情報構造化環境におけるスマートグラスと Faster R-CNN を用いた日常物品登録システムの開発

Automatic Houseware Registration System using Egocentric Vision and Faster R-CNN for Informationally-Structured Environment

○ 中嶋 一斗（九大） 岩下 友美（九大） 高嶺 朝理（九大） 正 倉爪 亮（九大）

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

In this paper, we present a system to register housewares in a room to database automatically to maintain an informationally-structured environment. We assume that housewares requested by a user are likely to be appeared in an egocentric vision of the user. The proposed system captures the egocentric vision by a smart glass, detects multi-class objects in images using CNN, and registers it to the database. We demonstrate the developed system enables to register several objects in a room automatically.

Key Words: Informationally-structured room, Convolutional Neural Networks, Object detection

1 はじめに

近年，急速な少子高齢化による労働力人口の減少が社会問題となっており，その打開策としてヒトと共生するサービスロボットに対する期待が高まっている．我々は，ロボット周囲の環境にセンサを分散配置することで生活空間全体の知能化を図る環境情報構造化（空間知能化）の概念に着目し，ロボットが安全で確実なサービスを遂行できる日常生活環境の構築に取り組んでいる [1]．特に，比較的容易かつ最も実現を期待されるサービスが，ロボットによる物品取り寄せである．物品取り寄せを実現するためには，対象物体が空間内のどこに存在し，どのような状態であるかといった情報が不可欠であり，空間内の各物体の位置・状態を常時計測する仕組みが必要である．

現在，我々の開発している生活支援アーキテクチャROS-TMS[2]では，日常物品に RFID タグや赤外線反射マーカを付与することで，空間内の位置情報を計測している．しかし，それぞれのセンサが特定物品の計測に特化していたり，計測範囲が限られる場合が多いため，完全に環境の状態を把握することは不可能である．計測範囲外に散在するものや，現状のシステムで必要な RFID やマーカの付与されていない物品を追跡することができない．この問題を解決し，より柔軟に物品の位置・状態を把握する一つの有効な手法として，一般物体認識が考えられる．

近年，画像や音声，文章といった多様なドメインのメディアを認識する上で，深層学習により特徴そのものを学習するアプローチが圧倒的な成果を収めている．その中でも，画像認識分野における研究は盛んで，一般物体認識や文字認識など応用先は多様である．さらに，最近では，実時間で数十カテゴリの物体を高精度に検出することのできる手法も開発され，生活空間内で取得できるカメラ画像から日用物品のカテゴリと位置を認識することは十分実現可能である．

また，ヒトの取り寄せ指示の対象となる物品は，そのヒト自身の関心が向いていることから，視界内に映る可能性が高い．そこで，居住者の装着するスマートグラスから一人称視点画像を取得し，画像内から検出される物品の種類・位置を実時間でデータベースに自動登録するシステムの開発を行う．本稿では，まずシステムの概要について述べ，日常物品自動登録の可能性を調査するための実験を行う．

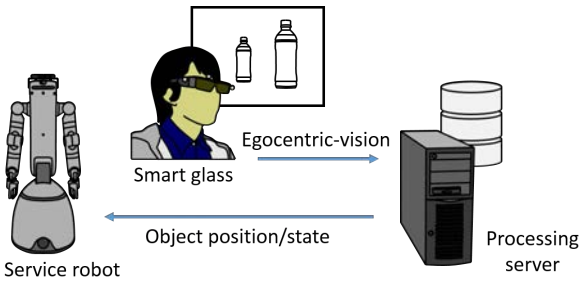

Fig.1 System configuration diagram

2 一人称視点内の日常物品検出

一人称視点から得られる画像や映像は，当該者の関心や意図を顕著に表すため，人物行動認識や意図推定の解析対象として利用されてきた．一般物体認識あるいは物体検出の対象として一人称視点画像を利用した研究も数多い．例えば，Pirsiavash ら [3]，Faith ら [4] は，ヒトの日常生活動作を正確に認識する手がかりとして，一人称視点映像中に観測される物体に着目している．また，原田ら [5] は，ウェアラブルカメラから得られた一人称視点画像に対してアノテーション及びその蓄積を行うことで，視覚記憶の拡張を図っている．

今想定している物品取り寄せタスクにおいては，対象物品の位置情報をいち早くロボットに知らせるために，居住者が指示する可能性の高いものを優先的に計測する必要がある．我々はその手がかりの一つとして，居住者の一人称視点に着目した．生活空間で活動中の居住者が関心を持つ物品は，その一人称視点内に映る可能性が高いと考えられる．指示される可能性の高い物品を一人称視点画像から積極的に検出し，物品情報データベースを逐次更新することで，即時性の高いサービスが期待できる．

近年，高性能なウェアラブルカメラが手軽に入手でき，頭部に装着することで容易に一人称視点画像を得ることができる．なかでも，スマートグラスと呼ばれるものの多くは軽量な OS をベースに動作し，マイクやスピーカ，ディスプレイといった外部デバイスとの入出力を可能にしており，可搬性の高い小型コンピュータとしての側面も強い．本研究では，ユーザインタフェースとしての応用も想定し，スマートグラスの内蔵カメラから一人称視点画像を取得する．

