Vol.2023-GN-118 No.18
Vol.2023-CDS-36 No.18
Vol.2023-DCC-33 No.18
2023/1/23

撮影物体の名称照合による画像内の物体位置推定手法の提案

酒井 航太[1]　吉野 孝[2]

概要：近年，様々な店舗や施設の紹介において，バーチャルツアーが多く利用されている．バーチャルツアーは，パノラマ画像をブラウザ上で 360 度表示することで，空間の雰囲気を伝えることが可能である．このバーチャルツアーでは，個々の物体に対して情報を付与することが可能なアノテーションが利用されている．このアノテーション付与作業には，複数のパノラマ画像に対して複数の物体画像を割り当てる作業が必要となり，大きな手間となっている．そこで，我々は画像処理による物体情報抽出および，取得した情報に対する自然言語処理によって，アノテーションの付与作業を半自動化し，効率化する手法を提案する．

Proposal of a Method for Estimating Object Locations in Images by Matching Names of Photographed Objects

1. はじめに

近年，様々な店舗や施設の紹介において，バーチャルツアーが多く利用されている．バーチャルツアーは，パノラマ画像をブラウザ上で 360 度表示することで，空間の雰囲気を伝えることが可能である．

このバーチャルツアーでは，個々の物体に対して情報を付与することが可能なアノテーションが利用されている．アノテーションを用いることでパノラマ画像の中の小さくて見づらい部分に対して別途撮影した画像を付け加えて見やすくしたり，特定の物体に対してパノラマ画像の撮影時以外の様子の写真や，文章による追加説明などを付け加えてわかりやすくしたりといったことが可能となる．

しかし，このアノテーションを用いた情報の付与において，施設の紹介に使用されるパノラマ画像は複数枚になることが一般的で，付与するアノテーションも大量に存在することがほとんどである．さらに，従来のアノテーションの付与作業は手動で行われるため，大きな手間となっている．

一方，近年では画像処理技術の発達により，1 枚の画像に映る物体の名称を取得できるようになってきた．この物体認識技術を利用することで，1 枚の画像からその画像内に映る物体の名称および，その位置データを取得することが可能となる．

[1] 和歌山大学大学院 システム工学研究科
[2] 和歌山大学 システム工学部

また，自然言語処理技術も発達しており，1 つの文章について，他の複数の文章との類似度を計算できるようになってきた．このような技術を使うことで，文章だけでなく，物体の名称同士の類似度も計算することができる．

そこで，本研究ではパノラマ画像内に映る物体および，アノテーションとして使用する撮影物体に対して，物体認識技術を使った物体の名称および，位置データの取得を行い，そこから自然言語処理による名称のマッチング処理を実行することで，バーチャルツアー作成におけるアノテーションの付与作業の効率化が可能になると考えた．本稿では，撮影物体の名称照合による画像内の物体位置推定手法の提案および，それらを実装したシステムについて説明する．

2. 関連研究

2.1 画像内の物体位置推定に関する研究

井上らは，全方位カメラを用いた物体検出とトラッキング関する研究を行った [1]．この研究では，視覚障害者に対する障害物の位置通知を目的とし，パノラマ画像中の障害物を検出し，その位置を追跡する手法を提案している．物体の検出については，物体認識アルゴリズムに Yolo v2 を，データセットの拡張を行うために COCO detection dataset および，ImageNet classification dataset を用いている．本研究とは，パノラマ画像内の物体を検出するという点で類似しているが，本研究は画像処理アルゴリズムに加え，自然言語処理によるマッチング処理を行っていると

いう点で異なる．

藤田らは，床指紋を用いた位置推定に関する研究を行った [2]．この研究では，人物の位置推定手法について，既存の GPS，Wi-Fi，ビーコンおよび，RFID などによる位置推定では，精度および，設置コストの問題などがあり，決定的な手法が普及するに至っていない点を指摘している．一方で，屋内では，床面の材質，木目調，石材調などが様々であり，使用によるキズや汚れも出てくる．このような床面の特徴を床指紋と名付け，人が接している 1 枚の床画像に対する画像フィルタリング，特徴点マッチングなどにより，人物の位置を推定する手法を提案している．本研究とは，人物が接する床の画像という，1 枚の画像の位置を推定する点で類似しているが，本研究は屋内のみならず，屋外での物体位置推定も対象としている．

