サービスロボットのための第４人称センシングの提案

Proposal of fourth-person sensing for service robots

○ 中嶋 一斗（九州大） 岩下 友美（九州大） ピョ ユンソク（九州大）
高嶺 朝理（九州大） 正 倉爪 亮（九州大）

Kazuto NAKASHIMA, Kyushu University, k_nakashima@irvs.ait.kyushu-u.ac.jp
Yumi IWASHITA, Kyushu University, Yoonseok PYO, Kyushu University
Asamichi TAKAMINE, Kyushu University, Ryo KURAZUME, Kyushu University

This paper proposes a new concept of "fourth-person sensing" for service robots. The proposed concept combines wearable cameras (the first-person viewpoint), sensors mounted on robots (the second-person viewpoint) and sensors embedded in the informationally structured environment (the third-person viewpoint). Each sensor has its advantage and disadvantage, while the proposed concept can compensate the disadvantages by combining the advantages of all sensors. The proposed concept can be used to understand a user's intention and context of the scene with high accuracy, thus it enables to provide proactive services by service robots. As one of applications of the proposed concept, we developed a HCI system combines the first-person sensing and the third-person one. We show the effectiveness of the proposed concepts through experiments.

Key Words: Service robots, fourth-person vision, TMS, activity recognition, spatio-temporal features

1 はじめに

高齢化の影響に伴い，介護現場における労働力不足が深刻化しており，人との共生を目指したサービスロボットの開発が進められている．一方で，サービスロボットが実際に生活支援サービスを計画・提供するためには，複雑に変動する生活空間の中で多くの環境情報を取得し，それらを実時間で処理する必要がある．そのため，センサの可搬能力や処理能力に限界のあるサービスロボット単体が全てを実行することは困難である．

この問題に対し，我々はサービスロボットの作業環境側に分散センサネットワークを構築する環境情報構造化アーキテクチャTown Management System (TMS) の開発を進めてきた [1]．TMS では，環境全体に分散配置したセンサにより空間内の人やロボット，物品の位置や状態といった情報を取得し，クラウド型データベースで統合管理する．サービスロボットは，作業を行う際にこれらの環境情報を利用することで，仮想的に拡大したセンシング能力を得ることができる．また，現在ではシステムのミドルウェアにRobot Operating System (ROS) を導入し，ロボットやセンサ，機能の追加に柔軟なアーキテクチャROS-TMS として開発を行っている [2].

従来の ROS-TMS で管理される環境情報を生活支援を受けるユーザの視点（1 人称）から整理すると，サービスロボットに搭載するセンサから得られる情報を 2 人称，環境全体に固定したセンサから得られる情報を 3 人称とすることができる．これら 2 人称・3 人称視点の情報は，環境全体を計測することができる反面，ユーザに近い環境に対しては，解像度や死角の存在などの問題が起きやすく，ユーザの指示や要求を信頼性高く認識することが困難な場合がある．

そこで本研究では，従来の 2 人称・3 人称視点による環境計測に加えて，ウェアラブルカメラによって得られる 1 人称視点情報を利用し，3 者を組み合わせた新たなセンシングシステム「第 4 人称センシング」を提案する．また，第 4 人称センシングの適用例として，曖昧性を含むサービスロボットへの物品取り寄せ指示に焦点を当て，1 人称視点映像により認識したユーザ行動と TMS の 3 人称センサで計測された物品情報を基に，物品特定を行うシステムを構築する．さらに，構築したシステムを用いた実験を行い，第 4 人称センシングが曖昧な指示に対する正確な理解に有効であることを示す．

2 第 4 人称センシング

2.1 概念

ここで述べる第 4 人称という言葉は，1 人称・2 人称・3 人称の 3 者の状態を客観的な立場から理解し，独自の解釈や分析を行う視点を指す．小説を例に挙げると，主人公を始めとした登場人物らが展開する世界を，物語として読み取る「読者」の視点に相当する．読者は，物語を読み進めていく中で，その世界とは完全に独立した視点から，通常では知り得ない主人公（1 人称），相手（2 人称），それを取り巻く人々（3 人称）の心の動きを把握し，独自の予測を立てることができる．第 4 人称による環境計測が目指す究極の目標は，3 つの人称視点を以って環境を分析することで，ユーザの心理状態からコンテキスト，環境の状態に至るまで包括的な空間の理解を行うことである．

