歩容に基づく個人識別における Kinect と OpenPose の多人数追跡評価

當麻 僚太郎[1,a)]　谷口 輝海[1,b)]　菊池 浩明[1]

概要：歩容は人の歩き方の特徴を表し，それゆえに個人情報に分類されている. 歩容を用いた個人識別には，深度センサなどのハードウェアを用いる方法や歩行のシルエット画像列を用いる方法が知られている. しかし，これらの研究では，単独での歩行に限られており，画角内に複数人が同時に映っている場合の評価が明らかにされていない. ハードウェアに対する依存度も不明である. そこで，本研究では，複数人を同時にリアルタイムで検出する機能を持つ Kinect と OpenPose に着目し，それぞれ同時に何人まで識別，追跡できるかを明らかにする.

Multiple person tracking based on gait identification using Kinect and OpenPose

Abstract: A gait provides the characteristics of a person's walking style and hence is classified as personal identifiable information. There have been several studies for personal identification using gait, including works using hardware such as depth sensors and studies using silhouette image sequences of gait. However, these methods were designed specialized for tracking a single walking person and the accuracy reduction when multiple people are simultaneously reflected in several angles of view is not clear yet. In addition, dependencies on hardware-based methods is not clarified yet. In this study, we focus on Kinect and OpenPose, the representative gait identification techniques with a function to detect multiple people simultaneously in real time. We investigate how many people can be identified for these devices and with the accuracy for tracking.

1. はじめに

人の歩き方の特徴を表す歩容は解像度の低いカメラ映像からでも取得できることから，犯罪捜査などの分野において，個人識別新たな手段として，近年注目されている. 歩容に基づく属性推定・個人識別手法には，ウェアラブルデバイスなどの特定のハードウェアを本人に装着する方法 [8] や，歩容のシルエット画像列などを用いて外部から観測する方法 [2] が知られている. しかし，特定のハードウェアを用いる方法は使用場面が限られ，シルエット画像列を用いる方法では服装や髪型，携帯品などの外乱の影響を受けやすいという問題がある. 今日，注目されているのは，街中に設置された防犯カメラ映像や深度センサー，および，LiDAR などのセンシングデバイスを用いて歩行などを観測する技術である. 従って，画角内に複数の人間が映っていることが想定される. しかしながら，多人数の同時追跡に対する性能は，用いるデバイスの特性に大きく依存すると考えられ，明らかにされていなかった. これらのデバイスには，次のような特徴がある.

Kinect　深度センサなど複数センサを備えており，正確に 3D 座標を測定できるが，距離や角度に制限がある.

OpenPose　画質さえ良ければ，距離や角度の変化に対して頑強であるが，2D 座標から 3D 座標を推定する必要があるため，精度に課題がある.

従って，多人数のオブジェクトを正確に追跡するにはどちらのデバイスが向いているのか，自明ではない.

そこで，本研究では，複数人数を同時にリアルタイムで検出する機能を持つ，Kinect[5] と OpenPose[3] に注目する. Kinect と OpenPose について，何人まで同時に識別できるか，どの程度の精度で識別できるかを明らかにすることを目的とする. 本研究の個人識別の構成と流れを図 1 に示す. 図 1 に示すように，本研究では同じオブジェクトを Kinect(緑) と OpenPose(青) の異なる方式で観測する.

[1] 明治大学 総合数理学部
School of Interdisciplinary Mathematical Science, Meiji Uniersity
[a)]ev200598@meiji.ac.jp
[b)]ev200594@meiji.ac.jp

