J-012

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
A Study of Hand Pointing Detection by Deep Learning using RGB-D Sensors Placed on the Ceiling

野田 雄希† 水谷 晃三†
Yuki Noda Kozo Mizutani

1. はじめに

人が行う指差しジェスチャを認識して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．筆者らは天井から下方に向けて設置した RGB-D センサを用いて，指差し方向にポインタを表示する指差しポインティングシステムの開発を行った[1]．センサを天井から見下ろす形で設置することで，立ち位置の制約を改善しながら複数人が同時にシステムを使用できることを示した．一方で，指差しジェスチャを認識できる高さ方向の範囲が固定であるため，使用者の身長の違いや起立/着席状態の違いが生じると使用できなくなる問題があった．そこで本研究では，RGB-D センサから取得した深度データをいくつかの深度範囲に分割し，分割したデータごとに深層学習により認識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 システム概要

指差しポインティングに関する既存研究として，ユーザの正面または上方のカメラと側方のカメラによってジェスチャを捉えるもの[2, 3]や，ユーザごとにセンサを卓上に配置することで使用位置に制限なくジェスチャを捉えることができるもの[4]が存在する．これらの手法では複数人で使用する上で，センサから見てユーザ同士が重なるオクルージョンが生じたり，機材をユーザごとに配置する必要があったりするなどの問題がある．そこで先行研究では，室内全体で複数人が同時使用できる指差しポインティングシステムの実現方法について検討した．

先行研究における指差しポインティングシステムの概要を図 1 に示す．天井から真下に見下ろす形で RGB-D センサを設置することで，オクルージョンの発生を軽減している．実際の試作システムでは図 2（右）のように床から 2.5m の位置に下方に向けてセンサを設置する．ユーザの指差しジェスチャの認識には，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値の範囲でグレースケール画像化した画像を用いる．これを深度画像と呼ぶ．深度画像中の指差しジェスチャを行っている手の領域をカスケード分類器で検出し，その検出領域を元に，画像中のユーザの人体部分の領域を決定する．その後，各ユーザの頭頂部・指先・指の付け根の座標を取得して指差しジェスチャを認識し，指が差しているスクリーン上の位置を推定し，ポインタを表示する．システムを使用するときは，図 2（左）のシステム画面で分類器に入力する深度画像の深度範囲やポインタを表示するスクリーンの位置を設定する必要がある．

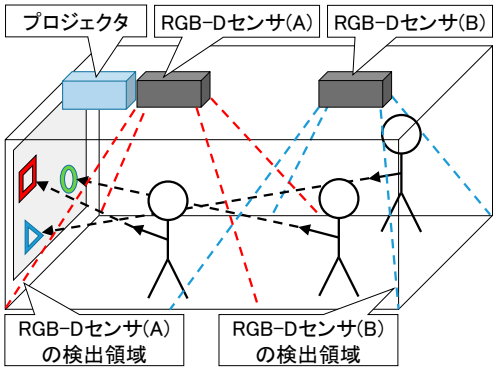

図 1　システム概要図

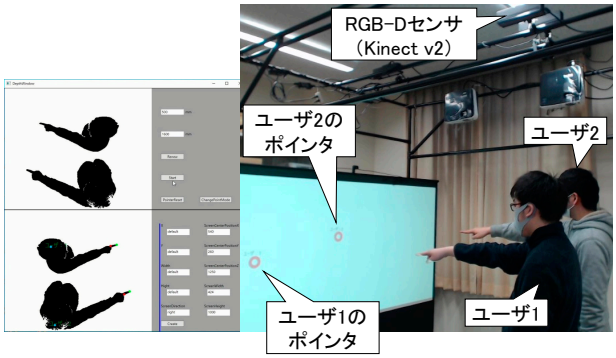

図 2　システム動作例
（左：システム画面，右：システム使用例）

2.2 カスケード分類器による手の領域の検出

指差しジェスチャを行っている手の検出には OpenCV のカスケード分類器を使用している．深度画像を分類器に入力して画像中の手を検出し，手の領域の位置と大きさを取得する．カスケード分類器は，検出対象が写った正例画像と検出対象が写っていない負例画像を使って学習する．正例画像には指差しジェスチャを上から捉えた深度画像，負例画像には指差しジェスチャが写っていない画像を用意する．先行研究ではユーザは立って使用することを想定していたため，正例画像は立っているユーザの上半身が写る範囲（センサから 750-1600mm）を深度画像化したものであった．そのため，使用者の身長の違いや起立/着席状態の違いが生じると使用できなくなり実用的でなかった．本稿ではこの解決手段として深層学習を適用する方法を検討する．

3. 深層学習を用いた手の領域の検出方法の検討

深層学習による手の領域の検出の方法には，RDB-D センサの深度データを直接ニューラルネットワーク（以下 NN

† 帝京大学大学院理工学研究科
Graduate School of Science and Engineering, Teikyo University

と略す）に与える方法と，先行研究と同様に深度画像を生成してからこれを NN に与える方法が考えられる．画像からオブジェクトを検出する方法は様々な分野に応用され実績が多いことから，本稿では後者の方法を取ることとした．

