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

Localization in 2D Semantic Map Using Object Recognition Information from Spherical Stereo Camera

○学　小野関　祐介 (中央大学)　　入山　真伍 (中央大学)　小笠　遼太 (中央大学)
Sarthak Pathak(中央大学)　正　梅田　和昇 (中央大学)

Yusuke ONOZEKI, Chuo University, onozeki@sensor.mech.chuo-u.ac.jp
Shingo IRIYAMA, Chuo University
Ryota OGASA, Chuo University
Sarthak PATHAK, Chuo University
Kazunori UMEDA, Chuo University

In this paper, we propose a system for localization using semantic information and distance between camera and each object from spherical stereo camera. Recently, demand for a servicing robot has been increased due to understaff for work and increasing industrial accidents in aging infrastructures. Thus, our final goal is to make general people be able to operate robots due to workforce diversification. In this paper, we aim to locate a robot more accurately by using distance information. In our method, we locate a robot from a set of object center-of-gravity points on a 2D semantic map prepared in advance and a set of corresponding object center-of-gravity points obtained from a spherical stereo camera. Through the experiments, when all matching between the point clouds was successful, a robot's rough localization was accurate.

Key Words: Spherical camera, Stereo camera, Localization, Semantic, 2DMap

1　序論

近年，人手不足の解消や労働災害回避などのために，インフラの老朽化修理などに用いる屋内作業ロボットの需要が急増している．労働力の多様化のために，作業ロボットに関して知識がない人でもロボットに直感的かつ的確な指示を出せるシステムの開発が有効である．直感的な指示を出すためには，ロボットの位置情報を，図 1 に示すように，どの「もの」の近くにあるかなどの感覚的情報として操作者にフィードバックする必要がある．位置姿勢推定には GPS を用いた手法が一般的であるが，屋内環境では電波が届かず機能しない場合がある．また他の位置姿勢推定手法として，LiDAR を用いた手法が挙げられるが，LiDAR は高価，重量が重い，給電コストが高いなどの欠点がある [1]．そこで Uygur[2] らは，1 台の様々な視野角のカメラから得られる「もの」の方位角情報と事前に準備した 2D 意味地図（「もの」の情報を含む地図）を基にロボットの 2 次元自己位置姿勢推定を行い，視野角が広いカメラの方が広範囲の物体を認識可能なため，位置姿勢推定精度が向上するといった結論を導いた．また Pathak[3] らは，視野角が全方位である単眼全天球カメラを用いて，Uygur らと同様の実験を行った．しかし単眼全天球カメラから得られる 1 枚の画像を用いているため，1 点での位置姿勢推定は困難であり，またその精度は時間とともに向上していくものの不十分であった．そこで本研究では，カメラから物体までの距離が得られる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精度の向上を目指す．

2　提案手法

2.1　概要

提案手法では，全天球ステレオカメラから得られた画像と事前に準備した 2D 意味地図情報を用いて，カメラの 2 次元自己位置姿勢 $\boldsymbol{s} = [x, y, \theta]^T$ を推定する．全天球カメラを用いる理由は，幅広い視野の方がより広範囲の物体を認識でき，精度向上が見込めるためである．本手法の流れを図 2 に示す．はじめに全天球カメラで得られた画像から物体検出を行う．次に，2 台の全天球ステレオカメラ間で認識された物体のマッチング及びラベリングを行う．そして，Bounding Box(BBox) で囲まれたラベル付けを行った物体の重心の画像座標を用いて，三角測量の原理により物体重心の 3 次元座標を得る．最後に，2D 意味地図からの点群情報と 3 次元計測で得られた点群情報から，2 次元の自己位置姿勢推定を行う．以下，それぞれの手法を説明する．

