# 超低解像度 FIR 画像内での人物位置と動作の違いに着目した骨格推定法の検討

○岩田 紗希 † 川西 康友 † 出口 大輔 † 井手 一郎 † 村瀬 洋 † 相澤 知禎 ‡

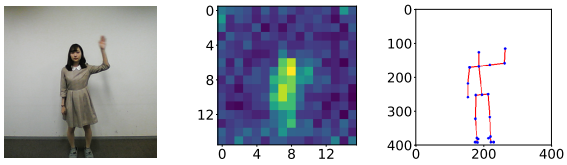
○ Saki IWATA † Yasutomo KAWANISHI † Daisuke DEGUCHI † Ichiro IDE †
Hiroshi MURASE † and Tomoyoshi AIZAWA ‡

†：名古屋大学， iwatas@murase.is.i.nagoya-u.ac.jp
{kawanishi, ide, deguchi, murase}@i.nagoya-u.ac.jp
‡：オムロン株式会社

＜要約＞近年，高齢化の進展により，独居高齢者に向けた見守りシステムが注目されている．しかし，見守りシステムにはプライバシーの問題が存在する．そこで我々はプライバシーの問題を軽減でき，さらに暗闇での撮影にも強い赤外線センサアレイを用いて撮影した超低解像度 FIR 画像系列から人物の骨格を推定する手法を提案してきた．本発表では特に超低解像度 FIR 画像内での人物位置と動作の違いに着目した骨格推定法について検討する．具体的には，超低解像度 FIR 画像内で人物位置に依存しない特徴を抽出するため，低解像度画像において影響が大きいと考えられる中間層での Pooling をせずに，特徴抽出の最後に GlobalMaxPooling を導入する．さらに動作の種類に合わせて，時系列情報と空間的情報を有効活用できるネットワークを提案する．実験では，赤外線センサアレイの画角内の様々な位置で人物を撮影したデータセットを用いて，従来手法と提案手法で特定の人物位置で学習を行ない，未学習位置で骨格推定精度を評価した．その結果，人物の動作を滑らかに推定でき，定量的にも精度が向上することを確認した．
＜キーワード＞ CNN，骨格推定，低解像度画像，FIR 画像，高齢者

## 1 はじめに

近年，日本では高齢化が問題となっている．2017 年 10 月 1 日時点での総人口に対して 65 歳以上が占める人口の割合は 27.7 %となった．さらに今後，少子化の影響もあり，高齢化率は上昇すると予測されており，2065 年には国民の約 2.6 人に 1 人が 65 歳以上となる社会が到来すると推定されている．その中でも 75 歳以上の人口の割合は 25.5 %を占め，約 3.9 人に 1 人が 75 歳以上になると推計されており，1 人暮らしをする高齢者が増加すると考えられている [1]．高齢者の健康で安全な暮らしのためには，身体機能の維持と緊急時への対応が必要であり，独居高齢者を対象とした見守りシステムが注目されている．見守りシステムには人感センサ型や緊急通報型，カメラ型などが存在するが，屋内に可視光カメラを設置して撮影した画像を用いて，人物の行動を認識することが一般的である．一方で高解像度で日常生活の様子を撮影することには，プライバシー上の懸念がある．




(i) 可視光画像 (ii) FIR 画像 (iii) 関節点位置

**図 1 赤外線センサアレイによる撮影例と関節点位置**

そこで我々は，特定の領域内の温度分布を計測できる赤外線センサアレイに着目している．赤外線センサアレイは複数の赤外線センサを格子状に集約したもので，特定の領域内の温度分布を計測することができる．非常に安価であることから，人感センサとして空調などの家電用品にも用いられている．赤外線センサを用いて撮影した画像（図 1 (ii)）は低解像度であり，プラ

DIA2020 動的画像処理実利用化ワークショップ(2020. 3. 9-10 沖縄)

イバシー上の懸念を軽減できる．また，暗闇でも熱源を感知できるという利点もある．以上の理由から我々は赤外線センサを用いて撮影した画像から図 1 (iii) に示すような人物の骨格を推定する手法を提案している [2]．しかしこの手法では，人物は常に FIR 画像中の同じ位置にいるという前提があるため，センサの画角内で人物の位置が変わると骨格を精度良く推定できない．また，最大でも隣接 2 フレームのみから推定するため，動作の違いをうまく扱えない．