一方で，各人称で得られる情報には，それぞれ長所と短所がある．1 人称センサは，ウェアラブルカメラ装着者の行動を認識したり，細かな変化からユーザの意図や興味を推定することができるが，計測範囲が狭く，局所的・断片的な情報になりがちである．2 人称センサは，サービスロボット自体が生活空間内を移動できることから，環境に固定されるセンサに比べて計測の自由度が高く，実際にサービスを受ける人とその周囲環境を計測するのに適している．一方で，可搬能力や処理能力に制約を受けるため，多くのセンサを搭載することはできず，生活支援に十分な情報を得ることができない．3 人称センサは，対象・ロボット・環境を俯瞰的に計測することができるが，計測対象から離れた位置に固定されていたり，何らかの計測のみに特化した配置になっていることが多いため，死角や解像度といった問題が起きやすく，空間内の人の要求や指示を高精度に理解をすることは困難である．

一方，これら 3 者を相補的に組み合わせることで，サービスロボットへの指示に関連して次のことが期待できる．1 つ目に，より正確な指示理解である．システムに対するサービス要請の手段としては，音声が広く利用される．音声による指示はユーザから自発的に明示されるため，サービスのトリガとしては有用である．しかし，自然な音声指示の中で，ユーザの意図や要求が充分に表現される場合は少ない．一方，ウェアラブルカメラによって得られる 1 人称視点には，装着者が何を見ているか，何をしているかといった情報が含まれている．1 人称視点映像の見えや動きの特徴を分析すれば，これらを行動情報や注視情報として抽出す

ることができ，音声指示が曖昧な場合でも，指示内容を明確にできる可能性がある．2 つ目に予見的なサービスの開始である．1 人称視点からは，2 人称・3 人称センサでは捉えることのできない細かな変化を計測することができる．これらには装着者の意図や興味といった心理的要因に依る動作も含まれており，直近で明示的な指示が行われる可能性が高い．それら特徴的な動作を検出した時点でサービスを開始すれば，従来の 2 人称・3 人称によるシステムよりも早く生活支援を提供することができる．以下の節では，1 人称，2 人称，3 人称の各種センサについて説明する．

2.2 1 人称センシング

近年，高性能なウェアラブルカメラが手軽に入手できるようになった．なかでも，一般的に smart glasses と呼ばれるものの多くは，それ自体が Android OS を搭載しており，ウェアラブルカメラとしての側面だけでなく，可搬性の高いコンピュータとして幅広い用途に利用できる．また，マイクやスピーカー，ディスプレイといったユーザインターフェースを内蔵しているため，Human-Computer Interaction (HCI) を担うデバイスとして，TMS アーキテクチャに導入することもできる．本研究では，Epson 社の Moverio BT-200AV (図 1) を利用した．

Fig.1 The first-person viewpoint: wearable camera

2.3 2 人称センシング

サービスロボットは，生活支援サービスを提供する側の立場にあり，自身に搭載するセンサから，生活支援対象の周囲環境を計測したり，作業に必要な環境情報を取得する．そのため，ロボットの視点から計測される情報を 2 人称情報と表現することができる．現在 TMS で稼働するサービスロボット SmartPal V (図 2) は，頭部に LRF と RGB-D センサ，胴体には同様の LRF とカメラを搭載している．

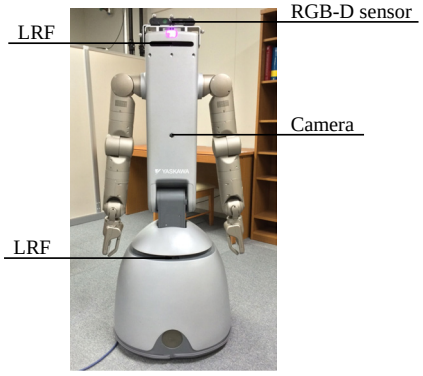


Fig.2 The second-person viewpoint: robot-mounted sensors

2.4 3 人称センシング

我々は，環境全体に分散センサネットワークを構築する TMS の開発を行っており，生活空間で計測された環境情報はクラウド型データベースで管理される．分散センサとしては，LRF や RFID タグリーダ，Load cell，RGB-D センサなどが挙げられ，空間全体の物体位置や状態を計測している．本研究では，生活支援を受けるユーザの 1 人称情報，生活支援を提供するロボットの 2 人称情報に対して，環境全体の分散センサから取得する情報を 3 人称情報と呼ぶ．