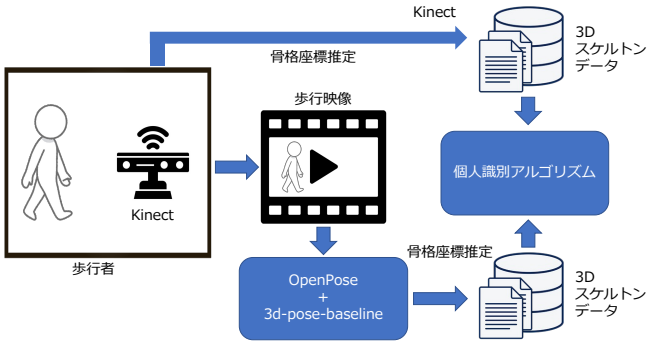

図 1　個人識別システム構成図

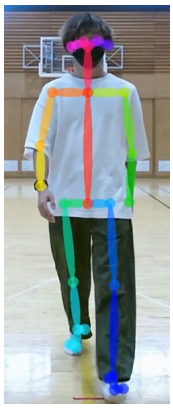

図 2　OpenPose の姿勢推定例

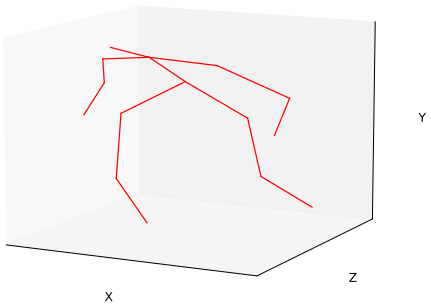

図 3　3d-pose-baseline の 3 次元姿勢推定例

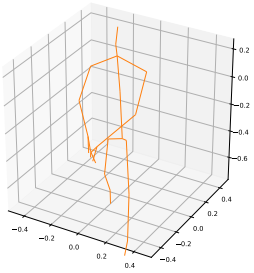

図 4　Kinect による 3D スケルトンデータの例

2. 準備

2.1 OpenPose

OpenPose[3] は，Zhe らによって開発されたオープンソースである．静止画像または動画からリアルタイムに複数人数の 2D 姿勢推定を行う深層学習モデルである．姿勢推定 (Human Pose Estimation) では，人の頭部，肩，肘，手，腰，膝，足を検出し，人がどのような姿勢を取っているかを推定する．OpenPose は深度センサなどの特別な機器を必要とせず，単眼カメラのみで姿勢の推定ができ，25 点を検出できる．図 2 に推定結果のプロット例を示す．多人数を同時に検出できる利点の一方で，大きな計算コストを要し，人数の増加に伴う推定精度の劣化は明らかではない．

2.2 3d-pose-baseline

3d-pose-baseline[4] は，Julieta らによって開発されたオープンソースであり，OpenPose の出力を入力とし，深度を推定する深層学習モデルである．2 次元画像や動画から 3 次元の姿勢を推定する．

本研究では，OpenPose と 3d-pose-baseline を用いて歩行映像から 3 次元の骨格座標データを取得する．図 3 に 3d-pose-baseline の出力を 3D プロットした例を示す．

2.3 Kinect

Kinect[5] は Microsoft 社によって開発されたゲーム向けデバイスである．RGB カメラと深度センサ，マイクを備え，姿勢推定や音声認識を提供する．姿勢推定で検出される関節の数は一人当たり 25 点であり，手指検出や手のポーズ検出ができる．個人を追跡する機能を持ち，最大で 6 人までを同時に検知する．

本研究では，Kinect for Windows v2[5] を用いて映像と姿勢情報の取得を行った．図 4 に Kinect で取得した姿勢情報の 3D プロット例を示す．

2.4 推定器の比較

表 1 に本研究で用いた姿勢推定ツール Kinect, OpenPose, 3d-pose-baseline の機能や特性を示す．表 1 より，Kinect はゲームでの利用を想定しているため，最大 6 人までを追跡できるように設計されている．深度センサのセンシング距離や角度にも制約がある．

一方，OpnePose には，カメラの解像度の範囲内で人数の制約はなく，距離にも頑強性があると考えられる．しかし，2D 画像から 3D 画像を推定するため，Kinect ほどの精度は期待できない．以上の考察より，両ツールの間には表 2 の関係があると考えられる．

なお，本研究では OpenPose と 3d-pose-baseline を組み合わせて姿勢の 3 次元情報を取得するため，この 2 ツールをまとめて OpenPose と呼称することとする．

表 1 ツールの機能比較

ツール	出力	特徴	用途	原理
Kinect	カラー画像: 1920×1080 深度画像: 512 × 424 FPS:30 1 人あたり 25 関節	処理が高速なため, リアルタイムでの追跡が可能. 最大で 6 人までを同時に追跡可能	リアルタイム骨格検出 (3d) 音声認識	深度センサを用いた RandomForest による推論
OpenPose	関節位置座標 2D データ 25 関節 各関節の推論信頼度	ハードウェア不要 多人数の同時検出が可能 最大追跡人数は不明 人物追跡ができない	2D 姿勢推定	深層学習
3d-pose-baseline	関節位置座標 3D データ 16 関節	OpenPose の出力から深度を推定する	3D 姿勢推定	深層学習