Fig.2 Example of detection. Faster R-CNN detected the bottle in the shelf with the score 0.999 in 0.089s

2.1 深層学習に基づく物体検出技術

一般物体認識を高精度に実現する強力な手法として，深層学習の実装の一つである畳込みニューラルネットワーク（Convolutional Neural Networks, CNN）が注目されている．さらに，CNN の入力に対象画像中の物体候補領域を与えることで，深層学習の成果を一般物体検出に応用することができる．Girshick ら [6] が提案した R-CNN(Region-based Convolutional Neural Networks) は，Selective Search と呼ばれるセグメンテーション手法で画像中から予め多数の物体候補領域を検出し，各領域のクラスを学習済み CNN で推論することで，多クラスを対象としながら高精度な物体検出を可能にしている．検出を高速化するための改善も進んでおり，Ren ら [7] の提案した Faster R-CNN では，従来ボトルネックとなっていた候補領域の検出処理さえも CNN に置き換えたことで，精度向上だけでなく，ほぼ実時間動作の物体検出を可能にした．本研究で開発するシステムでは，スマートグラスから得られる一人称視点画像に対し，Faster R-CNN による物体検出を行い，画像内の日常物品の位置と種類を取得する．また，現在の深層学習の進展を支える重要な要素の一つが，GPU をベースにした並列計算環境である．本研究では，Faster R-CNN による物体検出処理をスマートグラスと独立した処理サーバに分散させ，GPU による高速な実時間処理の実現を図る．

図 2 に，処理サーバによるボトルの検出例を示した．別のクライアント PC から UVC カメラで撮影した画像を処理サーバに送信している．この例では，処理サーバが画像を受信してから 89ms で画像内のボトルを検出した．Faster R-CNN のソフトマックス出力層によるスコアは 0.999 である．

3 スマートグラスを用いた日常物品の自動登録

一人称視点画像から検出した物品を情報構造化環境のデータベースに登録するまでの処理手順について述べる．本システムがデータベースで管理する物品情報は，物品のカテゴリと空間内の位置である．それらの物品情報を更新するために，スマートグラスの位置・姿勢と一人称視点画像内の検出物品のカテゴリ・位置を利用する．

図 3 に使用するスマートグラスを示す．スマートグラスには反射マーカを付与しており，ROS-TMS 空間内に配置した光学式モーショントラッカを用いて，位置・姿勢を計測することができる．そのため，一人称視点画像内の検出領域の位置が分かれば，スマートグラスの位置・姿勢から物品の空間内相対位置を推定できる可能性がある．

本稿では，まず日常物品の検出からデータベース登録まで実時間で可能かどうかを調査するため，データベースに登録される同一物品の重複を許し，物品情報は物体カテゴリのみとする．

4 実験

日常生活を想定した一人称視点映像を対象として，物体検出とデータベース登録の実験を行う．実験は，カメラから取得する複数フレームの一人称視点映像中に図 4 のように机の上に離れて置かれた 3 種類のボトル (2 種類のガラス瓶とペットボトル) を出現させ，検出率とフレームレート，データベースの物品登録状況を観測する．また，カメラ画質が検出精度に及ぼす影響を観測するために，スマートグラス内蔵カメラと UVC カメラの二種に対して，カメラ自体が静止状態と運動状態の時の一人称視点映像を撮影する．ただし，運動状態とは 3 種類のボトルを常に視野に含

Fig.3 Smart glass (Moverio BT-200AV, EPSON) with markers

Fig.4 Example of video captured by the smart glass

むように，カメラをランダムに動かすことである．この実験の目的は以下の二つである．

- スマートグラスの撮影画像から高精度に物体を検出できるか確認する．
- 検出した物体の識別クラスに基づいて，データベースの物品情報を実時間で更新できるかを確認する．

4.1 システム構成

システムは，スマートグラスと物体検出用処理サーバからなり，それぞれローカルネットワーク上で画像の送受信と各種処理を行うためのプロセスを実行する．スマートグラスは，図 3 に示した Moverio BT-200AV を利用した．前面に搭載した内蔵カメラから，一人称視点画像を撮影し，ネットワークに配信することで処理サーバの物体検出処理を呼び出す．処理サーバでは，画像の受信をトリガとして，Faster R-CNN による物体検出を実行する．Faster R-CNN のネットワークパラメータは，PASCAL VOC 2007 データセットで学習済みのもの [8] を利用し，検出対象クラスはデータセットに含まれる bottle クラスのみとする．また，アーキテクチャは，Zeiler and Fergus モデル [8] を利用し，GPU による並列計算を行った．UVC カメラによる撮影には，画像配信用のクライアント PC を用意し，スマートグラスと同様にローカルネットワーク上で接続している．