本発表ではこの 2 つの問題点に対処するために

- 畳み込みネットワークの中間で Pooling をせず，最後に GloabalMaxPooling を用いる超低解像度 FIR の画像内での人物位置の違いに頑健な特徴抽出
- 行動の違いに対し，空間的情報と時系列情報を選択的に活用するネットワーク

の 2 つのアプローチによって，人物位置によらない特徴量を抽出し，また動作の違いに合わせて時系列情報を有効活用できる骨格推定法を検討する．

## 2 関連研究

人物の骨格を推定する研究として，高解像度可視光画像を用いる手法が提案されている．また赤外線センサを用いた研究として，手振り動作の認識や行動認識が提案されている．

### 2.1 人物骨格推定に関する研究

人物の骨格推定のモデルには画像中のキーポイントを抽出し，人物ごとにつなぎあわるボトムアップ型と，人物を検知した後にそれぞれの人物において骨格推定を行なうトップダウン型がある．ボトムアップ型では計算量を削減できるが，画像全体の情報を十分に考慮できていないため，関節点の接続の精度が低い傾向がある．一方，トップダウン型は人数に比例して計算量も増加してしまうが，各人物に対して骨格推定を行なうので，高精度に骨格推定ができる．

トップダウン型の手法として，Chen らの Cascaded Pyramid Nework（CPN） [3] がある．CPN は明確なキーポイントを抽出する GlobalNet と，GlobalNet で生成した特徴量をアップサンプリングして統合することで見つけにくいキーポイントの推定を可能にする RefineNet という 2 つのネットワークから構成される．

一方，ボトムアップ型の手法として，Cao らは OpenPose[4] を提案している．OpenPose では入力画像に対し，関節点の座標を示すヒートマップである Part Confidence Map（PCM）と関節点間のつながりを示すベクトル場である Part Affinity Fields（PAF）を計算する．その後，推定した関節点の座標とそのつながりから，人物の骨格を推定する．

上記で紹介した手法ではいずれも情報が多い可視光画像を対象としており，多人数の複雑な骨格に対し，高精度な推定を実現している．しかし，これを超低解像度 FIR の画像に直接適用し，骨格推定をすることは難しい．

### 2.2 赤外線センサアレイを用いた研究

鳥山らは人物の正面や頭上に設置した赤外線センサを用いて，手振り動作を認識する手法 [5] を提案している．低解像度かつノイズが多いという赤外線センサの特性から生じる問題を軽減するため，人体温度に注目した温度の絞り込みと，手振りの動作領域のみを切り出す空間的な絞り込みを合わせ，手振り動作認識精度を向上させている．

川島らはこのセンサを用いて，FIR 画像から特徴抽出を行ない，日常行動（歩行，着席，起立）と異常行動（転倒）の 4 種類のクラスからなる人物行動認識手法 [6] を提案している．この手法では FIR 画像から人物領域のみを切り出し，フレーム間差分画像を求める．そして畳み込みニューラルネットワーク（CNN）による視覚特徴と再帰型ニューラルネットワーク（RNN）による時系列情報を学習をすることで高精度な行動クラス分類を可能にしている．これにより，独居高齢者に対しての緊急時への対応が実現できる．

これらの研究をふまえ，我々はこれまでに図 1 (ii) のような超低解像度 FIR 画像から図 1 (iii) に示すような人物の骨格を推定する手法を提案している [2]．この手法は，FIR 画像を入力として関節点の座標を直接回帰することで，人物の骨格位置を 2 次元座標で出力する．また赤外線センサアレイと可視光カメラを同期させて学習データを撮影し，可視光画像に対して OpenPose を適用した推定結果から自動的に教師信号を取得して学習する手法も提案している．しかし，超低解像度 FIR 画像内における 1 画素のずれが大きいため，人物の位置が変わってしまうと推定精度が下がり，この手法では骨格を正しく推定できない．