表 2 個人識別における各ツールの特性の予測

予想	正面	側面	小人数	大人数 (6 人以上)
Kinect	○	×	○	×
OpenPose	△	△	△	△

2.5 関連研究

2.5.1 歩容データからの性別属性推定

三好ら [9] は，Kinect で得た歩容データから身体の各部位に関する特徴量を定義し，男女の平均の中央値を閾値として性別の推定を行なった. また，定義した特徴量を統合することにより，最大 99.86%の推定精度を達成した.

2.5.2 GEI を用いた CNN による年齢推定

阪田ら [10] らは，歩容のシルエット画像列 GEI(Gait Energy Image)[2] を用いて年齢の推定を行なった. まず，CNN(Convolutional Nural Network) を用いて性別や，年代を推定し，そのあとで年齢を推定する多段階のモデルを構築している. その結果，年齢の平均絶対誤差が 5.83 歳となり，既存研究を大きく上回る性能を示した.

2.5.3 深度センサ Kinect を用いた個人識別

森ら [11] の研究では，合計 145 名の外乱を含む 3 次元歩容データを Kinect を用いて取得し，DTW 距離と kNN を用いた個人識別手法を提案している.

また，この手法では，6 つの各関節の特徴量を統合する方法を提案し，特徴量の統合によって EER(Equal Error Rate)=0.048 で個人識別が可能であることが実証されている. さらに，箱を運んでいたり，歩きスマホなどをしていて，意図的に追跡防止行為がされている場合の頑強性も従来手法の精度を上回る結果となっている.

2.5.4 カメラ画像から複数物体追跡

Nicolai ら [12] は，複数物体追跡アルゴリズムである Simple Online and Realtime Tracking (SORT) に外観情報を付加することで，物体同士の交差や遮蔽に頑強な複数物体追跡アルゴリズム Deep SORT を提案している. この手法では，実験的評価により，交差による追跡 ID の入れ替わりを従来手法から 45%減らすことが示された.

3. 個人識別

3.1 個人識別手法

本研究では，森ら [1] の手法に従い，個人識別を行なう. 森手法では，Kinect と OpenPose を用いて取得した関節の 3 次元座標をそれぞれ測定し，一歩分の時系列データの DTW（Dynamic Time Warping）距離を算出して個人識別を行う. 識別手法は以下の 4 ステップから成る.

(1) サイクル切り出し

(2) 関節座標の相対座標化

(3) DTW 距離の計算

(4) 個人識別

3.2 サイクル切り出し

身体の部位 ℓ の時刻 t における 3 次元空間の絶対座標を $a_\ell(t) = (x, y, z)$ とする．ここで，時刻 t の単位はフレームレートに対応する. 測定時間の絶対座標の時系列データ $\langle a_\ell(t_1), a_\ell(t_2), \ldots \rangle$ から歩行の 1 サイクル分を抽出する.

まず，時刻 t の左右の足の絶対座標 $a_{LF}(t), a_{RF}(t)$ から，両足の間の間隔を

$$\Delta(t) = sign \cdot \|a_{RF}(t) - a_{LF}(t)\|$$

により計算する. ここで，$sign$ は $\{-1, +1\}$ の値を取る符号であり，右足が前の状態を正とする.

次に，両足間の距離 $(\Delta(1), \ldots, \Delta(n))$ の時系列データにフーリエ変換を適用し，全周波数成分の 1/30 の低周波数成分のみを残して，残りを 0 とする．すなわち，ローパスフィルタをかけることでノイズを除去し，そのピーク間を 1 サイクルとする.