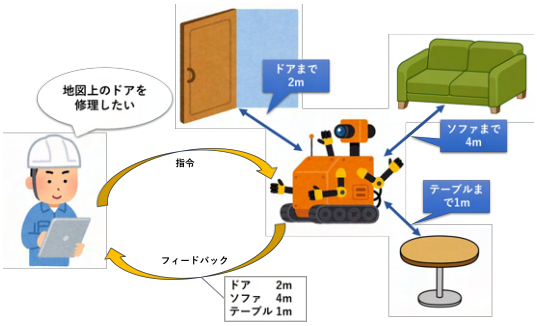

Fig.1 Overview of the final target system

2.2　深層学習を用いた物体検出

物体認識には，YOLOv4(You Only Look Once)[4] を用いる．YOLO で検出された BBox の中心座標及び BBox で囲まれる領域の一部における物体の色相 H，彩度 S 値の平均値を得る．また，認識したい物体が画像端で切れてしまい正しく認識されない，または重複して物体が認識されてしまうという問題があった．そこで，図 3 に示すように正距円筒画像の左端 1/4 を元画像の右端に付加した画像に対して，物体が検知できる範囲を設定し，物体認識を行うことでこの問題を解決した．

2.3　2 台のカメラで認識された物体間のマッチング

2 台のカメラ間のマッチングは，エピポーラ拘束と Accelerated KAZE(AKAZE) 特徴量 [5] を用いて行う．具体的には，図 4 に示すように，左右カメラより得られた画像から検出される同種類の物体に対して，左カメラから検出される物体と右カメラから検出される物体同士を，BBox 内で得られる AKAZE 特徴量に対し，エピポーラ拘束式を考慮しながら総当たり的にマッチングをする．そして，AKAZE 特徴量によるマッチング点の数が最大の組み合わせを左右カメラ間において同一物体と判断する．このマッチング情報を基に，2 台のカメラから検出される同一物体に対してラベル付けを行う．

