論文

超低解像度 FIR 画像系列中での人物位置と行動の違いに着目した骨格推定法 *

岩田紗希 ** 川西康友 ** 出口大輔 ** 井手一郎 ** 村瀬洋 ** 相澤知禎 ***

Human Skeleton Estimation Method Focusing on the Difference of Human Position and Action from an Extremely Low-Resolution FIR Image Sequence

Saki IWATA, Yasutomo KAWANISHI, Daisuke DEGUCHI, Ichiro IDE, Hiroshi MURASE and Tomoyoshi AIZAWA

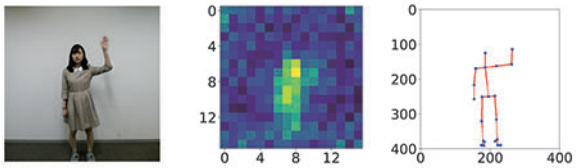
Elderly monitoring systems are gaining attention in the modern aging society. For the purpose, Far-InfraRed (FIR) sensors are often used, because they can avoid privacy concerns and are robust to environmental lightings. The authors have previously proposed several methods for human skeleton estimation from an extremely low-resolution FIR image sequence whose resolution is 16 × 16 pixels. For more accurate estimation, this paper proposes a method that is robust to variations of human positions and actions in the FIR sequences. Specifically, to extract features robust to the human positions from the images by using a Convolutional Neural Network (CNN), a global max-pooling layer is inserted into the last layer instead of multiple pooling layers which are not suitable for low-resolution inputs. Also, a network with two branches is introduced that focuses on capturing spatial and temporal information respectively. Moreover, the network has a weighted sum mechanism of their outputs, which depends on the human actions. For evaluation, a dataset was created by capturing action sequences of a human at various positions in the FIR images. Through an experiment, we confirmed that the human motion can be smoothly estimated and that the estimation accuracy is improved by the proposed method.

Key words: CNN, human skeleton estimation, low-resolution image, FIR image, elderly monitoring

1. 緒　　言

近年，日本では高齢化社会が問題となっている．内閣府によれば，2018 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 28.1 %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，人物の行動を認識するカメラ型が一般的である．しかし，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．このような見守りシステムを高齢者の身体機能の低下の評価へと発展させることを考えた場合，より細かな関節点の情報が得られる骨格推定の技術が必要となる．

そこで我々は，図 1 (i) のようなシーンにおいて，図 1 (ii) のような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 [2]．赤外線（Far-InfraRed ; FIR）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)）は非常に低解像度であり，このような「超低解像度」の画像を用いることで，プライバシー上の懸念を軽減できる [3) 4)]．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高精度に推定できない．また，1 フレームまたは隣接 2 フレームのみから推定するため，行動の違いをうまく扱えない．



(i) RGB image　(ii) FIR image　(iii) Joint points

Fig. 1 Example of a human in FIR image and its joint points

本論文ではこれら 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 global max-pooling 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

*　原稿受付　令和 2 年 5 月 21 日
　　掲載決定　令和 2 年 9 月 15 日
**　名古屋大学　（愛知県名古屋市千種区不老町）
***　正 会 員　オムロン（株）
（京都府木津川市木津川台 9-1）
（現，オムロンソーシアルソリューションズ（株）
滋賀県野洲市市三宅 686-1）

用するネットワーク

の 2 つの工夫によって，人物位置によらない特徴量を抽出し，また行動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．なお，本論文は我々の先行研究 [5] における論点を整理し，関連研究及び詳細な実験を追加して論文としてまとめたものである．

以降，2 章では関連研究を紹介する．3 章では，FIR 画像からの人物骨格推定法について提案する．4 章では，提案手法の有効性を確認するために行なった実験について報告し，5 章ではその考察を述べる．最後に，6 章で本論文をまとめる．

2. 関 連 研 究

人物の骨格を推定する研究は，高解像度可視光画像を用いる手法が一般的であり，多くの手法が提案されている．一方，本研究が対象とする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．本章では，これら 2 つの観点で関連研究をまとめる．

2.1 人物骨格推定に関する研究

人物の骨格推定モデルには，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と，画像中のキーポイントを抽出し，人物ごとにつなぎ合わせるボトムアップ型がある．

トップダウン型の手法として，Chen らの Cascaded Pyramid Nework（CPN）[6] 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．