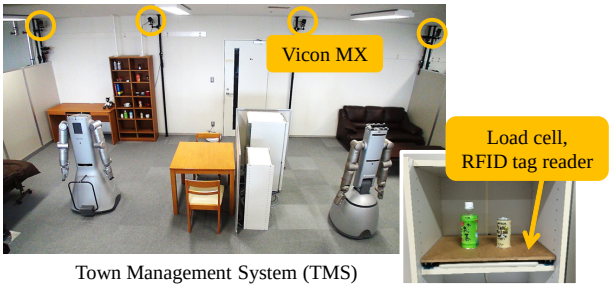


Fig.3 The third-person viewpoint: embedded sensors in the environment

3 第 4 人称センシングの適用例

本章では，知能化空間内で想定される曖昧な物品取り寄せ指示に着目し，第 4 人称センシングの適用例を示す．図 4 では，食事中のユーザがペットボトルの取り寄せを希望し，水の取り寄せを指示している．一方で，図 4 の生活空間では，水に関する物品はペットボトルだけではなく，園芸のためのじょうろや，掃除のためのバケツが存在する．この場合，TMS やサービスロボットが，ユーザの音声内容のみから適切な対象物品を判断することは困難である．

一方，水といった抽象的な指示を理解する手掛りとしてユーザの行動情報がある．この場合，ユーザは食事に必要な飲料水としてペットボトルの取り寄せを意図した．システムがユーザの行動を認識することができれば，複数の候補から食事に関連する物品を優先的に選択することができる．こうした行動を捉える方法として，1 人称視点映像の利用が適している．1 人称視点は，計測範囲が狭い分，装着者の能動的動作や近辺で起こっている事象を捉えやすい．そのため，食事や読書といった行動は，1 人称視点画像中の見えや動きの特徴として現れる可能性が高い．

本研究では，以上のシナリオを想定し，1 人称センシングによる行動情報と 3 人称センシングによる物品情報を相補的に組み合わせたシステムの構築を行った．また，実際の生活空間で想定される多様な指示や行動に対応するのは困難であるため，TMS の実験環境で想定される水の取り寄せに焦点を当てた．