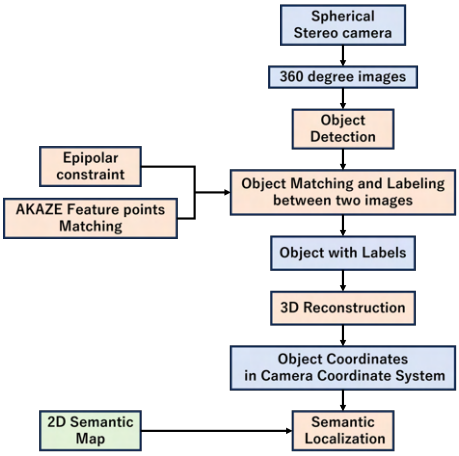

Fig.2 Flow of the proposed method

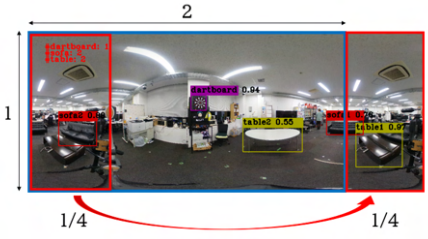

Fig.3 Object detection on a image

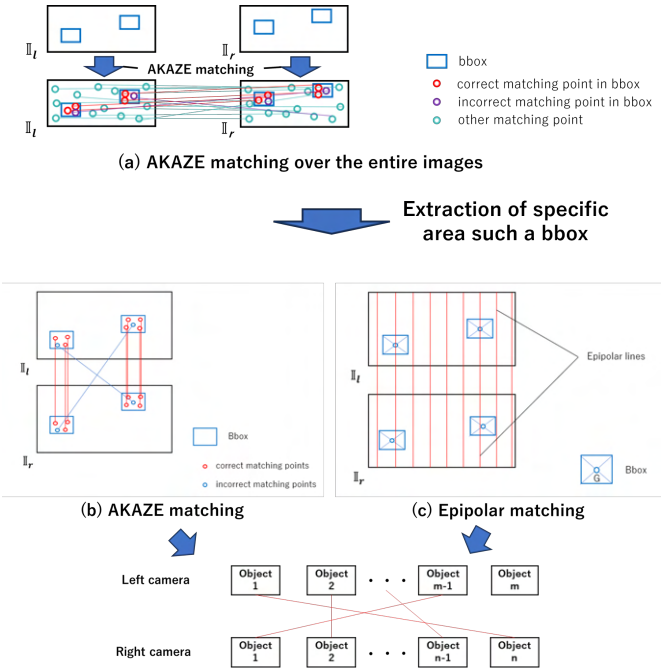

Fig.4 Macthing and labeling of detected objects

2.4 全天球ステレオカメラによる 3 次元計測

全天球ステレオカメラによる 3 次元計測の流れを，図 5 に示す．2.3 節でラベル付けした各々の物体を囲む BBox の画像上の 2 次元重心座標を，単位球面画像上のカメラ座標系で表される 3 次元座標 $\hat{\boldsymbol{x}}$ に変換することで，カメラ中心から物体までの単位方向ベクトルを算出する．また，三角測量の原理により全天球ステレオカメラの左カメラから検出された各々の物体の重心までの距離 d を算出する．以上より，カメラ座標系における各々の物体の 3 次元重心座標 $\boldsymbol{X}$ を，$\boldsymbol{X} = d \cdot \hat{\boldsymbol{x}}$ より算出する．

2.5 カメラの点群と地図データの点群間のマッチング

地図データ $\boldsymbol{M}$ には，各々の物体の種類 $\boldsymbol{c}$，ワールド座標系における 2 次元重心座標 $\boldsymbol{X_w}$，物体全体に対する H，S 値の平均値の 3 種の物体情報が含まれているとする．また全天球ステレオカメラから得られる情報 $\boldsymbol{Z}$ として，2.2 節から 2.4 節までの手法を用いて，地図データと同様の物体情報が得られる．ただしこの場合，各々の物体の重心情報はカメラ座標系を基準としたものである．以上の物体情報を含む，全天球ステレオカメラから得られた点群と事前地図から得られた点群を総当たりさせ，式

$$penalty \equiv p\_obj + p\_h \times \frac{\Delta h}{360} + p\_s \times \frac{\Delta s}{100} + p\_d \times \frac{|\,\|\Delta \boldsymbol{x_c}\| - \|\Delta \boldsymbol{x_m}\|\,|}{\|\Delta \boldsymbol{x_m}\|} \tag{1}$$

に定義するペナルティ関数の値が最小となる組み合わせを見つける．この処理を点群同士が 1 対 1 対応になるまで繰り返すことで，点群間のマッチングを行う．

ここでペナルティ関数の p_{obj}, p_h, p_s, p_d は，それぞれ物体の種類，色相 H，彩度 S, 距離 d に関する penalty 関数全体に対しての重みであり，これらは手動で定める．$\Delta h, \Delta s$ は，それぞれカメラと地図データから得られた点群の各々の H,S 値の差の絶対値である．また，カメラ座標系における検出された物体重心座標 $\boldsymbol{x_c}$，検出されたすべての物体重心の重心座標 $\boldsymbol{x_{G\_camera}}$ に対し，$\Delta \boldsymbol{x_c}$ はこれらの差分，すなわち $\Delta \boldsymbol{x_c} = \boldsymbol{x_c} - \boldsymbol{x_{G\_camera}}$ を表す．更に，ワールド座標系における 2D 意味地図上の物体重心座標 $\boldsymbol{x_m}$，地図上すべての物体重心の重心座標 $\boldsymbol{x_{G\_map}}$ に対し，$\Delta \boldsymbol{x_m}$ はこれらの差分，すなわち $\Delta \boldsymbol{x_m} = \boldsymbol{x_m} - \boldsymbol{x_{G\_map}}$ を表す．

2.6 自己位置姿勢推定

カメラの自己位置姿勢推定を行う流れを，図 6 に示す．2.5 節で得られる対応付けられた点群同士を用いて，カメラの自己位置姿勢推定を行う．自己位置姿勢推定の手法として，Umeyama らの手法 [6] を用いて点群間の平均 2 乗誤差が最小となる 2 次元の回転行列 $\boldsymbol{R}$，並進ベクトル $\boldsymbol{t}$ を算出する．以上より，カメラのワールド座標系における 2 次元位置姿勢 $\boldsymbol{s} = [x, y, \theta]^T$ を推定する．