Xiao らによって提案された Simple baseline method [7] もトップダウン型の 1 つであり，近年複雑化してきている骨格推定モデルの精度比較を簡単化した．このネットワークは ResNet に逆畳み込み層を結合した単純な構造からなり，各解像度の情報を保持するスキップ接続を有する骨格推定モデル [8] と比較しても，高精度な骨格推定を実現している．

一方，ボトムアップ型の手法として，Cao らは OpenPose [9] 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．

また他に Papandreou らの PersonLab [10] が挙げられる．PersonLab ではヒートマップ, Short-range offsets, Mid-range offsets の 3 つを推定する．ここでヒートマップは関節点を中心とした半径が一定の円とそれ以外の 2 クラス分類を，Short-range offset はヒートマップ内での関節点の座標の回帰を，Mid-range offset では各関節点のつながりを学習しておくことで得られる推定器により推定する．Personlab は姿勢推定に加え，インスタンスセグメンテーションも同時に行なうマルチタスク学習であるが，他のボトムアップ手法である OpenPose と比較してもより高精度な骨格推定を達成している．

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，これらの手法は入力画像上で推定した骨格のヒートマップに基づくため，それらを超低解像度 FIR の画像に直接適用し，骨格推定をすることは難しい．

また深度センサを用いた骨格推定も存在するが，見守りシステムの実現として各部屋にセンサを設置することを考慮した場合，センサ単体のコストが高く，実現が難しいと考えれる．

2.2 赤外線センサアレイを用いた人物の行動に関する研究

本研究で使用する赤外線センサアレイを用いた研究として，特定の行動検知や行動クラス分類が存在する．ここでは，この分野における既存研究をまとめる．

岡田らは頭上に設置した 8 × 8 画素の赤外線センサを用いて，FIR 画像内の人数と 2 種類の行動を検出する手法 [11] を提案している．この手法では，撮影した FIR 画像に対し，サポートベクトルマシン（Supprt Vector Machine ; SVM）による機械学習で分類を行ない，いずれにおいても，高精度な検出を実現した．

鳥山らは人物の正面や頭上に設置した本論文と同じ赤外線センサを用いて，手振り動作を認識する手法 [12] 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．

Fujita ら [13] や川島ら [14] は頭上に設置した赤外線センサを用いて，日常行動と異常行動の人物行動認識手法を提案している．

このうち Fujita らは 8 × 8 画素の赤外線センサを用いて，日常行動（起立，着席，横たわり，ベッドで寝る）と異常行動（転倒）の 5 種類のクラスからなる行動認識手法 [13] を提案している．格子状の画素で表現された FIR 画像をグラデーションで表現し，それをファインチューニングしたネットワークで学習を行なうことで，高精度な行動クラス分類を実現した．

一方，川島らは本論文と同じ赤外線センサを用いて，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行動認識手法 [14] 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（Convolutional Neural Network ; CNN）による視覚特徴と再帰型ニューラルネットワーク（Recurrent Neural Network ; 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．

上記の手法は，独居高齢者に対するプライバシーに配慮した見守りシステムにおいて，緊急時対応の実現に貢献している．

これらの研究をふまえ，我々はこれまでに FIR 画像からの行動認識を用いた上で，より詳細な情報を推定するため，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きた [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．そこで，本論文では人物位置に頑健な特徴を抽出する，赤外線センサアレイを用いた骨格推定についてまとめる．