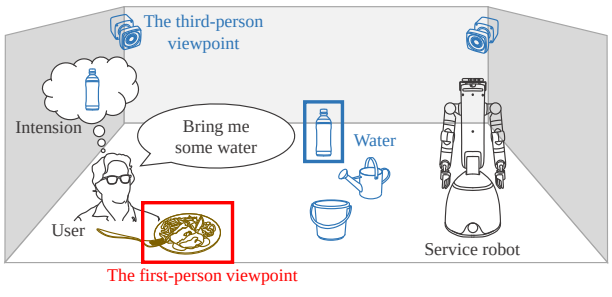


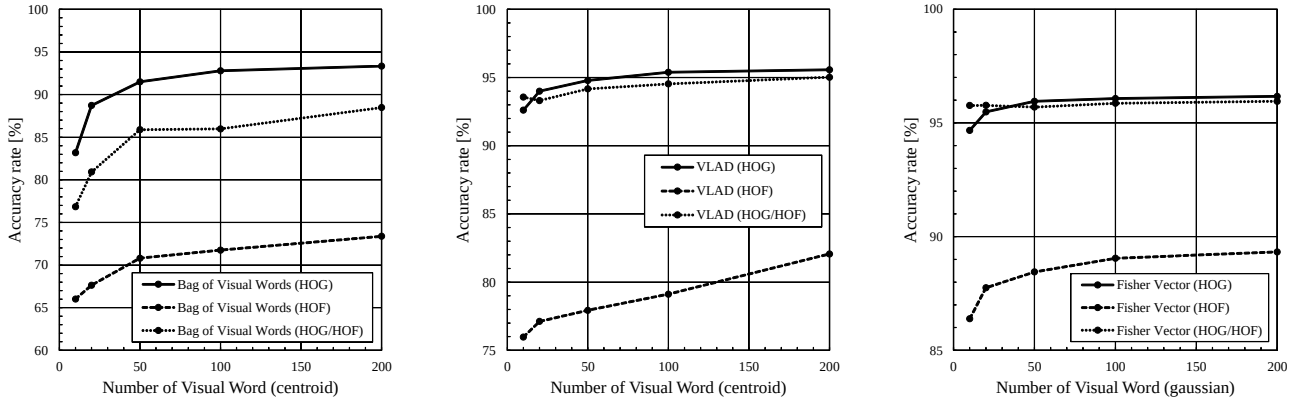
Fig.4 Service scenario

3.1 システム構成

ウェアラブルカメラと処理サーバによる分散システムを構築した．

3.1.1 ウェアラブルカメラ

本研究で使用する Moverio BT-200AV は Android OS を搭載しており，1 人称視点画像と音声指示の取得，システムユーザへの情報提示を行う．1 人称視点画像は，前方に搭載したカメラか



(a) Bag of Visual Words (b) VLAD (c) Fisher Vector

Fig.5 Accuracy rate for various numbers of Visual Word k: (a) Maximum rate is 93.3% with HOG descriptor, k=200 (b) Maximum rate is 95.6% with HOG descriptor, k=200 (c) Maximum rate is 96.2% with HOG descriptor, k=200

ら取得した後，画像圧縮を施し，動画像のフレームレートに相当する一定周期で処理サーバに送信する．音声指示は，搭載するマイクから不定期に受け付け，認識できれば処理サーバへ送信する．また，サーバの処理状況やサービス実施状況を搭載ディスプレイから適宜提示する．

3.1.2 処理サーバ

処理サーバは，次の手順に従い，定期的な行動認識処理を行う．

1. 受信する 1 人称視点画像を固定サイズバッファに適宜格納し，動画像を生成
2. 動画像から局所特徴を抽出
3. 動画像の特徴ベクトルを計算
4. Support Vector Machine (SVM) によるカテゴリ識別

また，ウェアラブルカメラから音声指示を受信した場合，次の手順に従って対象物品の特定を行う．

1. 音声指示内容から，特定の物品名検索
2. 関連する物品候補リストをデータベースから取得
3. 物品候補リストをその時点の行動情報に基づいて，ソート
4. 優先順位の高い物品から，サービスロボットへ発話指令を送信

第 4 章，第 5 章では，それぞれの処理の詳細について説明する．

4 1 人称視点映像による行動認識

本章では，1 人称視点映像による行動認識を実現するための手法をいくつか検討し，それらの識別性能から最適な認識プロセスを決定する．まず，動画像から局所特徴を抽出する手法，それら多くの局所特徴を 1 つのベクトルにエンコーディングする手法について述べ，識別評価の方法と結果について述べる．また，識別対象とする行動カテゴリは，読書，食事，植木を注視，ロボットを注視，辺りを見回すの 5 つとした．

4.1 特徴抽出

本研究では，動画像の時空間変化に基いて局所特徴点を検出する手法として，Laptev が提案した Space-Time Interest Points (STIP)[3] を利用した．検出した特徴点に関しては，Histogram of Oriented Gradients (HOG) と Histogram of Optical Flow (HOF)，及び両者を結合したヒストグラムに従って特徴記述を行い，局所特徴ベクトルとした．また，抽出した局所特徴ベクトルに対し，次元数の削減を行った．最終的な動画像の特徴ベクトルは，次節の処理により局所特徴の次元数に比例した高次元ベクトルとなる．そのため，ここでは計算コストの削減と情報の圧縮を目的として主成分分析を行い，累積寄与率 95%の主成分を利用した．

4.2 局所特徴のエンコーディング

1 つの動画像から抽出された多数の局所特徴の統計的分布に基づいて，動画像の特徴を表現する 1 つのベクトルにエンコーディングする．本研究では，局所特徴のエンコーディング手法としては一般的な Bag of Visual Words[4] に加えて，より高次の統計量を利用する Fisher Vector[5]，Vector of Locally Aggregated Descriptors (VLAD)[6] の 3 手法を適用した．