3 実験

提案手法の有用性を検証するために，カメラから得られる情報を基に，カメラの位置姿勢推定をする実験を行った．実験は，図 7 に示す環境で行った．本実験では，図 6 に示す全天球カメラ，RICOH 社 THETA Z1 を 30cm 間隔で 2 台縦に並べた，全天球縦ステレオカメラを用いた．計測点は表 1 に示す 10 箇所で，1 カ所につき 2 つの姿勢で計測した．また，カメラの地面からの高さに応じて 2 種類の実験を行った．それぞれの実験を，実験 1，実験 2 とする．実験 1，2 において，左カメラの高さはそれぞれ 132cm，144cm，右カメラの高さはそれぞれ 102cm，114cm であった．実験の評価方法として，点群間のマッチングの仕方に応じて得られたデータを，表 2 に定義する correct，incorrect，unable to calculate，All の 4 種類に分割し，unable to calculate を除いた各々の場合において計測値と真値の絶対差分をとり，本手法の評価を行った．

表 3 に示す実験 1 の結果から，点群間のマッチングがすべて成功している場合（correct），x, y, θ 方向の絶対平均誤差がそれぞれ 0.23m，0.44m，11.92deg とロボットの大まかな位置姿勢推定を行うのに十分な精度であると考えられる．

更に，表 4 に示す実験 2 の結果においても，点群間のマッチングがすべて成功している場合（correct），x, y, θ 方向の絶対平均誤差がそれぞれ 0.19m，0.20m，2.79deg と比較的良い精度でロボットの位置姿勢推定が行えている．

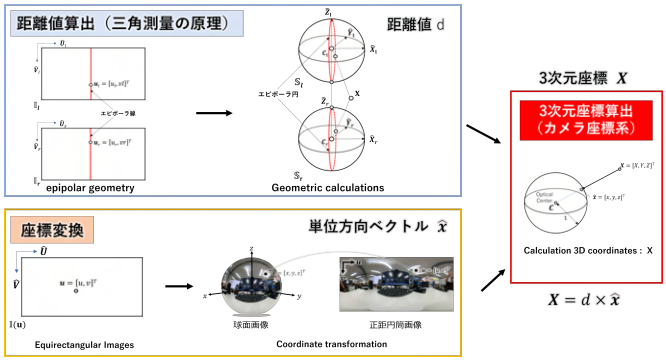

Fig.5 3D measurement

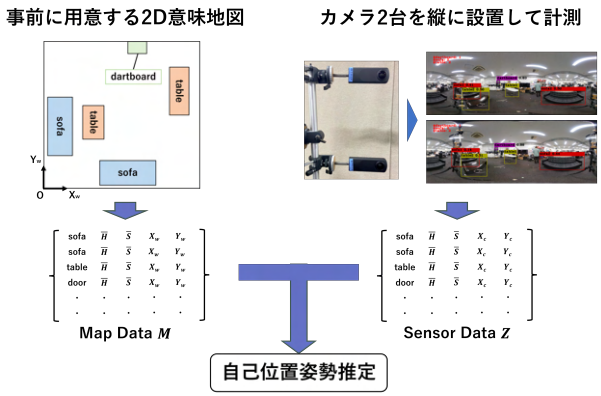

Fig.6 Localization

しかし，どちらの場合も全体のデータ（All）での精度は，点群間のマッチングが 1 つでも失敗している場合（incorrect）の影響を大きく受けるといった課題が存在する．この事から，マッチング成功率が本手法の位置姿勢推定精度に大きく依存すると考察できる．ここで，本実験 1，2 いずれの場合も表 5 に示すように，マッチング成功率は 65%であった．これは，マッチング失敗の組の大半が似た色のソファの組であったことから，点群間のマッチングを行う際に用いるペナルティ関数の色相情報におけるペナルティの値の差が，ソファ間でほとんどなかったためであると考えられる．

実験 1 と 2 の結果を比較すると，カメラ位置の高い実験 2 の方が位置姿勢推定精度が良いことが読み取れる．この理由として，カメラ位置が高い方が物体全体を捉える事が可能なため，撮影された画像から算出される物体重心が実際の物体重心位置とより近くなるからであると考えられる．