2.2 アノテーション作業の効率化に関する研究

石曽根らは，ユーザ参加型アノテーションにおける UI 及びデータオーグメンテーションのデザインに関する研究を行った [3]．この研究では，視覚障害者向け屋外移動支援システムの開発を目指す過程で，COCO detection dataset などの画像認識の際のデータセットに，横断歩道や信号機といったもののデータが不足している点が指摘されていた．そのような画像認識用のデータセット作成では，既存のアノテーション付与ツールを使って手作業で登録しなければならず，膨大な時間を要することがわかっている．そこで画像認識用の障害物のデータセットを，一般ユーザから提供できるようなアノテーション付与システムの開発を行っている．このアノテーション付与システムでは，スマートフォンなどのカメラで取得した映像に映る物体に対して，その場で指を使ってバウンディングボックスを描画し，画像とバウンディングボックスの位置をリアルタイムで書き出していくことで，効率化を測っている．本研究とは，アノテーション付与作業の効率化という点で類似しているが，この研究はアノテーション付与そのものを手作業で行っているのに対し，本研究ではアノテーション付与作業の半自動化を目的としているという点で異なる．

2.3 自然言語処理に関する研究

自然言語処理による文字列同士の類似度を計算する手法として，BERT がある [4]．BERT は自然言語処理のためのオープンソースの機械学習フレームワークであり，文章同士の類似度の計算などが可能である．BERT では，ラベルなしデータを用いた複数のタスクでの事前学習と，事前学習時の重みを初期値とした状態でのラベルありデータを使ったファインチューニングを行う．この BERT は，単語単位での類似度も計算することが可能であり，本研究では名称データのマッチング処理にこの BERT を使用している．

2.4 画像処理と自然言語処理を融合した研究

Alec らは，画像処理技術と自然言語処理技術を融合した画像の分類モデルを開発した [5]．これは，まずインターネットから収集した 4 億の画像および，テキストのペアのデータセットに対して，どの説明情報がどの画像に対応するかを学習した事前学習モデルを作成する．その上で自然言語を用いた学習した概念を参照することで，未知の画像を分類することを可能にしたものである．本研究とは，画像処理技術と自然言語処理を融合しているという点で類似しているが，本研究では事前学習などによるモデルの作成は行っていない．

3. 本手法について

3.1 物体位置推定手法の概要

本手法は，画像処理技術と自然言語処理技術を融合した，撮影物体の名称照合による画像内の物体位置推定手法である．

図 1 に本手法における物体位置推定の大まかな流れを示す．

(1) パノラマ画像内物体の名称と位置の取得 (図 1(1))
パノラマ画像内に存在する物体領域について，それぞれの物体領域に含まれる物体の名称および，位置を取得し，データベースに保存する．

(2) 撮影物体の名称取得 (図 1(2))
パノラマ画像と同様に，スマートフォンなどで撮影した物体画像についても名称を取得する．

(3) 物体名称のマッチング (図 1(3))
撮影した物体画像から取得した名称について，パノラマ画像内の物体から取得した名称データとの類似度を自然言語処理で計算し，最大類似度を持つ名称データを取得するマッチング処理を行う．

(4) マッチングデータを使った位置の可視化 (図 1(4))
マッチング処理によって取得した最大類似度を持つ名称データについて，データベースからパノラマ画像内の物体名称および，位置のデータを参照し，位置データを取得する．取得した位置データを，撮影した物体のパノラマ中の位置として扱い，可視化する．