4.3 カテゴリ識別

本研究では，5 つの行動カテゴリを識別するための学習モデルとして Linear Support Vector Machine（Linear SVM）を用い，前節までに求めた動画像の特徴ベクトルから帰属カテゴリを出力する．

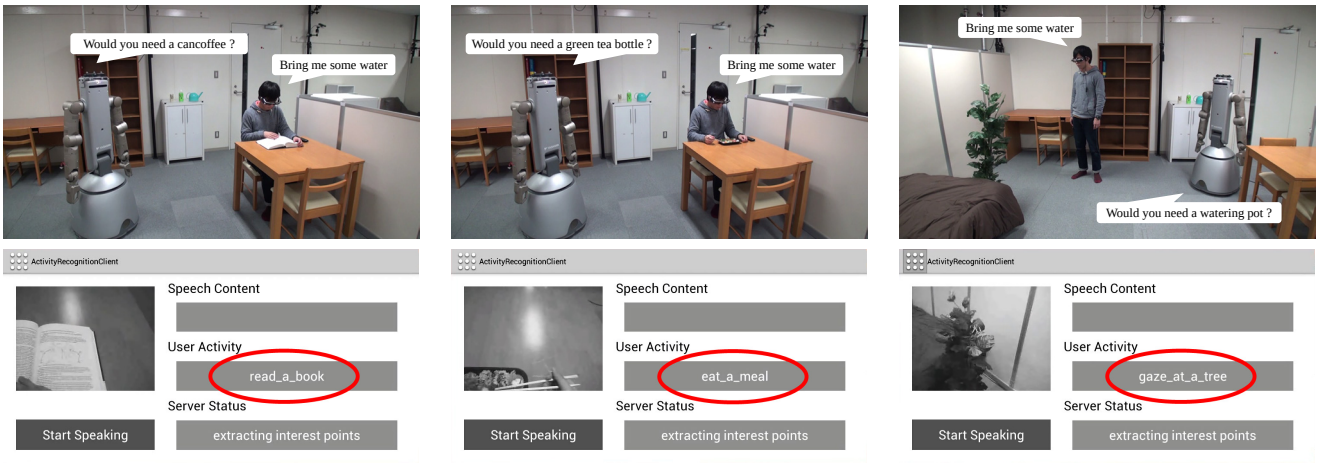
4.4 識別評価と結果

1 シーケンスを 10 秒，画像サイズを 320×240，フレームレートを 30 fps とし，各カテゴリ 50 シーケンスの動画像をウェアラブルカメラにより撮影した．識別評価を行う際には，まず，各行動カテゴリ 50 シーケンスから半数の動画をランダムサンプリングする．これらを学習データセットとし，Visual Words，PCA の主成分，SVM のパラメータを学習する．また，残り半数の動画をテストデータセットとし，特徴量計算，識別を行う．以上の手順を 100 回試行し，平均正解率を算出する．また，各エンコーダの Visual Words の数を 10，20，50，100，200 と変化させた．

識別結果を図 5 に示す．いずれのパターンにおいても，特徴記述子に HOG を選択した場合が最も高い識別率を示す傾向にある．また，Fisher Vector が全体的に最も高い識別性能を示した．そこで以降の実験では，特徴記述子として HOG，エンコーダとして Fisher Vector を選択し，また Fisher Vector の Visual Words の数を 100 とする．

5 指示対象の推定法

音声指示から複数候補が列挙された場合の行動情報による対象推定法に述べる．本システムでは，TMS データベースの物品情報の項目の 1 つ「タグ」を利用して，候補物品と認識された行動情報との関連度を比較する．タグ情報は，物品に関連するキーワードを羅列した項目である．例えば，お茶の入ったペットボトルには，「drink」，「tea」，「water」といった情報がタグとして登録される．各行動情報にもデータベースと同様のタグを複数割り当てておく．物品候補が与えられると，各物品に登録されている



(a) Reading a book (b) Eating a meal (c) Gazing at a tree

Fig.6 Experiment: Figures on upper row shows actual images and a user did some activities. Figures on lower row shows the screen of wearable camera. Recognized results are shown as a "User Activity" (red circles)

Table 1 Tags associating to activities

Activity	Tag
read a book	drink, coffee
eat a meal	drink, tea
gaze at a tree	pot

Table 2 Objects stored in the database

Category	Name	Tag
Coffee	cancofee	drink, coffee, water
Tea	greentea_bottle	drink, tea, water
Tea	soukentea_bottle	drink, tea, water
Watering Pot	watering_pot	pot, water

