ロバストな機器操作のための操作者特定・追跡システム

Operator Identification and Tracking for a Robust Appliance Control System using Pointing Gestures

○学　横田雅恵 (中央大)　　Sarthak Pathak（中央大）
正　梅田和昇（中央大）

Masae YOKOTA, Chuo University, yokota@sensor.mech.chuo-u.ac.jp
Sarthak PATHAK, Chuo University
Kazunori UMEDA, Chuo University

In recent years, research has been conducted on how to operate devices based on gesture recognition in order to achieve spatial interaction with devices. In this study, to make gesture recognition more robust in multi-person situations, we aim to extract only the operator's image after identifying and tracking the operator. In the proposed method, after identifying the operator by a hand-raising gesture, the operator is tracked using the coordinates of the center of the operator's head, and then the image is extracted so that only the operator's whole body is captured. In experiments, we evaluated the success of operator identification, tracking accuracy, and image extraction accuracy.

Key Words: Intelligent room, Human machine interface, Human identify and tracking, Control system of appliances

1　序論

現在，日常生活での機器操作の主な手段は機器に対応するリモコンでの操作である．しかし，リモコン操作の際には人がリモコンのある位置まで移動してから手に取って行うという負荷がある，また，どの位置からでも機器操作を行える音声認識での操作手法も普及しているが，口頭指示のみでは三次元的な指示の自由度は制限される．そこで，近年では空間的な指示を直感的に機器に伝えるため，ジェスチャ認識での操作手法について研究が行われている．

ジェスチャ認識での機器操作では，様々な手法が検討されているが，ジェスチャの中には文化や文脈によって意味が変わるものもあるため，操作と紐づけるジェスチャを適切に設定する必要がある [1]．そこで，状況や文化が異なっていても同じ意味を持つジェスチャの中から，モノを指し示すジェスチャを採用した研究が多く行われている [2]-[4]．その内，従来研究 [4] では，家電の位置情報を予め必要としない中で，物体検出と骨格点検出を融合させる手法を導入した．この手法により，腕で指し示した先の家電の電源操作を行うシステムを構築し，認識率とユーザビリティの面から良い結果が得られた．しかし，このシステムでは操作者を特定しておらず，複数人状況下でのジェスチャ認識や骨格点の誤検出に弱い．

そこで，本研究では操作者を特定・追跡することで，操作者のジェスチャのみを認識可能にし，腕指し家電操作システムをよりロバストにすることを目的とする．

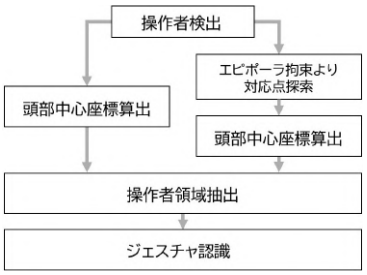


Fig.1: Flowchart of this system

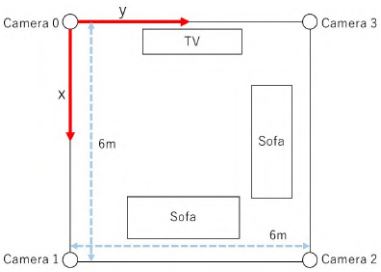


Fig.2: The overview of the room

2　提案手法

2.1　提案手法概要

提案手法の大まかな流れを図 1 に示す．本システムの環境は図 2 のようになっており，天井四隅に設置した各カメラから本システム環境の撮影を行う．得られた画像に対して，YOLO[5] を用いて挙手ジェスチャの検出，人物やその身体部位の検出，家電の検出を行う．操作者は，挙手ジェスチャを数フレーム連続で行うことで操作者として特定され，家電操作を始めることが出来る．その後，骨格点抽出を用い，腕差しジェスチャ認識により家電操作を行う．本論文では操作者領域の抽出までについて述べる．

2.2　YOLO による検出

本システムでは YOLOv4[6] を用い，挙手ジェスチャ(operator) を検出する検出器 A と，人物とその頭部，TV，ソファを検出する検出器 B を作成した．検出器 A では本システムの空間を撮影した写真でアノテーションを行い，データを拡張して 1235 組のデータセットを得た．このうち 987 組を学習に用い，平均適合率 91.5%であった．検出器 B では本システムの空間を撮影した 122 枚の写真でアノテーションを行い，データを拡張し，1128 組のデータセットを得た．このうち 902 組を学習に用い，226 組を検証に用いた．これにより，人物全体 (full)，頭部 (head)，TV(tv)，ソファ(sofa) の平均適合率はそれぞれ 89.6%，98.3%，97.8%，98.6%となり，高精度の検出器が作成できた．これらの検出器を用いることで検出した物体の周りに図 3 のようなバウンディングボックスが描画される．