3.2 実装システムの構成

図 2 に本手法を実装したシステムの構成を示す．本システムは，図 2 の左下（ア）に示すブラウザ，図 2 の中央下（イ）に示す Web サーバ，図 2 の中央上（ウ）に示す画像処理システム，図 2 の右下（エ）に示すデータベースおよび，図 2 の右上（オ）に示す自然言語処理システムから構成される．

(1) Web サーバへのパノラマ画像送信 (図 2(1))
パノラマ画像の撮影は，パノラマ画像を撮影することが

Vol.2023-GN-118 No.18
Vol.2023-CDS-36 No.18
Vol.2023-DCC-33 No.18
2023/1/23

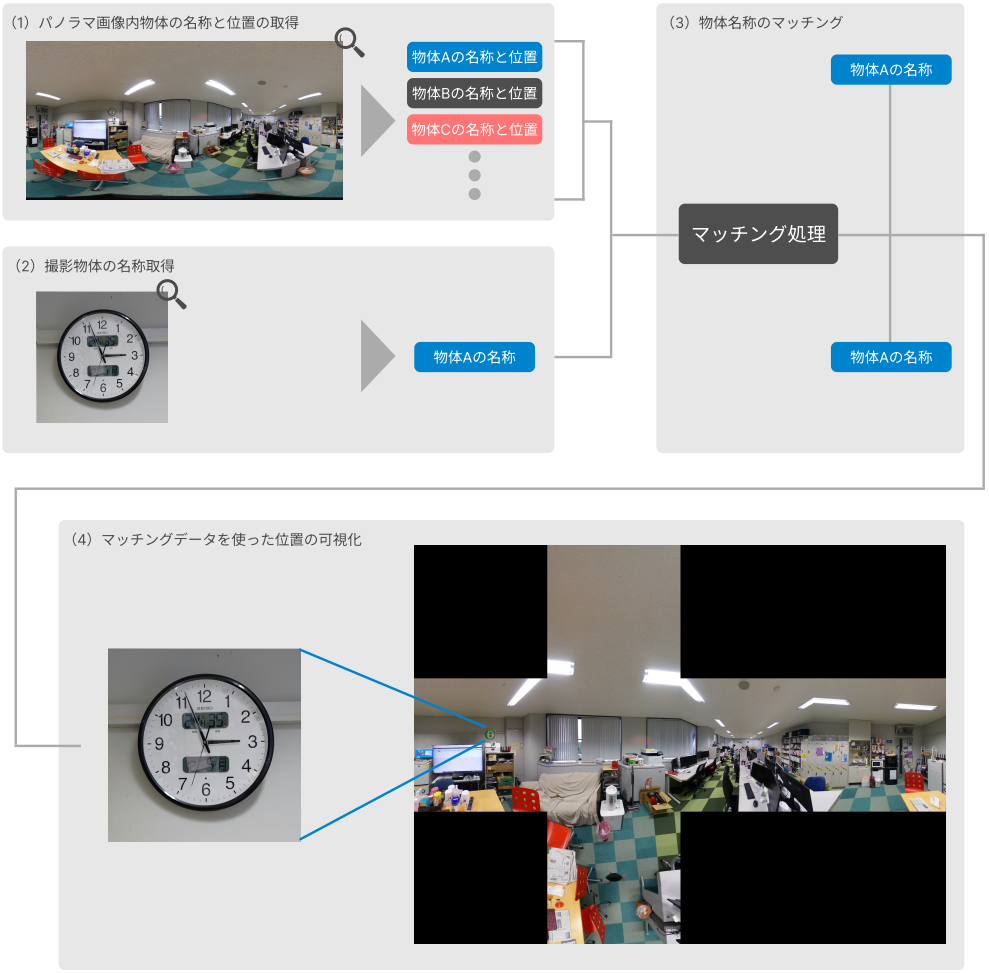


図 1 物体位置推定の流れ

できる RICOH THETA Z1 を用いた[*1]．この RICOH THETA Z1 は，専用のスマートフォンアプリからパノラマ画像を撮影することが可能で，撮影したパノラマ画像はスマートフォン内に保存される．そのため，パノラマ画像のアップロードはスマートフォンを使い，ブラウザから行う．パノラマ画像内物体の名称を取得する際，まず図 2 の（ア）ブラウザから，図 2 の（イ）Web サーバへ，パノラマ画像を送信する．