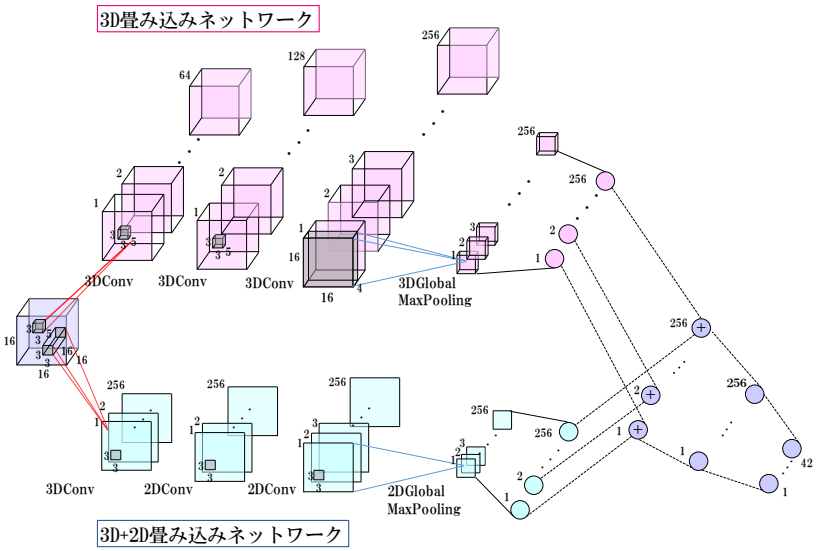



図 2　提案手法のネットワーク構造

## 3　人物位置と動作の違いに着目した骨格推定法

本研究では骨格推定のために $N$ フレームの FIR 画像系列 $X_i = (\mathbf{x}_{(i-(N-1))}, ...\mathbf{x}_{(i-1)}, \mathbf{x}_i) \in \mathbb{R}^{N\times16\times16}$ を入力とし，その最終フレームにおける骨格として 21 個の関節点の 2 次元座標を並べた 42 次元ベクトル $\mathbf{y}_i \in \mathbb{R}^{42}$ を出力とする CNN を設計する．提案手法のネットワークでは，扱う FIR 画像が低解像度であり，さらに位置に依存しない特徴を抽出するために，畳み込み層毎に Pooling はせず，特徴抽出の最後に GlobalMaxPooling を用いる．

一方で，人物の動作に着目すると，突発的な動作と連続した動作が存在する．各動作中の人物の骨格を推定する場合，突発的な動作に対しては現在フレームの空間的情報が有効であるのに対し，連続した動作に対しては時系列情報が有効であると考えられる．そこで，動作の種類に応じて時系列情報と空間的情報を選択的に利用するネットワークを提案する．提案手法は次の 2 つのネットワークを並列に用いて特徴抽出を行なう．3D 畳み込みネットワーク（図 2 上段）は時系列情報と空間的情報に着目したネットワーク，3D+2D 畳み込みネットワーク（図 2 下段）は空間的情報に注目したネットワークである．最後に 2 つのネットワークから抽出した特徴を結合して，骨格推定に用いる．図 3 に提案手法の処理手順を示す．

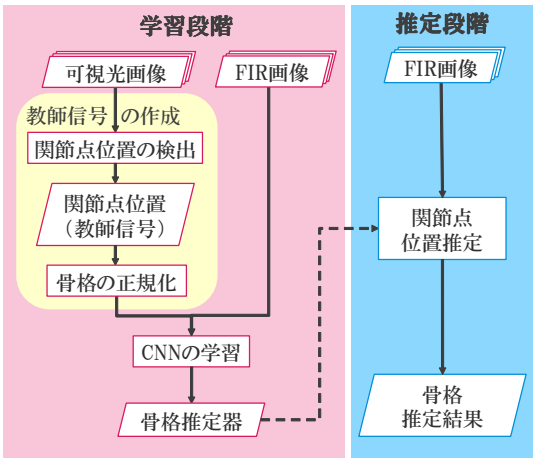



図 3　提案手法の処理手順

### 3.1　3D 畳み込みネットワーク

FIR 画像を $N$ 枚入力し，枚数 5 ×高さ 3 ×幅 3 のカーネルサイズのフィルタを用いて，移動幅 1 で 3D 畳み込みを行ない，64, 128, 256 の 3 段階のチャネル数で特徴を抽出する．時間方向に 5 枚の画像を畳み込むことで，動きの情報を学習し，時系列情報を活用することを目的とする．最後に枚数 4 ×高さ 16 ×幅 16 のカーネルサイズのフィルタで GlobalMaxPooling を行ない，人物位置によらない情報を抽出し，全結合層により，256 次元のベクトル $\boldsymbol{a}_i = f_{3D}(X_i)$ を得る．