タグとその時点の行動情報に結びつけられたタグとでマッチする個数をカウントし，その個数の大きい順から指示対象としての優先度を与えていく．

5.1 サービス実験

本章では，設計したシステムを用いたサービス実験について述べる．ウェアラブルカメラを装着したユーザが,「読書」,「食事」,「植木を注視」の 3 つの異なる行動の最中に，水の取り寄せを指示し，それに応答してサービスロボットが提示する物品名を確認する．各行動において水の取り寄せを指示するユーザは,「読書」の場合にコーヒー,「食事」の場合にお茶,「植木を注視」の場合にじょうろを意図しているものとし，これらに相当する物品が TMS データベースから提示されれば，システムの応答は適切である．表 1 に各行動に関連付けたタグ情報を示し，表 2 に TMS データベースで管理されている水に関連する物品を示す．水の取り寄せを行った場合，これらが候補となる．

5.2 実験結果

各行動で，水の取り寄せを指示した際の実験結果について述べる．図 6(a) は，読書している場合の様子である．1 人称視点から得た動画像のカテゴリは,「読書」であると識別された．また，水の取り寄せを指示したところ，サービスロボットから「cancoffee」の提示を受けた．図 6(b) は，食事している場合の様子である．1 人称視点から得た動画像のカテゴリは,「食事」であると識別された．また，水の取り寄せを指示したところ，サービスロボットから「green tea bottle」の提示を受けた．図 6(c) は，植木を注視している場合の様子である．1 人称視点から得た動画像のカテゴリは,「植木を注視」であると識別された．また，水の取り寄せを指示したところ,「watering pot」の提示を受けた．実験結果から，共通の曖昧な音声指示に対しても，その時点の行動情報によって適切な物品を特定できていることを確認した．

6 まとめ

従来の知能化空間で利用されてきた分散センサを人称の観点から分類し，新たに 1 人称センサとしてウェアラブルカメラを統合する第 4 人称センシングを提案した．また，第 4 人称センシングの適用例として，曖昧性を含む物品取り寄せ指示に焦点を当て，1 人称センシングと 3 人称センシングを相補的に組み合わせたシステムを構築した．サービス実験により，構築したシステムの有用性を確認し，第 4 人称センシングの概念がより正確な指示の理解に有効であることを示した．

本研究では 1 人称による計測と 3 人称による計測を相補的に組み合わたシステムを構築したが，未だ 2 人称情報の導入には至っていない．今後は，1 人称・2 人称・3 人称の 3 者を統合し，より正確な空間の理解とそれを応用した新たなシステムの構築を課題とする．

謝辞

本研究は文部科学省科学研究費補助金挑戦的萌芽（課題番号 26630099）の支援を受けた．

References

[1] 村上剛司, 長谷川勉, 木室義彦, 千田陽介, 家永貴史, 有田大作, 倉爪亮, “情報構造化環境における情報管理の一手法”, 日本ロボット学会誌, vol.26, No.2, pp.192–199, 2008.

[2] ピョユンソク, 辻徳生, 橋口優香, 永田晃洋, 中島洸平, 倉爪亮, 長谷川勉, 諸岡健一, “情報構造化アーキテクチャの提案とサービスロボットのオンライン動作計画の実現”, 第 19 回ロボティクスシンポジア講演予稿集, 6D2, pp.624–630, 2014.

[3] I. Laptev, “On Space-Time Interest Points,” Int. J. of Computer Vision, Vol.64, No.2-3, pp.107–203, 2005.

[4] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, “Visual Categorization with Bags of Keypoints,” Proc. of ECCV Workshop on Statistical Learning in Computer Vision, pp.59–74, 2004.

[5] F. Perronnin, J. Sanchez, T. Mensink, “Improving the fisher kernel for large-scale image classification,” In Computer Vision-ECCV 2010, Springer Berlin Heidelberg, pp.143–156, 2010.

[6] H. Jegou, M. Douze, C. Schmid, P. Perez, “Aggregating local descriptors into a compact image representation,” In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, pp.3304–3311, 2010.