2.3　操作者の特定・追跡

図 4a に示すように，検出器 A での検出結果と，検出器 B での検出結果を融合させることで操作者を特定する．operator バウンディングボックス中心が移動せず，検出が数フレーム連続で

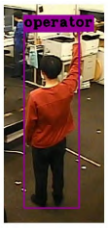


Fig.3: Object detection results: Bounding boxes

維持されることで操作者を確定する．operator バウンディングボックス中心と，検出器 B で検出した全ての full バウンディングボックス中心との 2 次元距離を計算し，距離が最も近い full バウンディングボックスを操作者のものと特定する．続いて，操作者の full バウンディングボックスの上辺中心の 2 次元座標と，全ての head バウンディングボックスの上辺の中心の 2 次元座標との距離を計算する，この距離が最も小さい head バウンディングボックスが，操作者の頭部であると特定する．

ここで，複数カメラ画像で操作者の頭部を特定できた場合，二つのカメラ画像で得られた操作者の head バウンディングボックス中心を用い，ステレオ視の原理 [7] により操作者の頭部中心の三次元座標を求める．この三次元座標を，透視投影行列を用いて各カメラ画像内の二次元座標に復元することで，操作者を検出できなかったカメラ画像内でも操作者頭部バウンディングボックスを特定する．一方で，一つのカメラ画像内でしか操作者を検出できなかった場合，エピポーラ拘束を用いて，図 4b に示す対応関係にあるカメラ画像内から対応する head バウンディングボックスを特定し，操作者の頭部とみなす．

操作者特定後は，head バウンディングボックス中心座標を用いて操作者追跡を行う．常にフレーム間での移動量から 2 次元的に追跡するのみでは人物すれ違いに対応できないため，数フレームに一度カルマンフィルタによる予測値による補正を行い，ロバストな追跡を実現する．

2.4　カルマンフィルタによる補正

ステレオ視の原理から算出した頭部中心三次元座標の値を保存し，カルマンフィルタによる予測を行う．毎フレームで予測値を各カメラ画像内での 2 次元座標に復元し，5 フレームに一度，復元した値と検出された値の差がしきい値を超えた時，予測値を復元した値を優先する．このしきい値は経験的に 40 ピクセルとした．

2.5　ジェスチャ認識

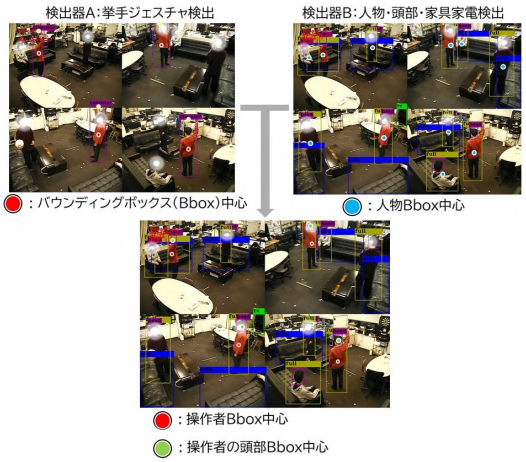
操作者の骨格点のみを検出するため，取得画像から図 5 のように操作者の画像を抽出する．ここで，操作者の伸ばした腕がバウンディングボックス外に出た際にもジェスチャ認識を行えるよう，操作者の full バウンディングボックスの高さの値を一辺とする正方形の領域で抽出を行う．

3　実験

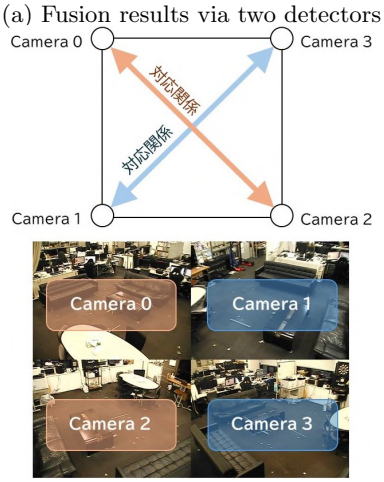
本手法の有用性を確認するため，操作者特定・追跡・操作者画像の抽出について実験を行った．実験はオフライン環境下で行った．複数人物状況下で，非操作者・操作者ともに立ち位置を指定していない．操作者にはしばらく挙手し，手を下ろして実験環境内を歩き回るよう指示した．この際，挙手を維持する時間を具体的に指示していない．5 本の動画を用いて本システムの評価を行う．5 本の動画のうち，実験環境内にいる人数での種別の内訳は，2 人の動画が 1 本，3 人の動画が 2 本，5 人の動画が 2 本である．