### 3.2　3D+2D 畳み込みネットワーク

FIR 画像を $N$ 枚入力し，まず枚数 16 ×高さ 3 ×幅 3 のカーネルサイズのフィルタを用いて，移動幅 1 で

3D 畳み込みを行ない，時系列情報を削減した 64 チャネルの特徴を抽出する．その後，高さ 3 ×幅 3 のカーネルサイズのフィルタで 2D 畳み込みを行ない，3D 畳み込みネットワーク同様に 128，256 チャネルの特徴を抽出する．最後に高さ 16 ×幅 16 のカーネルサイズのフィルタで GlobalMaxPooling を行ない，人物位置によらない情報を抽出し，全結合層により．256 次元のベクトル $\boldsymbol{b}_i = f_{\mathrm{2D+3D}}(X_i)$ を得る．

### 3.3　骨格の回帰

3D 畳み込みネットワークと 3D+2D 畳み込みネットワークから取得したそれぞれのベクトル $\boldsymbol{a}_i, \boldsymbol{b}_i$ を，$\alpha : 1-\alpha$ の比で重みをつけて足し合わせ，256 次元のベクトル $\boldsymbol{c}_i = \alpha \boldsymbol{a}_i + (1-\alpha)\boldsymbol{b}_i$ を得る．ただし，$\alpha \in [0, 1]$ は学習により決定する．この $\boldsymbol{c}_i$ から，ネットワーク $f_p$ により，骨格推定結果 $d_i = f_p(\boldsymbol{c}_i)$ を得る．

### 3.4　ネットワークの学習

損失関数には関節点の教師信号と推定関節点位置の平均 2 乗誤差（MSE）を用いる．またネットワークの重み $\alpha$ を 0 又は 1 に近づくような制約を加える．従って，以下の損失関数を最小化するようにネットワークを学習する．

$$L = \sum_i ||d_i - \mathbf{y}_i||^2 + \lambda \| \frac{1}{2}\alpha(\alpha - 1) \| \qquad (1)$$

ここで，$\lambda = 0.01$ とした．この $\alpha$ に関する制約項により，学習したデータに含まれる FIR 画像内における動作の違いに適応した学習ができることが期待される．

## 4　実験

FIR 画像内の人物位置によらない特徴を抽出し，行動によって時空間情報を選択的に利用する骨格推定法の有効性を示すための実験を行なった．

データセットとして，まず赤外線センサアレイ（オムロン製 D6T-1616L）と可視光カメラ（バッファロー製 BSW20KM11BK）を用い，従来手法 [2] に倣って図 4 のように同期撮影を行ない，撮影中の被撮影者の行動が異なる 2 種類のデータセットを構築した．行動の種類は, 着座状態から 1 度立って座り直す（行動 A）と左右の手を交互に上げる（行動 B）とした．各行動の撮影対象は 1 人で，センサに対する人物の位置は図 5 に示す 9 箇所とした．1 箇所での 1 回の撮影あたりのフレーム数は行動 A において約 220 フレーム，行動 B において約 440 フレームとした．

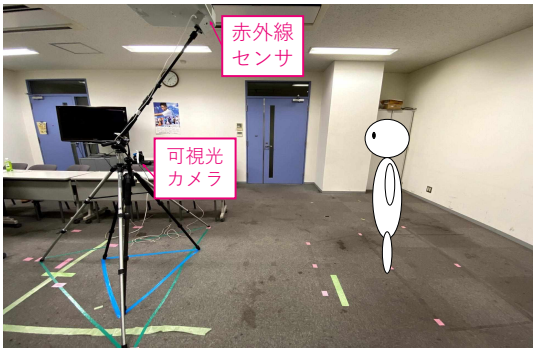
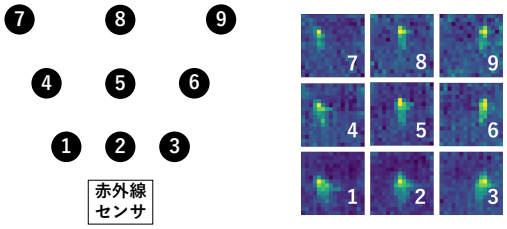