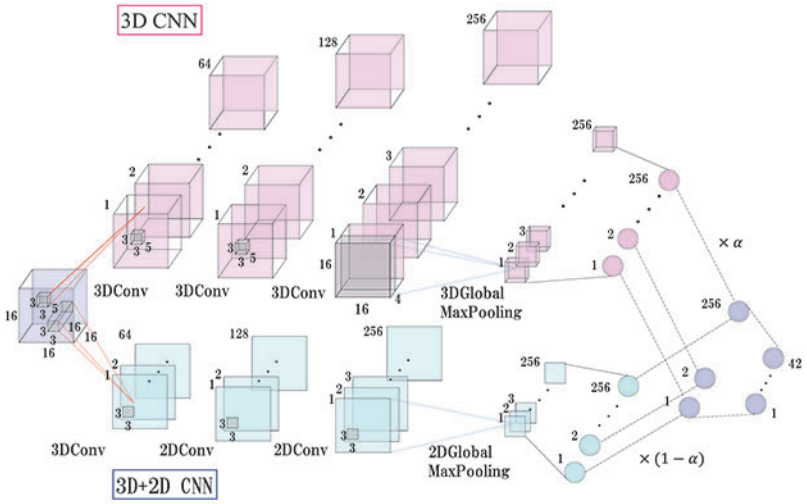


Fig. 2 Network structure of the proposed model

3. 人物位置と行動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 フレームからなる 16 画素四方の FIR 画像系列 $X_i = (\mathbf{x}_{(i-(N-1))}, \cdots, \mathbf{x}_{(i-1)}, \mathbf{x}_i) \in \mathbb{R}^{N\times16\times16}$ を入力とする．その最終フレームにおける骨格として入力画像とは異なる骨格を表現するための座標系（骨格空間）で表現した骨格 $\mathbf{y}_i \in \mathbb{R}^{2J}$ を出力する．ただし J は骨格を構成する関節点の個数である．これによって，FIR 画像から関節点などの細かな特徴抽出をするのではなく，FIR 画像の見えと画像中に映っている異なる空間で表された人物の骨格を対として学習を行なうことで，超低解像度 FIR 画像からの人物骨格推定を実現する．この骨格は，21 個の関節点の 2 次元座標を並べた 42 次元ベクトル $\mathbf{y}_i \in \mathbb{R}^{42}$ として表現する．提案手法のニューラルネットワークでは，扱う FIR 画像が低解像度であり，複数回の pooling 処理を適用すると，特徴マップにおいて人体部位間の位置関係が失われてしまうため，畳み込み層ごとの pooling はしない．一方で，画像内での人物位置に関する情報は削除して，特徴抽出の最後に global max-pooling を適用する．

人物の行動に着目すると，複数種類の静止状態（静止状態：起立状態など，全ての関節点の位置がそれぞれ人体の大きさに比べ十分小さい範囲内に収まっている状態）から構成される行動と，連続的な動き状態（関節点が連続的に移動している状態）から構成される行動がある．本論文では前者を行動クラス A，後者を行動クラス B と定義する．例えば，行動クラス A では椅子での起立・着席のような 2 つの静止状態が大部分を占める行動，行動クラス B では手をゆっくりと振り続けるなど連続的な動きが大部分を占める行動などが挙げられる．ここで，行動 A では起立状態と着席状態の 2 種類の静止状態が存在し，行動 B では静止状態は存在せず，手が連続的に動く行動となっている．各行動中の人物の骨格を推定する場合，行動クラス A では，各静止状態での骨格の学習が十分できれば，時系列情報を重視しないほうが正しく推定できると考えられる．一方で，行動クラス B では動きを捉えるために，時系列情報に注目し，複数フレームを用いて推定することが有効であると考えられる．そこで，学習データに含まれる行動に応じて、時系列情報と空間的情報の重みが学習時に決定されるネットワークを提案する．なお，本論文では行動認識 [14] 等により，観測対象の人物がとる行動は与えられているものとし，学習は行動ごとに行ない，行動に合わせて自動的にネットワークの重みを選択する．

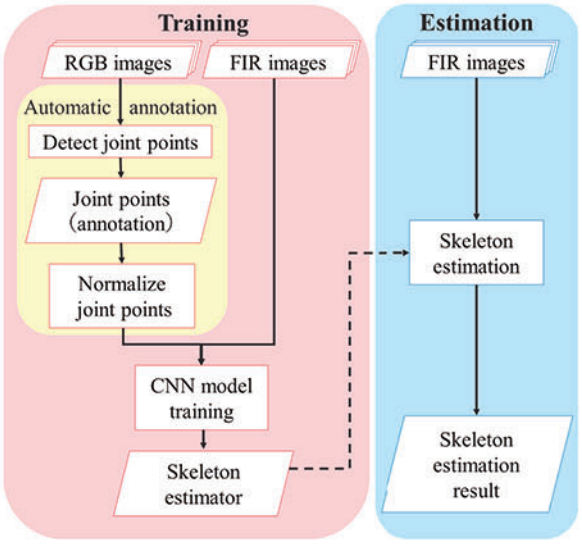


Fig. 3 Process flow of the proposed method

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を重視したネットワーク，3D+2D 畳み込みネットワーク（図 2 下段）は空間的情報を重視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行動ごとに学習した比率で混合して，骨格推定に用いる．図 3 に提案手法の処理手順を示す．なお，ここで示す骨格は 21 個の関節点から構成される．