3.1　操作者特定の精度

検出器の融合による操作者の特定の精度を評価する．検出された operator バウンディングボックス内の人物と，中心点距離計算により特定された full バウンディングボックス内の人物が同一であることを評価する．5 本の動画のうち，4 本の動画で成功し，1 本の動画で失敗した．その内訳を表 1 に示す．この結果のうち，操作者の服装と上下とも同じ色の服装である非操作者もいたが，特定に影響はなかった．



(a) Fusion results via two detectors



(b) Correspondence relationship of cameras

Fig.4: Elements for operator identification and tracking

操作者特定に失敗した動画の一つでは，図 6 に示すように Camera0 で認識した操作者の頭部が，対応関係にある Camera2 では写っていないため，対応点の探索に失敗したとわかった．このことから，操作者が検出されたカメラが一つしかない時，対応関係にあるカメラ内から対応点を探索するのでなく，head バウンディングボックスが最も多いカメラ内から対応点を探索する方法を試す必要があるとわかった．

また，もう一つの失敗例では，操作者の頭部が検出されているのにも関わらず，操作者を特定できないという失敗が観測された．これは，どのカメラでも挙手ジェスチャが認識されない際，判定のリセットを待つ猶予フレーム数が短すぎたことが原因だと考えた．そこで，この猶予フレーム数を大きくした結果，5 人という多人数状況下でも操作者を特定することが出来た．一方で，同じ猶予フレーム数を環境内人数 3 の動画に適応すると誤特定が発生した．従って，環境内人数の大小によって猶予フレーム数を可変の値にする必要があるとわかった．

Table 1: Result of success or failure of operator identification

環境内人数	2	3	3	5	5
操作者特定	成功	成功	成功	成功	失敗

3.2　操作者追跡の精度

操作者頭部中心の軌跡の俯瞰図を図 7 に示す．青色の線が検出結果による三角測量で得られた軌跡であり，オレンジ色の線がカルマンフィルタによる予測値の軌跡である．操作者特定に成功した動画を目視で確認したところ，追跡中に操作者以外の人物を操作者と誤認した誤追跡は観測できなかった．移動方向の転換や他

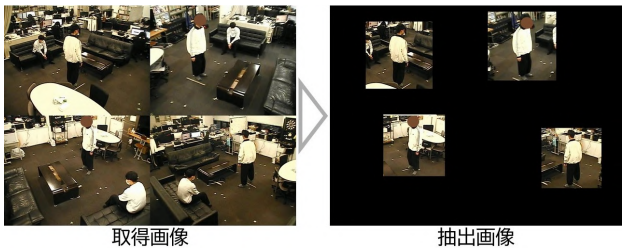


Fig.5: Extracted result: Operator's ROI

Fig.6: Example of operator identification failure

の人物とのすれ違いが発生した際にも，操作者を追跡することが出来た．また，色情報を使っていないため，上下とも同じ色の服装をした人物の有無に追跡精度は影響を受けなかった．加えて，カルマンフィルタでの補正により，操作者の移動の速度の変化にも対応できていることを確認した．

これは，身体部位の中でも部位自体の面積が小さく，バウンディングボックス中心が大きく変化しにくい頭部に着目したことが要因と考えられる．また，三次元座標を用いてカルマンフィルタによる補正を行ったことで，すれ違いの際に，止まっている人物との差別化が上手くいったと考えられる．