図 4　撮影風景




(i) 撮影位置　(ii) FIR 画像の例

図 5　赤外線センサアレイと被撮影者の位置関係

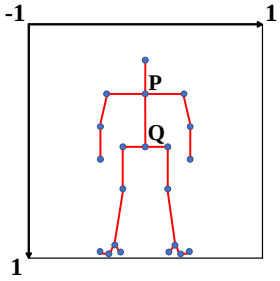
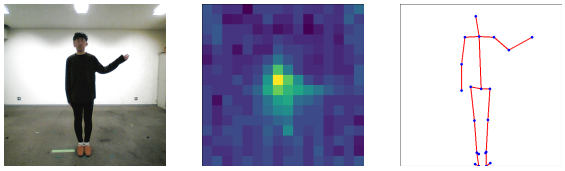



図 6　骨格空間の例



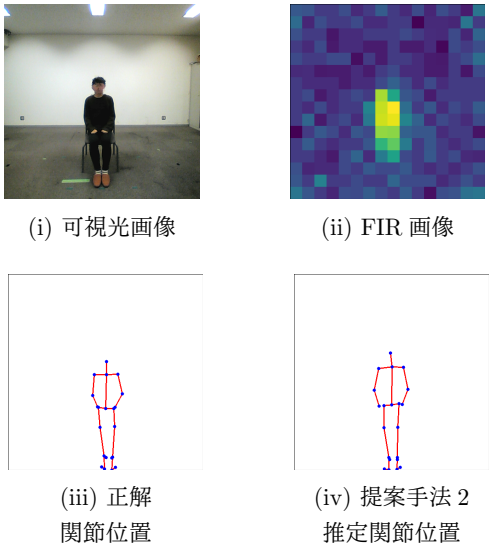
(i) 可視光画像　(ii) FIR 画像　(iii) 関節点位置

図 7　撮影データ例と関節点位置

撮影した可視光画像に対して OpenPose を用いて骨格の教師信号を取得した．その関節点位置を，図 6 に示す $[-1, 1]$ の大きさの骨格空間に対し，関節点 P，Q 間の長さが行動 A では 0.625，行動 B では 0.333 とな

表 1　真値と推定結果の RMSE （$\times 10^{-2}$）

| 位置 | | 2 | 4 | 5 | 6 | 8 | 平均 |
|---|---|---|---|---|---|---|---|
| 行動 A | 手法 [2] | 3.75 | 5.32 | 3.74 | 5.01 | 8.05 | 5.17 |
| | 提案手法 1 | **3.30** | 4.51 | 3.55 | 4.01 | **3.96** | 3.87 |
| | 提案手法 2 | 3.67 | **3.51** | **3.42** | **3.61** | 4.31 | **3.70** |
| 行動 B | 手法 [2] | 4.91 | 6.78 | 3.39 | 7.45 | 4.96 | 5.50 |
| | 提案手法 1 | **3.07** | 5.58 | **3.17** | 5.94 | 4.57 | 4.47 |
| | 提案手法 2 | 3.85 | **4.74** | 3.63 | **5.09** | **4.29** | **4.32** |



(i) 可視光画像　(ii) FIR 画像

(iii) 正解関節位置　(iv) 提案手法 2 推定関節位置

図 8　実験結果例

るように正規化した．また図 6 に示す骨格空間の左上の座標を（$-1, -1$）としたときに，関節点 Q の $x$ 座標が空間の中心（$x = 0$）となるように，また $y$ 座標の最大値が $y = 1$ となるように平行移動して位置を揃えた．撮影した可視光画像，FIR 画像，それらに対応する教師信号を図 7 に示す．

実験では以下に示す 3 種類の手法をを比較した．提案手法 1 は，手法 [2] に対し畳込み層毎に Pooling をせず，特徴抽出の最後に GlobalMaxPooling を用いるネットワーク，提案手法 2 では図 2 のネットワークを用いた．比較手法は，手法 [2] とした．提案手法 1 と比較手法では 1 フレームの FIR の画像を，提案手法では 16 フレームの FIR 画像を入力した．

各行動の画像系列において，学習していない人物位置に対しての推定精度を比較するため，学習には図 5 中の人物位置 1, 3, 5, 7, 9 のデータを用い，評価には学習に含まれていない人物位置 2, 4, 6, 8 のデータと，位置 5 だが学習には用いていないデータを用いた．