3.3 関節座標の相対座標化

歩行中の各関節の座標について，身体の中心付近に位置する比較的安定した関節が原点となるように相対座標化を行う．

関節 ℓ の時刻 t における絶対座標を $a_\ell(t)$，中心の関節の時刻 t における絶対座標を $a_c(t)$ とすると，相対座標 r は

$$r_\ell(t) = a_\ell(t) - a_c(t)$$

と定める．本研究において，身体の中心 c は Kinect で Spine-Mid，OpenPose で Spine(脊椎) を用いた．

3.4 DTW 距離の計算

DTW 距離 [7] は，2 つの時系列データ間の類似度の 1 つである．各時系列データの類似度を DTW 距離を用いて定める．本研究では，2 点間の距離として 3 次元ベクトルのユークリッド距離

$$\|\mathbf{p}_i - \mathbf{q}_i\| = \sqrt{(p_{i,x} - q_{i,x})^2 + (p_{i,y} - q_{i,y})^2 + (p_{i,z} - q_{i,z})^2}$$

を用いる．歩行 1 サイクルの関節 ℓ の 2 つの時系列データ $R_\ell = \langle r_\ell(t_1), ..., r_\ell(t_n) \rangle$ と $R'_\ell = \left\langle r'_\ell(t_1), ..., r'_\ell(t_{n'}) \right\rangle$ の DTW 距離 $d(R, R')$ を R と R' の類似度とする．DTW 距離の性質から，$R = R'$ ならば $d(R, R') = 0$ であり，n と n' は一致する必要はない．

また，複数の関節を用いたときの類似度は次のように定める．異なる関節 m と関節 ℓ について 2 つの時系列データ (R_ℓ, R_m) と (R'_ℓ, R'_m) があるとき，統合 DTW 距離 $D((R_\ell, R_m), (R'_\ell, R'_m))$ は，ℓ と m についての DTW 距離の L2 ノルム (ユークリッド距離)，すなわち，$\sqrt{d(R_\ell, R'_\ell)^2 + d(R_m, R'_m)^2}$ とする．同様に，k 種の関節を統合した場合も，k 次元のユークリッド距離で類似度を定める．

3.5 個人識別

3.5.1 単独歩行

ある単独歩行のデータに対して，その他の単独歩行のデータ全てとの間で DTW 距離を計算し，最も DTW 距離が小さかった歩行データの該当者を識別結果とする．すなわち，サイクル切り出しと相対座標化を行った単独歩行のデータ N 組のうち i 番目の歩行データを W_i と表したとき，次の問題の解 j^* が識別結果である．

$$j^* = \underset{i \neq j \in \{1,...,N\}}{\arg\min}\ D(W_i, W_j)$$

3.5.2 複数人歩行

単独歩行の識別と同様にして行う．ある複数人歩行データの識別結果は，全ての単独歩行テンプレートとの間で DTW 距離を計算し，最も DTW 距離が小さかった歩行データの該当者を表す id である．

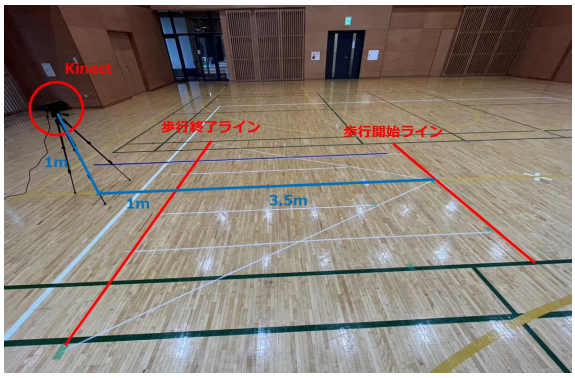

図 5 実験環境

表 3 実験参加者の情報

項目	環境
実験日	2022 年 7 月 16 日
実験時刻	9:30 から約 2 時間
場所	明治大学 中野キャンパス 多目的ホール
年齢	20 代前半
性別	男性 4 名 女性 3 名
人数	7 名

4. 実験

4.1 実験目的

Kinect と OpenPose の多人数同時個人識別の精度を比較すること．

実験 1 センシング評価

実験 2 DTW による個人識別

4.2 実験データ

Kinect によって得た映像と関節座標を保存するためのシステムは Processing を用いて開発した．Processing で Kinect for Windows v2 を扱うためのライブラリとして KinectPV2[5] を利用した．

単独歩行と複数人歩行をそれぞれ観測した．同時歩行人数を m とする．$m = 2 \sim 6$ 人が同時に歩行し，そのデータを得る実験である．実験参加者の詳細を表 3 に示す．実験環境は図 5 の通りである．