(2) 画像処理システムへのパノラマ画像送信 (図 2(2))
パノラマ画像を受信した Web サーバは，図 2 の（ウ）画像処理システムへパノラマ画像を送信する．

(3) パノラマ画像内物体の名称保存 (図 2(3))
図 2 の（ウ）画像処理システムは，図 2 の（イ）Web サーバからパノラマ画像を受信すると，画像内の物体について名称および，位置を取得し，図 2 の（エ）データベースに保存する．

(4) Web サーバへの撮影物体の画像送信 (図 2(4))
撮影物体の名称を取得する際，まず図 2 の（ア）ブラウザ上で物体画像を撮影し，図 2 の（イ）Web サーバへ送信する．

(5) 画像処理システムへの撮影物体の画像送信 (図 2(5))
図 2 の（イ）Web サーバは，撮影物体の画像を受信すると，図 2 の（ウ）画像処理システムへ送信する．

(6) 撮影物体の名称送信 (図 2(6))
図 2 の（ウ）画像処理システムは，図 2 の（イ）Web サーバから撮影物体の画像を受信すると，撮影物体の名称を取得し図 2 の（オ）自然言語処理システムへ送信する．

(7) パノラマ画像内物体の名称取得 (図 2(7))
図 2 の（オ）自然言語処理システムは，図 2 の（ウ）画像処理システムから撮影物体の名称を受信すると，パノラマ内物体の名称を取得するため，図 2 の（エ）データベースからパノラマ画像内物体の名称データを取得する．

(8) アノテーションデータ保存 (図 2(8))
図 2 の（オ）自然言語処理システムは，図 2 の（エ）データベースからパノラマ画像内物体の名称データを取得する．データの取得が完了すると，撮影物体の名称とパノラマ画像内物体の名称との類似度を計算し，最も類似度が高いデータを撮影物体と同一物体のデー

[*1] RICOH THETA Z1：入手先 ⟨https://theta360.com/ja/about/theta/z1.html⟩

タとするマッチング処理を実行する．このマッチングによって推定されたデータを，アノテーションデータとして図 2 の（エ）データベースに保存する．

(9) アノテーションデータの取得 (図 2(9))
図 2 の（イ）Web サーバは，図 2 の（ア）ブラウザからバーチャルツアー閲覧のリクエストが来ると，図 2 の（エ）データベースからアノテーションデータを取得する．

(10) アノテーションデータの送信 (図 2(10))
図 2 の（イ）Web サーバは，図 2 の（エ）データベースからアノテーションデータを取得すると，図 2 の（ア）ブラウザへアノテーションデータを送信し，ブラウザ上でアノテーションが表示される．

3.3 パノラマ画像内物体の名称取得について

図 3 にパノラマ画像内に存在する物体の名称を取得する流れを示す．

(1) パノラマ画像をキューブマップに変換 (図 3(1))
パノラマ画像は一般的に図 3 の（ア）パノラマ画像のように，歪みがあるなどの理由で扱いづらい．これを解消するため，図 3 の（イ）のように，キューブマップと呼ばれる形式に変換する．また，図 3 の（ウ）キューブマップの面のように，キューブマップへの変換で得られた画像のうち，パノラマ画像の領域に該当する正方形の領域を面と呼ぶ．

(2) キューブマップの各面について物体領域を検出 (図 3(2))
キューブマップの各面について，物体領域提案アルゴリズムである Selective Search を利用し，物体である可能性が高い領域を取得する．図 3 の左下に，キューブマップの正面画像について，物体領域検出を行った結果を示す．なお，Selective Search のみでは取得した物体領域の多くが重なっていたり，断片のみが検出されていたりすることがあるため，Non-Maximum Suppression によって信頼度が高い領域のみを残している[*2]．