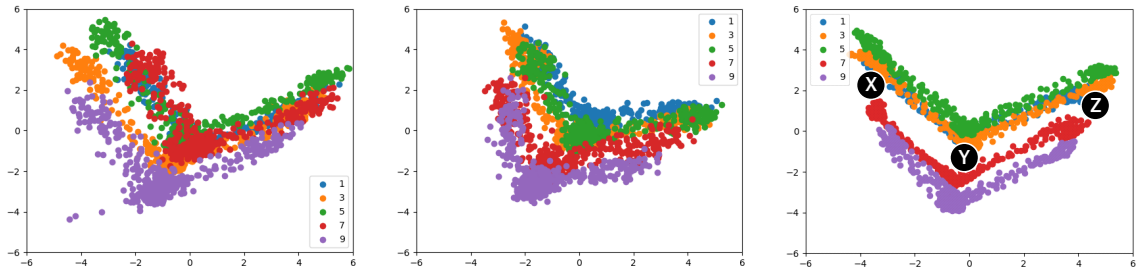
人物位置 3 において，比較手法と提案手法 2 により行動 A の時の骨格を推定した結果を図 8 に，各行動における真値と推定結果の平方根平均 2 乗誤差（RMSE）を表 1 に示す．両方の行動において平均的に提案手法の方が高精度であった．手法 [2] では，提案手法と比較して定量的にも定性的にも大きく精度が低かった．これは，各畳込み層の後に Pooling 層を挿入することで，様々な位置のデータに対して，有用な情報を抽出できず，精度良く骨格推定をできなかったためと考えられる．一方，表 1 より，提案手法 1 及び提案手法 2 では人物位置の違いに頑健になり，さらに提案手法 2 では時系列情報を動作に合わせて有効活用できた．ただし本実験ではデータセット中の撮影対象は 1 人の人物であったため，今後はデータセットを被験者人数，行動種類において拡張し，精度を確かめる必要がある．

## 5　考察

### 5.1　特徴抽出について

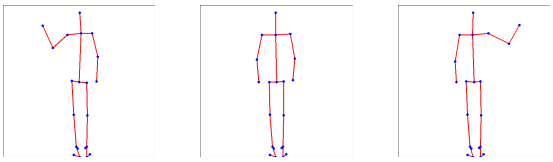
一般的な高解像度画像では，ネットワークの中間層で Pooling を行なうことで，位置ずれに頑健になり，計算量を削減できるという利点がある．しかし，本研究で扱っている超低解像度 FIR 画像では 1 画素が画像全体に対して非常に大きな割合を占めることから，手法 [2] では人物の位置ずれに対してうまく推定ができなかった．

図 9 にそれぞれの手法で，行動 A の推定結果について，学習した人物位置 1, 3, 5, 7, 9 における FIR 画像を入力した時の中間層の出力の特徴量を主成分分析（PCA）を用いて可視化したものを示す．ここではそれぞれの手法において，最終層の直前の層の出力値を用いる．手法 [2]（図 9(i)）と提案手法 1（図 9(ii)）を比較して，異なる人物位置でのデータの特徴間距離がより近くなったことが分かる．また，提案手法 2 ではそれぞれの位置ごとの分布の幅が狭くなり，より類似した形状になった．さらに図 9(iii) に示す X，Y，Z 付近にあるフレームの骨格推定の例を図 10 に示す．これにより，GlobalMaxPooling と行動に適したネットワークを用いることで超低解像度である FIR 画像から人物の位置の違いに頑健な特徴を得られたことが分かった．




(i) 手法 [2]　(ii) 提案手法 1　(iii) 提案手法 2

**図 9　PCA による特徴量の可視化**



(i) X　(ii) Y　(iii) Z

**図 10　PCA による各特徴点の骨格例**

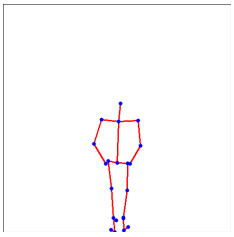
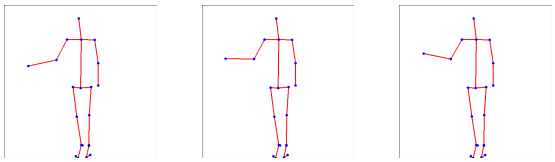