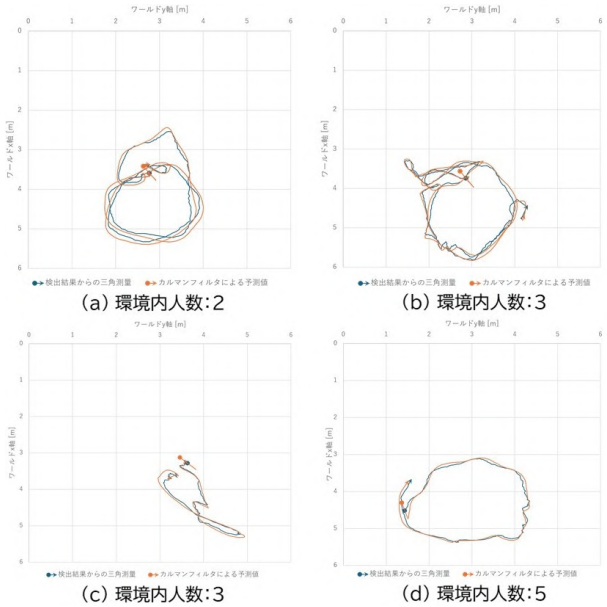


Fig.7: Overhead view of operator tracking

3.3 操作者領域抽出の精度

追跡結果をもとに，操作者を抽出する．ここでは操作者の full バウンディングボックスの高さを一辺とする正方形領域で抽出を行う．操作者の頭部中心の 2 次元座標とすべての人物の full バウンディングボックス中心の 2 次元座標を比較し，x 軸方向距離が最も近い full バウンディングボックスを操作者の full バウンディングボックスとみなす．図 5 で示した抽出画像において，操作者以外の人物を抽出してしまうとジェスチャの誤認識の要因となるため，操作者以外の人物を抽出しているカメラが一つでもある場合を失敗フレームとした．抽出画像をフレームごとに目視で確認し，失敗フレームの数をカウントした．操作者特定後の動画のフレーム数から失敗フレーム数を引いたものを成功フレーム数とし，その割合を抽出の成功率とした．特定に成功した動画の中での操作者領域の抽出の成功率は表 2 に示すとおりである．また，全フレーム中失敗したフレームを赤で示した図が図 8 である．

環境内人数が増えるほどすれ違いの頻度が増えるため，抽出成功率は下がると考えられ，妥当な結果といえる．しかし，どの失敗フレームでも 2 つ以上のカメラで操作者領域の抽出が成功していた．

Table 2: The accuracy of operator's image extraction [%]

環境内人数	2	3	3	5
操作者特定	93.8	89.1	88.4	70.7

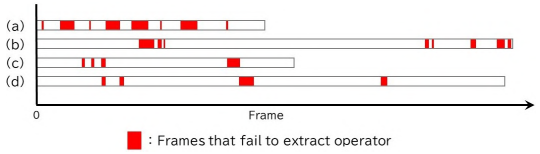


Fig.8: Frames in the video where operator extraction fails

4 結論

本研究では，腕指しでの空間的機器操作システムのロバスト化のため，従来手法での課題であった複数人状況下での操作者の特定と追跡の手法を提案した．操作者の骨格点のみを抽出するために，追跡結果から操作者の画像の抽出を行った．実験により，特定の状況を除いて操作者を特定することができ，かつ高精度の追跡を行えたことを確認できた．また，2 5 人の複数人環境下で 70%以上の精度で操作者画像の抽出が行えたことを確認した．今後の課題と展望を以下にまとめる．

- 特定失敗の際のリカバリー
  実験では操作者の特定の失敗が観測できた．これはエピポーラ拘束を適用するカメラの組み合わせを変えることで対応できると考えられる．
- 操作者画像の抽出について
  人数が増えるにつれ操作者のみを抽出することは困難になる．追跡に用いた点以外の情報を利用し，重みづけを工夫することで対応できる可能性がある．
- 今後の展望
  移動し続ける機器を腕差しした地点へ向かわせ，タスクを割り振る研究への発展が見込める．操作者を特定することで，個人専用の機器の操作なども可能になる．また，現在はカメラの位置情報を既知のものとして利用しているが，今後は震度カメラなどを用い，カメラの位置関係が未知でも本システムによる追跡手法が対応できることを目指す．

参考文献

[1] M. S. Verdadero, C. O. Martinez-Ojeda and J. C. D. Cruz, "Hand Gesture Recognition System as an Alternative Interface for Remote Controlled Home Appliances," 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM), Baguio City, Philippines, pp. 1-5, 2018.

[2] H. Asano, T. Nagayasu, T. Orimo, K. Terabayashi, M. Ohta and K. Umeda, "Recognition of finger-pointing direction using color clustering and image segmentation," The SICE Annual Conference 2013, Nagoya, Japan, pp. 2029-2034, 2013.

[3] A. I. D. Viaje, P. S. Bernardo, K. N. Manuel, G. M. Pacheco, K. -R. C. Barroma and R. E. Tolentino, “Selection of Appliance Using Skeletal Tracking of Hand to Hand-tip for a Gesture Controlled Home Automation,” 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, pp. 575-580, 2020.

[4] M. Yokota, S. Majima, S. Pathak and K. Umeda, ”Intuitive Arm-Pointing based Home-Appliance Control from Multiple Camera Views,” 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2023.

[5] 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.

[6] A. Bochkovskiy, C. Wang and H. M. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” arXiv preprint arXiv:2004.10934, 2004.

[7] R. Hartley and A. Zisserman, “Multiple View Geometry in Computer Vision,” Cambridge University Press, ISBN: 0521540518, second edition, 2004.