単独歩行では Kinect に対して直進する方向とそこから ±30◦ 傾いた方向の 3 パターンについて，1 人当たり 2 回ずつ測定した．複数人歩行では，人数 m を変えながら，全員が直進するパターンを 3 回，交差が発生するパターンを 3 回測定した．

4.3 実験方法

4.3.1 実験 1

Kinect で推定した関節座標データに対して Kinect のセンシング誤りによるデータの誤差がどの程度存在しているのかを調べる．歩行の向きや人数の変化による Kinect の推定

表 4　各歩行方向に対するセンシング状態

歩行方向 (Kinect 基準)	正常	一部異常	全体異常
正面	1.0 (21 / 21)	0.0 (0 / 21)	0.0 (0 / 21)
正面以外	0.48 (20 / 42)	0.33 (14 / 42)	0.19 (8 / 42)
合計	0.65 (41 / 63)	0.22 (14 / 63)	0.13 (8 / 63)

表 5　歩行人数に対するセンシング状態

歩行人数 m	正常	一部異常	全体異常
1	0.65 (41 / 63)	0.22 (14 / 63)	0.13 (8 / 63)
2	0.67 (8 / 12)	0.17 (2 / 12)	0.17 (2 / 12)
3	0.64 (9 / 14)	0.29 (4 / 14)	0.071 (1 / 14)
4	0.36 (5 / 14)	0.43 (6 / 14)	0.21 (3 / 14)
5	0.29 (4 / 14)	0.50 (7 / 14)	0.21 (3 / 14)
6	0.43 (6 / 14)	0.29 (4 / 14)	0.29 (4 / 14)

精度を評価する．

データからランダムにフレームを選んで可視化し，手動でラベル付けを行う．データの選び方は単独歩行と複数人歩行で異なる．

単独歩行については，実験参加者 $n = 7$ 人の歩行データについて，それぞれ 3 種類の歩行方向から 3 フレームずつランダムに選ぶ．可視化するフレームの合計は $7 \times 3 \times 3 = 63$ フレーム分である．

複数人歩行については，各歩行人数のデータに対して 1 人あたり 2 フレーム，合計 $7 \times 2 = 14$ フレームをランダムに選んで評価する．*

ラベル付けに関しては，正常に見えるもの，一部に異常が見られるもの，全体的に異常なものの 3 種類に分類する．各ラベルのサンプルを次の図 6, 7, 8 に示す．

4.3.2　実験 2

単独歩行・複数人歩行に対する Kinect と OpenPose の精度を調べる．

3.5 節の手法に基づき，各歩行データに対して個人識別を行う．また，$1 \leq k \leq 5$ の top と $topk$ の推定を含めた精度 top_kAcc を用いて，Kinect と OpenPose の間で精度を比較する．ただし，top_kAcc の計算においては，DTW 距離が最小のものから昇順で該当個人 id を並べたものを識別結果としている．ここで，top_kAcc の定義は以下の通りである．

