5ZD-05

RGB-D センサの深度データを入力とした 深層学習による指差しジェスチャの検出方法の検討

野田雄希† 田村律起‡ 水谷晃三†‡

帝京大学大学院理工学研究科† 帝京大学理工学部情報電子工学科‡

1. はじめに

人が行う指差しジェスチャをセンサで検出し，その指が差す方向にポインタを表示したり，ジェスチャに応じてコマンドを入力したりする方法がある．筆者らは天井に下方に向けて設置した RGB-D センサを用いてこれを実現する方法を検討し，指差しポインティングシステムを試作した[1]．本研究では，試作システムの指差しジェスチャの検出精度を向上させるために，RGB-D センサの深度データを直接物体検出モデルに与え，検出する方法を検討する．

2. 天井に下方に向けて設置した RGB-D センサによる指差しポインティングシステム

2.1 概要

指差しポインティングに関する既存研究には，ユーザの正面や側方のセンサで指差しジェスチャを捉える手法[2][3]がある．これに対して，筆者らの方式では RGB-D センサを天井から下方に向けて設置する（図 1）．既存研究の手法と比べて複数ユーザ同士が重なるオクルージョンの発生が軽減されるため，指差しの向きによらず複数人の指差しジェスチャを同時に捉えることができる（図 2）．試作システムでは RGB-D センサ（Kinect v2）を床から 2.5m の高さに設置した．

2.2 深度画像を用いた指差しジェスチャの検出

試作システムでの指差しジェスチャの検出には，RGB-D センサから取得した深度データをあらかじめ決めた深度範囲でグレースケール画像化した深度画像を用いる．これを OpenCV のカスケード分類器に入力して，画像中の指差しジェスチャを行っている手の領域の位置と大きさを取得する．

深度画像は各画素の深度値が 256 段階の濃度で表現される．分類器が学習した深度画像は，ユーザの上半身が写る深度範囲（センサから 750mm～1600mm）を 256 段階で表現したものである．従って，約 3mm の深度値の差は画像上で同じ濃度となり，深度情報が欠落することになる．深度範囲が広いほど，この欠落による認識精度への影響が大きくなる．実際，ユーザの身長の違いや起立/着席状態を含むように深度範囲を広げると，分類器が学習した深度画像より手の凹凸形状の情報が欠落した画像が入力され，検出精度が低下する問題があった．

そこで，深度範囲を複数の層に分け，各層で深度画像を作成して深層学習により検出する方法を検討し，実用化を踏まえた深度範囲（600mm～2190mm）における検証したところ精度の改善が確認された[4]．しかし，この方式では層の数だけ深度画像の生成や検出処理が必要であり，これによる処理負荷の増大が試作システムへの実装における問題となっていた．

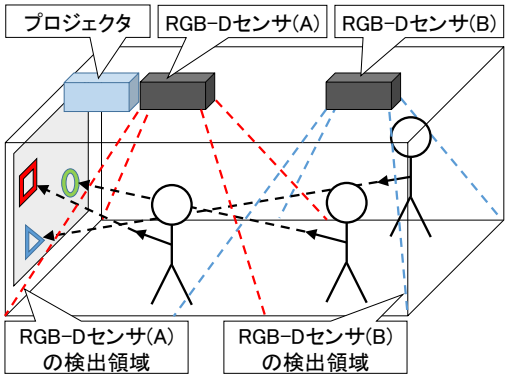

図 1. システムの概要図

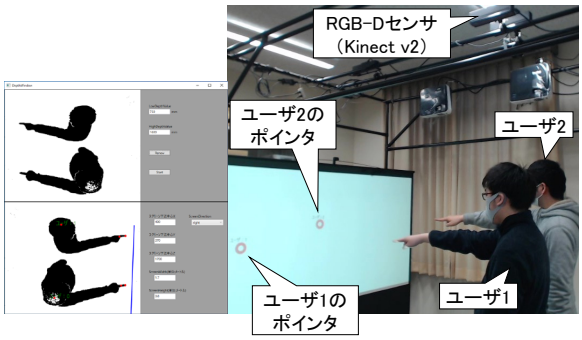

図 2. 複数人でのシステムの動作例
（左：システム画面，右：使用時の様子）

A Study of Hand Pointing Detection by Deep Learning with Depth Data from RGB-D Sensors
Yuki Noda†, Ritsuki Tamura‡, Kozo Mizutani†‡
† Graduate School of Science and Engineering, Teikyo University.
‡ Department of Information and Electronic Engineering, Faculty of Science and Engineering, Teikyo University.