(3) 検出した物体領域内に存在する物体の名称を取得 (図 3(3))
検出した物体領域に対して，物体認識技術を利用し，物体名称を取得する．物体認識技術として，今回は Google Cloud Vision API を利用した[*3]．Google Cloud Vision API を利用することで，画像内に存在する物体名称および，画像内での位置データを取得することができる．図 3 の右下に，キューブマップの正面画像について物体名称を取得した結果を示す．また，このように取得した物体名称および，その位置データは，データベースに保存する．

3.4 撮影物体の名称取得

撮影物体についても，パノラマ画像と同様に Google Cloud Vision API を利用し，物体名称を取得する．**図 4** に撮影物体の物体名称を取得した結果を示す．今回は所属している研究室に設置されている時計および，電子レンジの画像を用いた．

3.5 物体名称のマッチング処理

パノラマ画像内の物体からの名称取得および，撮影物体からの名称取得を行ったあとは，物体名称のマッチング処理を行う．物体名称のマッチング方法について，２通りの手法を検討した．１つ目は，文字列が一致している名称を探す方法．２つ目は，名称同士の類似度を計算し，最大類似度を持つ名称を探す方法である．

前者について，Google Cloud Vision API で物体名称を取得した際，同一物体であっても撮影時の状況によっては，異なる名称になる場合が存在した．**図 5** にパノラマ画像から取得した物体名称と，撮影物体から取得した名称がずれている様子を示す．図 5 の上（ア）パノラマ画像における時計の認識結果では，時計が「Clock」として認識されている．一方，図 5 の下（イ）撮影物体画像における時計の認識結果では，時計が「Wall clock」として認識されている．このような場合，単純な文字列の一致による探索が難しいと考え，２つ目の名称同士の類似度の計算による探索を採用した．

名称同士の類似度の計算には，BERT を使用した．BERT を使用した理由として，複合語を扱うことができるという点がある．パノラマ画像から取得した物体名称の中には，Packaged goods のような複合語が多数存在しており，この名称のうち goods は名詞だが，Packaged は名詞ではない．このような複数の品詞から構成される複合語が含まれているため，複合語の状態で類似度を計算できる必要がある．

表 1 に，パノラマ画像内の物体名称のうち，図 4 の（ア）時計および，（イ）電子レンジとの類似度が高い名称上位 10 件を示す．表 1 の右に記載している類似度について，最大値が 1.0，最小値が 0.0 の小数となっている．この結果から，時計も電子レンジも類似物体の名称が取得できていることがわかる．

以上のように取得した類似度データから，最大類似度を持つ名称のデータを，撮影物体と同一物体のデータとして出力する．

[*2] Non-Maximum Suppresion：入手先 ⟨https://cvml-expertguide.net/terms/dl/object-detection/non-maximum-suppression/⟩
[*3] Cloud Vision API：入手先 ⟨https://cloud.google.com/vision?hl=ja⟩