4.2 実験結果

実験より得られた結果を表 1，表 2 に示す．表 1 では，処理サーバと無線接続された二種のカメラについて，静止状態あるいは運動状態の時のボトルの検出率を示している．二種共に，静止状態の検出率が高い．図 5 に本システムによる検出例を示したが，(b) のようにカメラの自己運動によってモーションブラーが生じた画像について検出漏れが多く見受けられた．また，スマートグラスで撮影した画像は UVC カメラに比べてノイズが多く，カメラの動きに関わらず検出率が低い．表 2 では，二種のカメラの解像度とフレームレートを示した．実装上の違いから両者のフレームレートに僅かな差が生じたが，物品検出からデータベースへの自動登録まで十分高速に行えている．なお，一人称視点画像から検出された物品は，データベースに全て正常に登録されていることを確認した．

5 まとめ

スマートグラスから得られる一人称視点画像中から日常物品を検出し，データベースを逐次更新するためのシステムの概要について述べた．また，スマートグラスを用いた物品検出実験を行

(a) Successful case with UVC camera (b) Failure case with UVC camera (c) Successful case with smart glass

Fig.5 Examples of bottle detection using UVC camera and smart glass

Table 1 Accuracy of detection

Camera device		Number of bottles/images	Detection Rate [%]
Smart glass	static	90/30	87.8
	in-motion	90/30	53.3
UVC camera	static	87/29	100.0
	in-motion	87/29	64.4

Table 2 Frame rates

Camera device	Resolution	FPS
Smart glass	320 × 240	15.3
UVC camera	320 × 240	13.2

い，ネットワークを介して受信した一人称視点画像から，実時間で高精度に特定クラスの物品を検出できることを確認した．以下に，現状のシステムの課題と今後の予定について述べる．

5.1 課題

一つ目の課題は，本システムに実装した Faster R-CNN の識別対象クラスが本研究の課題に即して設定されていないことである．本稿で述べたネットワークは，一般に公開されたデータセットにより学習されたもので，日常物品に特化したものではない．そのため，日常物品を対象としたデータセットを独自に構築し，ネットワークを再学習する必要がある．

二つ目の課題として，スマートグラスの計算能力・カメラ性能が挙げられる．現状のシステムでは，スマートグラスで撮影される一人称視点画像は処理サーバに向けてネットワークに配信する必要があり，プライバシーの観点から好ましくない．理想的には，スマートグラス単体で物体検出からデータベースへの自動登録まで完結できると良いが，物体検出処理の計算コストが非常に高く，スマートグラス上での実時間動作は不可能である．また，実験結果に示したようにクライアント側のカメラ画質による検出精度の影響は無視できない．特に，居住者の頭部が静止していない状態では，一人称視点内の物品を追跡できない可能性がある．

5.2 今後の予定

今後は次の二点について取り組む予定である．

一つ目は，検出物品の生活空間内位置の推定である．生活空間に設置した光学式モーションキャプチャにより，スマートグラスの位置・姿勢を計測することができるため，一人称視点画像中の検出位置を使って物品までの方向を推定することができる．また，我々の開発する ROS-TMS では実際の環境と同期した仮想モデル空間を構築しているため，推定された方向とモデル表面を組み合わせて空間内の物品位置を推定できる．最終的には，本稿で述べた物体カテゴリと位置情報を併せてデータベースに自動登録する予定である．

二つ目は，検出物品に対するインスタンスレベルの識別である．現状のシステムでは，一人称視点画像中に検出される物品について登録済み物品との重複を判定することができない．データベースに登録される位置情報あるいは視覚情報を利用して，日常物品の識別機能を実装する．

謝辞

本研究は，国立研究開発法人科学技術振興機構の研究成果展開事業センター・オブ・イノベーション（COI）プログラムにより，助成を受けたものである．

参考文献

[1] Pyo, Y., Nakashima, K., Kawahata, S., Kurazume, R., Tsuji, T., Morooka, K., Hasegawa, T., “Service Robot System with an Informationally Structured Environment”, Robotics and Autonomous Systems, vol.74, Part A, pp.148–165, 2015.

[2] `https://github.com/irvs/ros_tms`

[3] Pirsiavash, H., Ramanan, D., “Detecting Activities of Daily Living in First-person Camera Views”, CVPR, pp.2847–2854, 2012.

[4] Faith, A., Farhadi, A., Rehg, J., “Understanding Egocentric Activities”, ICCV, pp.407–414, 2011.

[5] 原田達也, 中山英樹, 國吉康夫, “AI Goggles: 追加学習機能を備えたウェアラブル画像アノテーション・リト リーバルシステム”, 電子情報通信学会論文誌, Vol.J93-D, No.6, pp.857–869, 2010.

[6] Girshick, R., Donahue, J., Darrell, T., Malik, J., “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, CVPR, pp.580–587, 2014.

[7] Ren, S., He, K., Girshick, R., Sun, J., “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” arXiv, 2015.

[8] `https://github.com/rbgirshick/py-faster-rcnn`