$$top_kAcc = \frac{\text{予測の上位 k 個に本人が含まれた数}}{\text{全歩行データ数}}$$

4.4　実験結果

4.4.1　実験 1

歩行方向を正面と正面以外に分けたときの各ラベルの占める割合を表 4 に示す．複数人歩行と単独歩行を合わせて歩行人数に関してラベルの割合を表 5 に示す．

*ただし 2 人歩行については 1 人参加していない参加者が存在するため，複数人歩行の合計可視化フレーム数としては 68 フレームである．

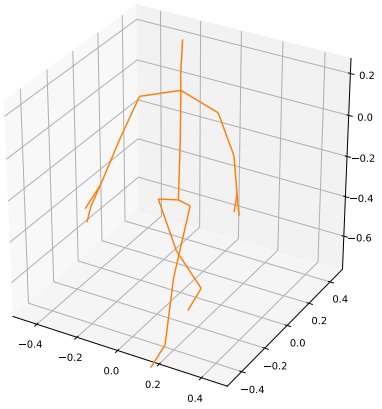

図 6　正常なフレーム

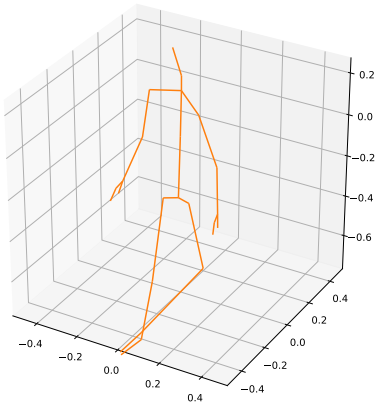

図 7　一部に異常が見られるフレーム

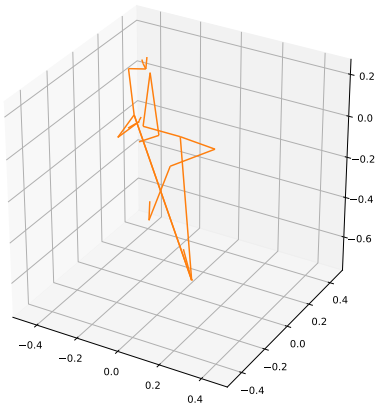

図 8　全体的に異常なフレーム

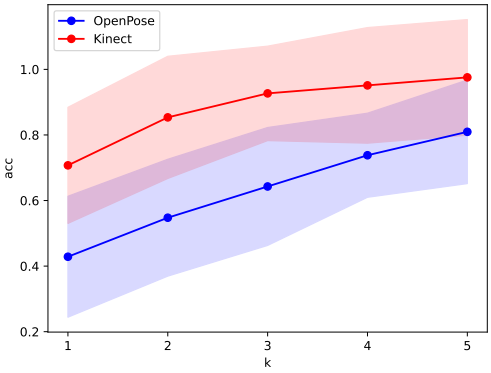

図 9　単独歩行精度

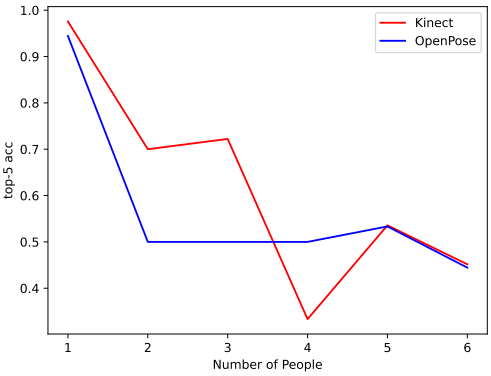

図 10　人数 m に対する精度 (top-5 acc)

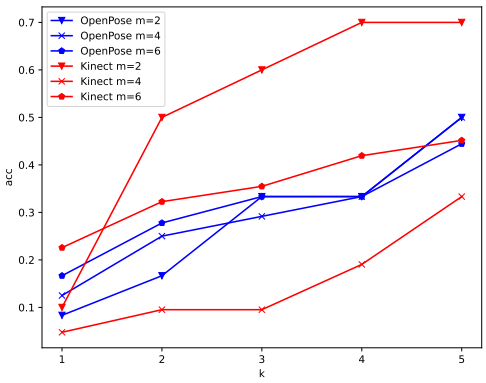

図 11　k に対する各人数 m における同時歩行精度

4.4.2　実験 2

単独歩行における top_kAcc $(k = 1, 2, 3, 4, 5)$ を図 9 に示す. また，歩行人数を増加させたときの識別精度を top_kAcc $(k = 1, 3, 5)$ を特に top_5Acc について図 10 に示す. Kinect と OpenPose の両ツールについて，k を変化させたときの各人数 m における同時歩行精度を図 11 に示す.

4.5　考察

4.5.1　Kinect のセンシング誤り

単独歩行について，Kinect の歩行角度に対するセンシング精度劣化した．Kinect は姿勢推定を行うとき，各画素に対して 25 関節のうちどの関節に該当するかを分類し，関節ごとに分類された画素の中心座標を求めることで姿勢推定を行っている．そのため，Kinect に対して角度が付くことで，Kinect から見えづらい関節が増えて姿勢推定は失敗しやすくなると考えられる．

また，歩行人数に対する Kinect のセンシング精度について，$m = 4, 5$ 人歩行のとき大きく 0.3 に半減していた. 従って，Kinect が問題なく性能を発揮できるのは 3 人同時姿勢推定までであると言える．また，6 人歩行のとき，4, 5 人歩行より精度が一時的に上がった．これは，そもそも 6 人を同時に捉えられず，歩行人数が少ない状態に等しくなっているからと考えられる．実際に，6 回の 6 人歩行のうち 2 回はそもそも 6 人分の歩容が得られていなかった.