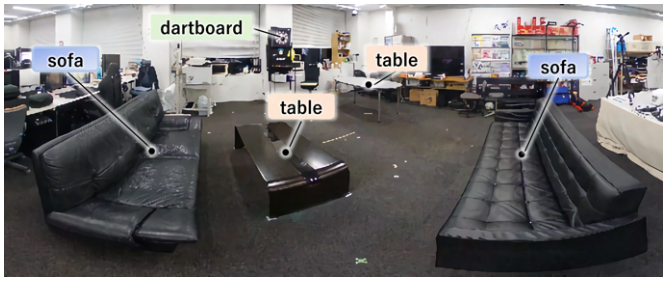

Fig.7 Experimental environment

Table 1 Measurement points

Measurement Points							
place	X[m]	Y[m]	θ [deg]	place	X[m]	Y[m]	θ [deg]
1	1.00	0.50	180	6	1.00	3.50	90
			-90				0
2	2.50	1.50	180	7	3.50	4.50	180
			90				90
3	3.00	2.50	180	8	4.50	4.50	180
			-90				90
4	2.50	3.50	180	9	4.00	1.50	-90
			-90				0
5	1.50	4.50	90	10	4.50	0.50	-90
			0				180

Table 2 Kind of data

カメラからの点群と2Dmapからの点群間のマッチング	
correct	全て成功しているかつx,yの誤差が2m以内
incorrect	1つでも失敗 又は x,yの誤差が2m以上
unable to calculate	1組以下のデータが計算に使用
All	unable to calculateを除いたすべてのデータ

Table 3 Experiment 1

	Mean Δ x [m]	Mean Δ y [m]	Mean Δ θ [deg]	std : x [m]	std : y [m]	std : θ [deg]
All	0.53	0.62	27.99	0.81	0.60	41.85
correct	0.23	0.44	11.94	0.19	0.41	25.42
incorrect	1.31	1.10	69.70	1.19	0.74	47.09

Table 4 Experiment 2

	Mean Δ x [m]	Mean Δ y [m]	Mean Δ θ [deg]	std : x [m]	std : y [m]	std : θ [deg]
All	0.59	0.31	14.75	0.98	0.29	25.34
correct	0.19	0.20	2.79	0.19	0.21	1.82
incorrect	1.65	0.59	45.85	1.34	0.29	31.05

Table 5 Matching results

(a) Experiment 1

	rate[%]
correct	65
incorrect	25
unable to calculate	10

(b) Experiment 2

	rate[%]
correct	65
incorrect	25
unable to calculate	10

4 結論

本研究では，全天球ステレオカメラによる物体認識情報及び 2D 意味地図から得られる，意味情報を含む疎な点群同士を用いた位置姿勢推定手法を提案した．評価実験から，全天球ステレオカメラから得られる距離情報により精度が向上したことを示した．また，時間に依存せずとも十分な精度が得られたため，本手法の有効性は示された．今後は，対象物体の選択や認識された物体をフレーム間追跡するなどの時間的な要素を組み合わせることで位置姿勢推定のロバスト化・高精度化を目標とする．

参考文献

[1] 須田 教明: “電磁波測距儀（改訂版）”, 森北出版, 1976 年

[2] I. Uygur, R. Miyagusuku, S. Pathak, A. Moro, A. Yamashita and H. Asama, “A Framework for Bearing-Only Sparse Semantic Self-Localization for Visually Impaired People,” 2019 IEEE/SICE International Symposium on System Integration (SII), Paris, France, pp.319-324, 2019

[3] S. Pathak et al., “Localization in a Semantic Map via Bounding Box Information and Feature Points,” 2021 IEEE/SICE International Symposium on System Integration (SII), Iwaki, Fukushima, Japan, pp. 126-131, 2021

[4] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp.779-788, 2016

[5] P. F. Alcantarilla et al.: “Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces”, Proc. of BMVC, pp.1-12, 2013

[6] S. Umeyama, “Least-squares estimation of transformation parameters between two point patterns,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, no. 4, pp.376-380, April 1991