Vol.2023-GN-118 No.18
Vol.2023-CDS-36 No.18
Vol.2023-DCC-33 No.18
2023/1/23

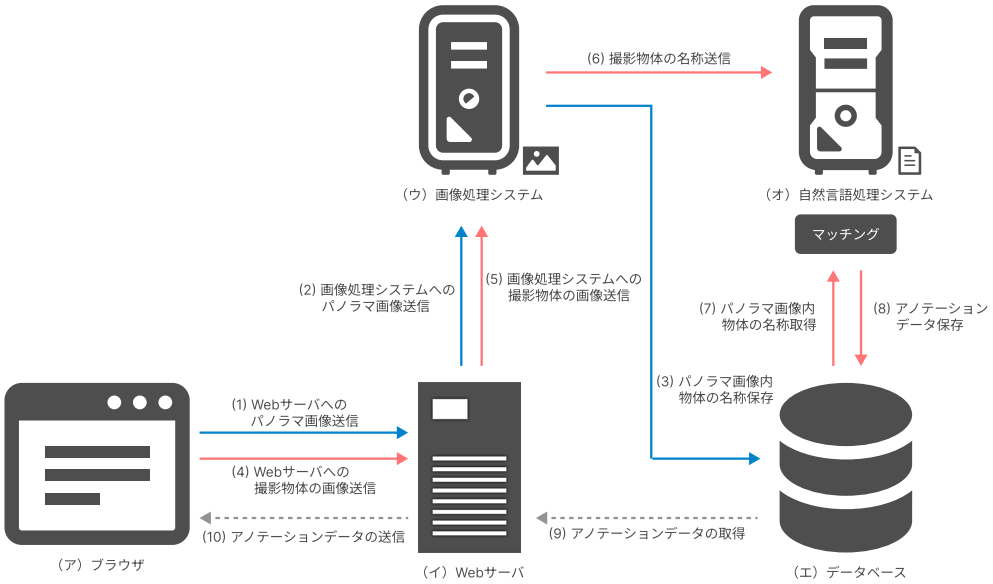


図 2　システム構成

図 3　パノラマ画像内物体の名称取得の流れ

Vol.2023-GN-118 No.18
Vol.2023-CDS-36 No.18
Vol.2023-DCC-33 No.18
2023/1/23

（ア）時計

（イ）電子レンジ

図 4　撮影物体の名称取得結果

4. 手法の動作確認

まず，成功例を示す．**図 6** に時計および，電子レンジについて，最大類似度を持つ名称データの取得および，その位置を可視化した結果を示す．図 6 の左上（ア）時計の位置は，図 6 の右上（イ）時計の位置をキューブマップ上に可視化した様子に，図 6 の左下（ウ）電子レンジの位置は，図 6 の右下（エ）電子レンジの位置をキューブマップ上に可視化した様子にそれぞれ可視化されており，位置推定が行われていることがわかる．

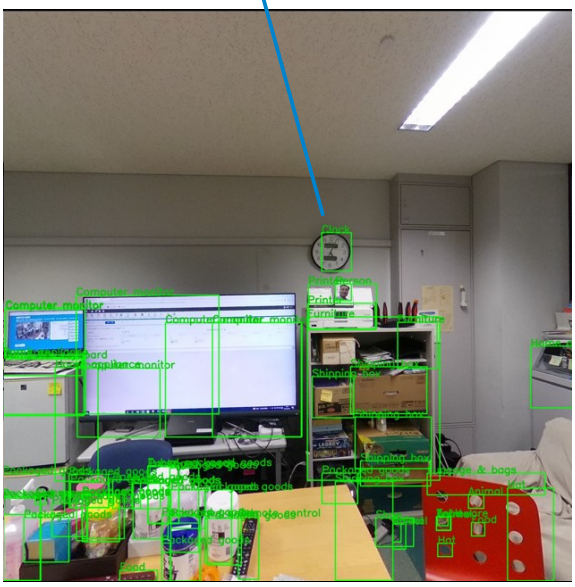
次に失敗例を示す．**図 7** に撮影物体の画像から名称が取得できない例を示す．図 7 の（ア）のように，物体画像の中で物体の上下左右にある程度の余裕がある場合，物体の名称を取得することができる．しかし，図 7 の（イ）や（ウ）のように，物体の周りに余裕がない場合や，他の物体が大きく映り込んでいる場合，物体の名称を取得できない場合がある．以上より，アノテーションを付与したい物体の画像を撮影する場合，撮影機能を工夫する必要がある．

5. おわりに

本稿では，撮影物体の名称照合による画像内の物体位置推定手法の提案および，それらを実装したシステムについて説明した．サンプルデータを使い本システムを実行した結果，撮影した物体のパノラマ画像内位置を推定することができた．