**図 11　行動 A の動作例**

**表 2　関節点ごとの真値と推定結果の RMSE （$\times 10^{-2}$）**

| 関節点 | 頭 | 手 |
|---|---|---|
| RMSE | 3.64 | 3.90 |



(i) $t$　(ii) $t+1$　(iii) $t+2$

**図 12　行動 B における連続するフレーム系列例**

### 5.2　行動の違いに着目したネットワークの有効性について

提案手法 2 では行動の違いに着目したネットワークを用い，骨格推定を行なった．その結果，最後の結合において，行動 A では 3D+2D 畳み込みネットワークの方に，行動 B では 3D 畳み込みネットワークの方の特徴を選択するように重み $\alpha$ の学習が進んだ．行動 A では図 11 のような立つ瞬間，座る瞬間など全体のフレーム数に対して，その動作が占める割合が非常に小さい突発的な動きが存在する．これに対し，現在のフレームでの空間的特徴を優先するため，3D+2D 畳み込みネットワークにより，特徴抽出の初期段階で時系列の情報を削減したことが有効だったと考えられる．一方，行動 B では図 12 のように全体的に連続的な動作を行なっていたため，時系列方向の 5 フレームの情報を使いながら特徴抽出することで，提案手法 1 と比較して，定性的にも滑らかな推定ができた．本実験で用いたデータセットは 2 種類の行動しか含まないため，骨格時系列変化に対して上記の仮説を立て実験を行なったが，今後より多様な行動に対して検証する必要がある．

### 5.3　関節点位置ごとの精度差について

行動 B で図 5 の人物位置 2, 4, 6, 8 において，提案手法 2 を用いた関節点位置ごとの真値と推定結果の RMSE の平均を表 2 に示す．ここで，頭は行動においてほとんど動かない部位，手は大きく動く部位を表しており，手は左右の平均値とする．定量的に見ると頭の誤差は小さく，動いている手の誤差が大きくなっているが，誤差の差は大きくはない．今後，行動において瞬発的に大きく動く部位をそうでない部位に分けて，重みを付けて学習をすることで関節点ごとに適した学習ができると考えられる．

## 6 むすび

本研究では，超低解像度 FIR 画像内での人物位置と動作の違いに着目した人物骨格推定手法について検討した．評価実験により，GlobalMaxPooling を用いることで人物の位置によらない特徴が得られることが確認できた．さらに行動の違いに着目することで，時系列情報を活用して平均的に精度が向上することを確認した．

今後の課題として，撮影対象者数や行動種類を増やした実験や，動作に応じて適切な特徴を抽出できるネットワークの提案，より正確に骨格推定精度を評価できる指標の検討などが考えられる．

## 謝辞

本研究の一部は，科学研究費補助金（17H 00745）による．

## 参考文献

[1] 内閣府， “令和元年度高齢社会白書，” https://www8.cao.go.jp/kourei/whitepaper/w-2019/zenbun/pdf/1s1s_01.pdf. （2020/1/8 参照）

[2] 岩田 紗希，川西 康友，出口 大輔，井手 一郎，村瀬 洋， 相澤 知禎 : “超低解像度遠赤外線画像からの人物骨格推定の検討”，第 25 回画像センシングシンポジウム，IS2-26， 6 月 2019.

[3] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu and J. Sun : “Cascaded pyramid network for multi-person pose estimation,” Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition, pp.7103–7112, June 2018.

[4] Z. Cao, T. Simon, S. Wei and Y. Sheikh : “Real-time multi-person 2D pose estimation using part affinity fields,” Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition, pp.7291–7299, July 2017.

[5] 鳥山 千智，川西 康友，出口 大輔，井手 一郎，村瀬 洋，相澤 知禎，川出 雅人 : “赤外線センサアレイを用いた温度と空間の絞り込みによる手振り動作認識に関する検討”，電子情報通信学会技術研究報告，PRMU2014-87， 1 月 2015.

[6] 川島 昂之，川西 康友，出口 大輔，井手 一郎，村瀬 洋，相澤 知禎，川出 雅人 : “赤外線センサアレイを用いた畳み込み RNN による人物行動認識”，精密工学会誌， vol.84，no.12，pp.1025–1032，Dec 2018.