュメント化, 2013 年度精密工学会秋季大会学術講演会講演論文集(2013)
- [2] 岡田 伸也, 鈴木 智, 石井 崇大, 藤澤 陽平, 飯塚 浩二郎, 河村 隆: RGB-D カメラを用いた移動ロボットののための 3 次元環境地図構築, ロボティクス・メカトロニクス講演会講演概要集(2013)
- [3] 小田巻 誠: 360 度カメラ THETA とその応用技術, *The Journal of the Institute of Image Electronics Engineers of Japan* Vol.51 No.4 (2022)
- [4] 小川 将範, 山崎 俊彦, 相澤 清晴: 適応的なサンプリングを用いた全天球ストリートビュー動画のためのハイパーラプス映像の生成手法, *電子情報通信学会論文誌 D* Vol.J101-D No.7 pp.1052-1060 (2018)
- [5] 小野関 祐介, 入山 真伍, 小笠 遼太, Sarthak Pathak, 梅田 和昇: 全天球ステレオカメラでの物体認識情報を用いた意味地図内位置姿勢推定, ロボティクス・メカトロニクス 講演会 2024
- [6] H. Surmann, M. Thurow, D. Slomma, N. Digakis, and N. Voigt, “PatchMatch-Stereo-Panorama, a fast dense reconstruction from 360° video images,” in *Proc. IEEE SSRR*, 2022, pp. 366–372
- [7] Wang, A., Liu, L., Chen, H., Lin, Z., Han, J., Ding, G., “YOLOE: Real-Time Seeing Anything,” *arXiv:2503.07465*, 2025.
- [8] Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., “Depth Anything V2,” *arXiv:2406.09414*, 2024.