今後は，様々な条件下での物体位置推定を行い，物体位置推定の精度について検証する．

（ア）パノラマ画像における
時計の認識結果「Clock」



（イ）撮影物体画像における
時計の認識結果「Wall Clock」

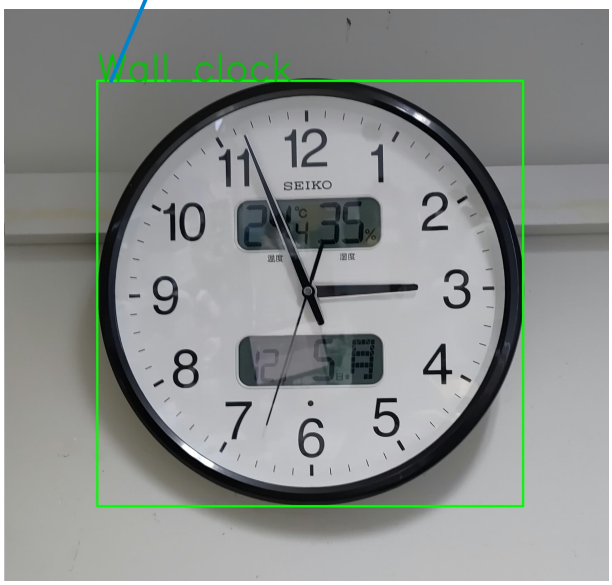


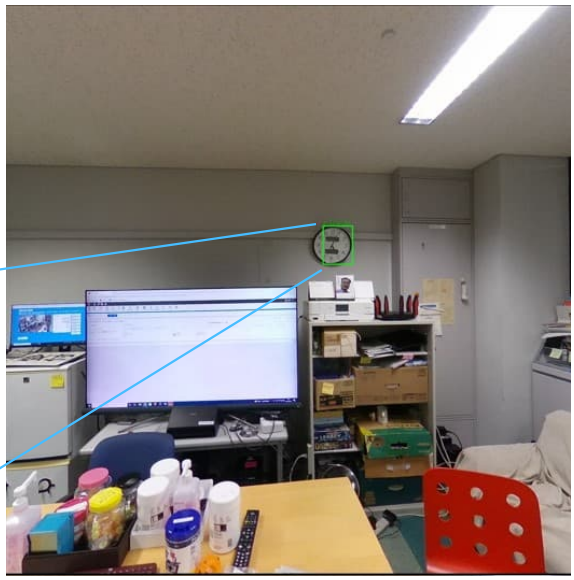
図 5　物体名称の取得結果のずれ

参考文献

[1] 井上慶彦，岩村雅一，黄瀬浩一：全方位カメラを用いた物体検出とトラッキング-視覚障害者支援システムの実現に向けて-，情報処理学会研究報告，No.20，Vol.2018，pp.1–6（2018）．

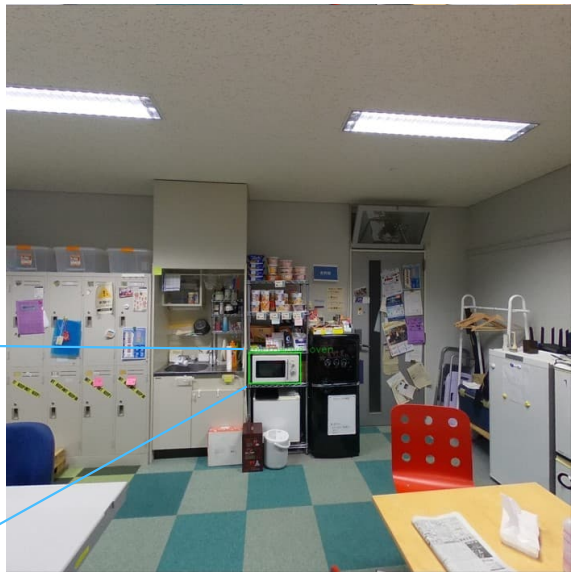
Vol.2023-GN-118 No.18
Vol.2023-CDS-36 No.18
Vol.2023-DCC-33 No.18
2023/1/23