深度データの画像化の際，画像化の対象となる深度値の範囲が狭いほど，物体の凹凸形状が濃淡で表現されやすくなる．本研究の目的においては深度値の範囲を広げる必要があるため検出精度の悪化が予想される．そこで，目的とする深度値の範囲（本研究ではセンサから 600-2190mm とした）を複数層に分割して作成した深度画像を使って手の検出モデルを生成することを試みる．層の分け方を検討するために，4・2・1 層で分けたときのそれぞれの深度画像を使用して手の検出モデルの作成，検証を行う．各層の深度値の範囲は表 1 の通りである．層が複数ある 4 層と 2 層ではいずれかの層の深度画像に必ず手全体が写るようにするため，各層を 150mm ずつオーバーラップさせてある．

4. 検出モデルの生成と検証

4.1 検出モデルの生成

手を検出するモデルの作成には，TensorFlow 上で利用できるフレームワークである TensorFlow Object Detection API を使用する．また，今回は同 API で公開されている物体検出の学習済みモデル EfficientDet D0 512x512 を使用した転移学習と転移学習なしの 2 パターンで学習を行う．各層のモデルにおいて 20000 ステップの学習を行った．

学習画像には複数層の深度画像のうち，手の写っている層の画像のみを使用する．深度データを床から 2.5m の位置に設置した Kinect v2 で取得して深度画像を作成した．学習画像の内容は，右手・左手，指差し方向が上・正面・下，起立状態・イスに座っている状態の組み合わせ全 12 通りの深度画像 750 枚であり，画像中には手が 1 つだけ写っている．深度画像の具体例を図 3 に示す．画像解像度は縦 424×横 512 ピクセルである．750 枚のうち，550 枚を学習用，200 枚を検証用にランダムに分けて学習・検証を行った．画像によっては濃淡の偏りがあったため，前処理として深度画像に正規化処理を行った画像を学習・検証に使用した．

4.2 検出モデルの検証

各モデルの検証結果について，転移学習による結果と，転移学習を用いずに用意した画像データのみを用いて学習した結果を表 2 に示す．検証は COCO Object Detection Challenge の評価指標とこれに基づく評価プログラムを用いた．AP はモデルが検出した手の領域（予測領域）と実際に手が存在する領域（正解領域）の重なり（IoU）を求め，正しく検出された領域の割合を算出したものである．IoU はモデルが予測した予測領域と実際に対象物体がある正解領域がどの程度重なっているかを表す指標であり，AP の算出において予測領域が正しく検出できたかどうか判定するための閾値として用いられる．本研究の目的においては高い精度で手の領域を検出することが必要であるため，IoU を 0.50-0.95 の範囲で変化させながらその平均を用いる COCO の評価指標に加え，0.75 のときの結果にも注目することにした．AR は正解領域に対して正しく検出できた予測領域の割合を求めたものである．使用した評価プログラムの仕様上，AR については IoU が 0.50-0.95，画像あたりの最大検出数を 100 としたときの結果を用いた．

表 1　各層の深度値範囲(mm)

	4 層	2 層	1 層
1 層目	600-1110	600-1470	600-2190
2 層目	960-1470	1320-2190	—
3 層目	1320-1830	—	—
4 層目	1680-2190	—	—

※同一フレームを層分けした各層のうち手の領域を含む深度画像の例(一部分)

図 3　学習に用いた深度画像の例

表 2　各モデルの評価結果（転移学習あり/なし）

	IoU	4 層	2 層	1 層
AP	0.50-0.95	0.680 / 0.651	0.644 / 0.613	0.668 / 0.625
	0.75	0.922 / 0.846	0.811 / 0.739	0.873 / 0.712
AR	0.50-0.95	0.733 / 0.698	0.690 / 0.685	0.716 / 0.683

5. 考察

転移学習あり・なしどちらにおいても 4 層の場合が最も良い評価となった．しかし，層の数が多いほど手の特徴を学習しやすく検出の精度が向上するという仮説に反して，転移学習ありにおいては 1 層の方が 2 層より良い結果となった．2 層は 1 層と同様に手の凹凸形状が濃淡の差として表現されにくいことや，多層に分けることで胴体部分が部分的に映り込むことで手の形状に似た部分が生じ，これが誤検出となることが影響していると考えられる．

6. おわりに

本稿では，天井から下方に向けて設置した RGB-D センサにより指差しジェスチャを検出する方法に関して，深層学習を適用する方法を検討した．手を検出させようとする範囲を複数の層に分割して検出モデルを作成し，その検証を行った結果，4 層モデルが最も評価の高い結果となった．本モデルを用いることで，先行研究で試作した指差しポインティングシステムの改善が期待できる．

謝辞

本研究の一部は JSPS 科研費 JP18K11580，21K12163 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] Shun Sekiguchi, Takashi Komuro: A Tabletop Projector-camera System for Remote and Nearby Pointing Operation, Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, pp. 1621–1626, 2015.