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．ここでは，骨格が滑らかに動き，3 回の畳み込みの後の RespectiveField

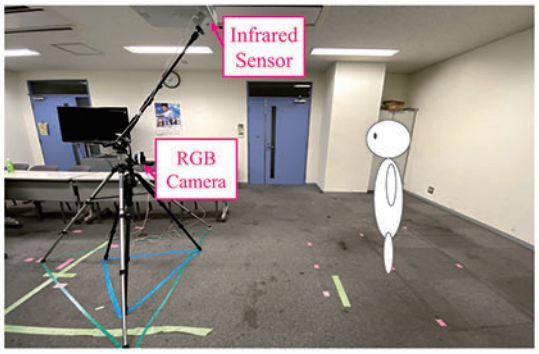
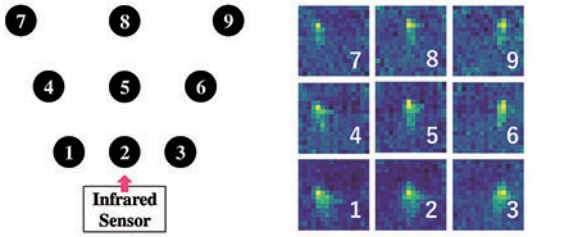


Fig. 4 Experiment environment



(i) Position of the sensor and the human position (ii) Example of the FIR image at each human position

Fig. 5 Positional relationship between the FIR sensor array and a human

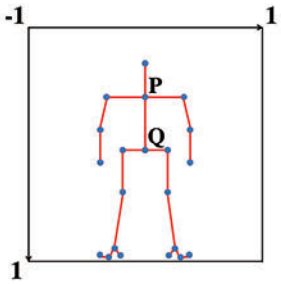
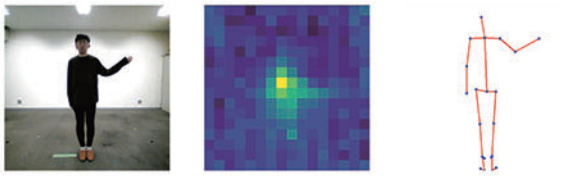
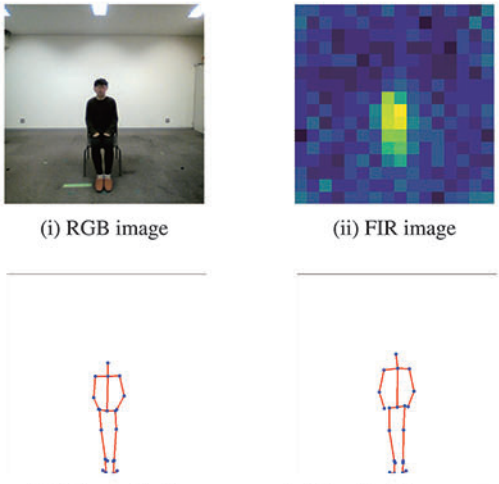


Fig. 6 Example of the skeleton space



(i) RGB image (ii) FIR image (iii) Ground truth in the skeleton space

Fig. 7 Example of corresponding images and joint point positions



(i) RGB image (ii) FIR image (iii) Ground truth (iv) Result of the proposed method

Fig. 8 Example of the skeleton estimation (Action A)

が最大限となるよう，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．最後に残った枚数 4× 高さ 16× 幅 16 の特徴マップに global max-pooling を適用し，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 N× 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 global max-pooling 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 $\boldsymbol{b}_i = f_{2D+3D}(X_i)$ 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 $\mathbf{a}_i, \mathbf{b}_i$ 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 $\mathbf{c}_i = \alpha\mathbf{a}_i + (1-\alpha)\mathbf{b}_i$ を得る．ただし，$\alpha \in [0,1]$ は行動ごとに学習を通して獲得される．この $\mathbf{c}_i$ から，全結合層で表現される関数 f_p により，骨格推定結果 $\mathbf{d}_i = f_p(\mathbf{c}_i)$ を得る．