3. 深度データを用いた深層学習による指差しジェスチャの検出方法

前述の問題に対し，本研究ではセンサから取得した深度データを物体検出モデルとしてのNeural Network（NN）に直接与えることで解決を試みる．深度画像を生成する方法に比べて，深度データを直接与える方法は画像化などの処理不要になるため比較的負荷が低くなると考えられる．また，手の凹凸形状の特徴を失わずにすむため検出精度が高まる可能性も考えられる．

本研究では，起立状態の上方向への指差しから着席状態の下方向への指差しまでが含まれる深度範囲500mm～2000mmを検出対象とする．指差しジェスチャの写った深度データを検出モデルに与えることで，指差しジェスチャを行っている手を検出する．深度データとの比較用として，試作システムの分類器への入力と同じ1層の深度画像による検出モデルも作成する．

4. 検出モデルの作成と結果

試作システムと同様に床から 2.5m の高さに設置した Kinect v2 から取得した深度データを用いて，各検出モデルの学習データを作成した．各モデルの学習データは同一深度データから作成し，3名の被験者から2268個のデータを得た．作成したデータは，手の形が指差しジェスチャ・グー・チョキ・パー，右手・左手，手の方向が水平・上・下，起立状態・着席状態の組み合わせ全 48 通りの内容となっている．各データは人物が1人のみ，手が1つだけ写っている．これらデータを学習用 1450 個，検証用 400 個，評価用418個に無作為に分けた．

検出モデルとして EfficientDet D0[5]を用い，転移学習ありとなしの2パターンの検証を行った．COCO Object Detection Challenge の評価指標を用いた各モデルの評価結果を表 1 に示す．1つのデータあたりの最大検出数を1として評価した．

5. 考察

本実験では画像用の物体検出を行うモデルに深度データを与える方法を試みた．表1に示すように，転移学習ありの場合は深度データの方が良い精度となり，本手法により指差しジェスチャを行う手の領域を検出できることを確認した．

一方，転移学習なしのときの結果を含めると，深度画像を用いるモデルが最も良い精度となった．この一因として，学習データのバリエーション不足が考えられる．深度範囲が広いため，手の高さや傾き，手の凹凸形状などの個人差が画像上の階調として表現されにくく，物体検出モデルにとっては学習しやすい画像となった可能性がある．一方，深度データを用いる方法では個人差が直接影響するため，本実験で用いたデータでは十分な精度を得られなかったものと考えられる．実用化を踏まえると，個人差に関わらず認識できることが望まれるため，データの見直しが必要である．また，深度データによる方法の精度を高めるためには，深度データに適したモデルの実装も有効であると考えられる．

表 1. モデルの評価結果（深度データ / 深度画像）

	IoU	転移学習あり	転移学習なし
AP	0.50～0.95	0.604 / 0.330	0.583 / 0.677
	0.75	0.703 / 0.237	0.637 / 0.883
AR	0.50～0.95	0.653 / 0.413	0.643 / 0.734

6. おわりに

本研究では，天井に下方に向けて設置したセンサにより指差しジェスチャを検出する方法として，物体検出モデルにセンサの深度データを直接入力する方法を検討した．深度データを直接用いることにより，処理負荷の低減や検出範囲の拡大が期待できることを述べた．

謝辞

本研究の一部はJSPS科研費 21K12163の助成を受けた．

参考文献

[1] 野田雄希, 水谷晃三: 天井から下方に向けて設置した RGB-D センサによる指差しポインティングの研究, 情報処理学会第 83 回全国大会講演論文集, 5ZB-08, 2021.

[2] Dai Fujita, Takashi Komuro: Real-time 3D Hand Pointing Recognition using Appearance Difference between Two Camera Images, The 3rd IAPR Asian Conference on Pattern Recognition (ACPR 2015) Program Booklet, pp.36-37, 2015.

[3] K. Hu, S. Canavan, L. Yin: Hand Pointing Estimation for Human Computer Interaction Based on Two Orthogonal-Views, Proceedings of International Conference on Pattern Recognition, pp. 3760–3763, 2010.

[4] 野田雄希, 水谷晃三: 天井部に設置した RGB-D センサを用いた深層学習による指差しジェスチャ認識方法の検討，第 20 回情報科学技術フォーラム 講演論文集， J-012, 2021.

[5] Mingxing Tan, Ruoming Pang, Quoc V. Le; EfficientDet: Scalable and Efficient Object Detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10781-10790, 2020.