（ア）時計



（イ）時計の位置をキューブマップ上に可視化した様子

（ウ）電子レンジ



（エ）電子レンジの位置をキューブマップ上に可視化した様子

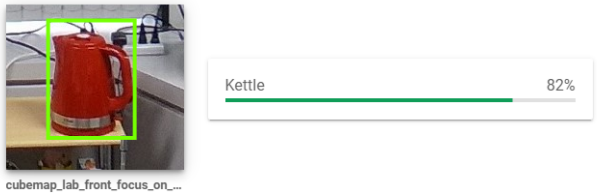
図 6 位置推定の結果

Vol.2023-GN-118 No.18
Vol.2023-CDS-36 No.18
Vol.2023-DCC-33 No.18
2023/1/23

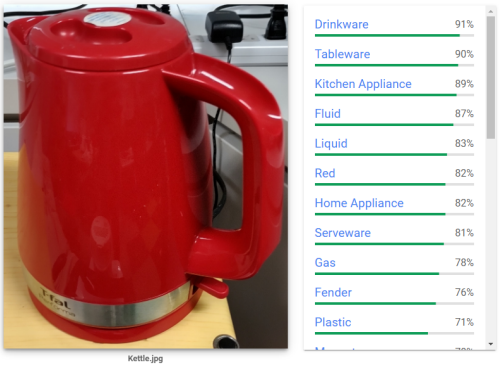
表 1　類似度上位 10 件の名称

撮影した物体	取得した物体名称	類似度上位 10 件の名称	類似度
時計	Wall clock	clock	0.8684
		racket	0.4931
		window	0.3668
		watch	0.3644
		furniture	0.3598
		light fixture	0.3532
		printer	0.3441
		computer monitor	0.3386
		picture frame	0.3379
		tableware	0.3212
電子レンジ	Microwave oven	microwave oven	1.0
		kitchen appliance	0.7377
		kettle	0.6495
		home appliance	0.5539
		refrigerator	0.5360
		light fixture	0.4390
		lighting	0.3956
		food	0.3427
		tableware	0.3385
		toilet	0.3332

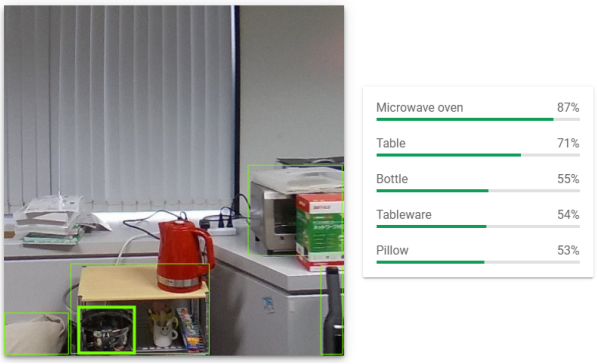
[2] 藤田悟，内田薫：床指紋を用いた位置推定，マルチメディア，分散協調とモバイルシンポジウム 2016 論文集，Vol.2016，pp.1244–1250（2016）．
[3] 石曽根奏子，馬場哲晃，渡邉英徳：ユーザ参加型アノテーションにおける UI 及びデータオーグメンテーションのデザイン，研究報告アクセシビリティ（AAC），No.1，Vol.2018，pp.1–4（2018）．
[4] Jacob Devlin，Ming-Wei Chang，Kenton Lee：BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding，Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies，Vol.1，pp.4171–4186（2019）．
[5] Alec Radford，Jong Wook Kim，Chris Hallacy：Learning Transferable Visual Models From Natural Language Supervision，Proceedings of the 38th International Conference on Machine Learning，Vol.139，pp.8748–8763（2021）．



（ア）撮影物体の認識が成功する場合



（イ）物体が認識されない場合



（ウ）情報が欲しい物体以外が検出される場合

図 7　位置推定の失敗例