4.5.2　個人識別

歩行人数が $m \leq 3$ 人のとき，k に関わらず OpenPose よりも Kinect の方が識別精度 top_kAcc が高かった．しかし，4 人歩行で精度が逆転した．6 人歩行に対しては Kinect の方がやや有利であったが有意な差ではない．このことは，Kinect のセンシング精度が 3 人までは性能が保たれ，4 人以降で精度が減少するというセンシング精度の考察と一貫している.

従って，6 人歩行について 6 人同時に捉えられていないことを考慮すると，表 6 のように 3 人以下の歩行は Kinect が有利であり，4 人以上の歩行は OpenPose がやや有利，7 人以上の歩行について Kinect の制約から OpenPose が有利である，と結論づける.

5.　おわりに

本実験により次を明らかにした. Kinect のセンシング精度は Kinect に対して直交歩行するとき高く，そうでないとき精度が低下する. また，3 人までは精度を落とさず姿勢推定が出来る．個人識別については，3 人までは Kinect が有利であり，4 人以上では OpenPose が有利である.

今後の課題として，歩行実験の参加者数を増やすことや，7 人以上の歩行に対する個人識別について調べることが挙げられる．Kinect は仕様として最大 6 人までの検出が可能としている．これは誤りでなく実際に 6 人を検知出来ていたが，一部のデータでは 6 人全員を捉えられていなかった．そのため，7 人以上の歩行に対しても全員を捉えられないことが考えられる．また，実際に多くの人が歩行している場においては一度に検出できる歩行者が多いほど良いため，Kinect よりも人数の増加に強いと考えられる OpenPose が相応しいと言える.

参考文献

[1]　森 駿文，菊池 浩明，“歩容データの DTW 距離に基づく個人識別手法の提案と外乱に対する評価”，情報処理学会論

表 6 歩行人数に対する 2 ツール間の優位性の比較

歩行人数	1, 2, 3	4, 5, 6
Kinect	有利	不利
OpenPose	不利	有利

文誌，Vol.60, No.9, 1538-1549, 2019.

[2] Ju Han, Bir Bhanu,“Individual recognition using gait energy image”, IEEE transactions on pattern analysis and machine intelligence, 28(2)，316-322，2005.

[3] Zhe，Gines，Tomas，Shih-En，Yaser,“OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”，CVPR，pp.7291-7299，2017.

[4] Julieta, Rayat, Javier, James,“A simple yet effective baseline for 3d human pose estimation”, ICCV，pp. 2640-2649, 2017.

[5] MicroSoft,“Kinect v2 library for Processing”(https://github.com/ThomasLengeling/KinectPV2)，2016.

[6] 渡邊宏,“「Kinect v2」はここがスゴい！新旧比較と Kinect による NUI 開発の最前線”，MONOist，2014.

[7] Sakoe, H. and Chiba, S,“Dynamic Programming Algorithm Optimization for Spoken Word Recognition, IEEE Transaction on Acoustics, Speech, and Signal Processing”,Vol.ASSP-26, No.1, pp.43–49，1978.

[8] Muaaz, M. and Mayrhofer, R.:“Smartphone-Based Gait Recognition: From Authentication to Imitation, IEEE Trans. Mobile Computing”, Vol.16, No.11, pp.3209–3221, 2017.

[9] 三好駿, 森駿文, 菊池浩明,“歩容データからの属性暴露リスクについて”，情報処理学会第 81 回全国大会， pp.3 421-3 422, 2019.

[10] 阪田 篤哉, 武村 紀子, 八木 康史,“多段階畳み込みニューラルネットワークを用いた歩容に基づく年齢推定”， 2018 年 5 月コンピュータビジョンとイメージメディア研究会，吹田，Vol. 2018-CVIM-212，No. 23，pp. 1-5，May 2018.

[11] 森　駿文, 菊池　浩明,“複数の歩容特徴量の k 近傍による「歩きスマホ」にロバストな個人識別手法の提案”, 暗号と情報セキュリティシンポジウム (SCIS 2019), pp. 1-7, 2019.

[12] Nicolai Wojke , Alex Bewley , Dietrich Paulus,“Simple online and realtime tracking with a deep association metric”，IEEE, ICIP, Pages.3645–3649, 2017.