3.4 ネットワークの学習

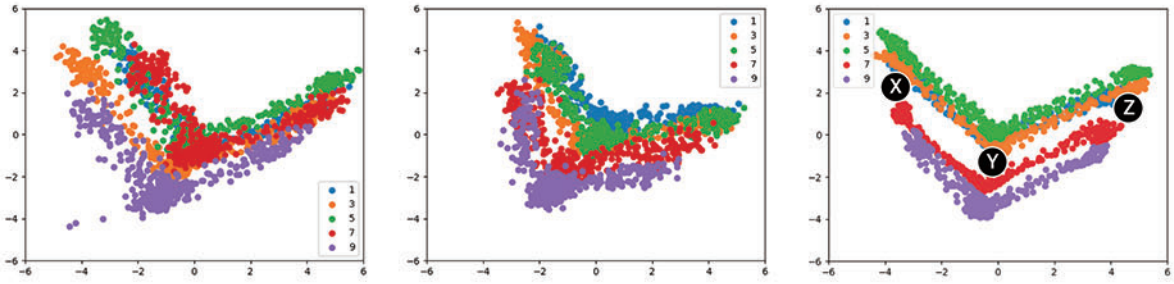
損失関数には関節点の教師信号と推定関節点位置の平均 2 乗誤差（Mean Squared Error ; MSE）を用いる．また学習したモデルにおける 2 つのネットワークの性質を明確にするため，全結合層に入力するベクトルを求める際の重みづけ和のための重み α が 0 又は 1 に近づくような制約を加える．従って，以下の損失関数を最小化するようにネットワークを学習する．

$$L = \frac{1}{N_{\text{total}}} \sum_i ||\mathbf{d}_i - \mathbf{y}_i||^2 + \lambda \left| \frac{1}{2}\alpha(\alpha - 1) \right| \qquad (1)$$

ここで，N_{total} はデータに含まれる画像の枚数，$\mathbf{y}_i$ は骨格空間内での各正解関節点位置の 2 次元座標を表し，また経験的に $\lambda = 0.01$ とした．この α に関する制約項により，学習したデータに含まれる FIR 画像内における行動に応じ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



(i) Comparative method ①　(ii) Comparative method ②　(iii) Proposed method

Fig. 9 Visualization of PCA feature in each frame at each human position (Action B)

めの実験を行なった．

データセットとして，まず 16 × 16 画素の赤外線センサアレイ（オムロン製 D6T-1616L）と 200 万画素の可視光カメラ（バッファロー製 BSW20KM11BK）を用い，以前に提案したデータ収集法 [2] に倣って図 4 のように FIR 画像に同期して可視光画像を撮影した．異なる行動における骨格推定精度を確認するため，2 種類のデータセットを構築した．対象とする行動は，行動クラス A（複数種類の静止状態により構成される）に含まれる，「着座状態から 1 度立って座り直す」（行動 A）と行動クラス B（関節点が連続的に移動している状態）に含まれる，「静止せずに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 人であり，センサに対する人物の位置の違いへの頑健性を検証するため，図 5 に示す 9 箇所で撮影した．本論文では人物の向きが変化しない状況を想定し，被験者をすべて正面から撮影した．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．そして各 FIR 画像に対応した可視光画像に既存の骨格推定法 [9] を適用することにより，人物の骨格を推定し，教師信号を取得した．得られる関節点数は $J = 21$ である．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おいて，関節点 P，Q 間の長さが行動 A では 0.625，行動 B では 0.333 となるように正規化した．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．

実験では図 2 のネットワークを用いる提案手法に対し，以下に示す 2 種類の手法を比較した．まずベースラインとして，我々の先行研究であり，FIR 画像から骨格を回帰により推定する手法 [2] を比較手法①とした．次に，pooling 層の除去及び global max-pooling の効果を確認するため，比較手法①に対し，畳込み層毎に pooling をせず，特徴抽出の最後に global max-pooling を用いる手法を比較手法②とした．比較手法①，②では 1 フレームの FIR 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する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 のデータとを用いた．また，学習データに含まれる人物位置での精度検証も行なうため，人物位置 5 でのデータを 2 セット撮影し，学習には用いていない人物位置 5 のデータを評価に用いた．

Table 1 RMSE of the ground truth and the estimation results ($\times 10^{-2}$)

Human position		2	4	5	6	8	Average
Action A	Comparative method ①	3.75	5.32	3.74	5.01	8.05	5.17
	Comparative method ②	3.30	4.51	3.55	4.01	**3.96**	3.87
	Proposed method	**3.08**	**3.54**	**3.50**	**3.69**	4.20	**3.60**
Action B	Comparative method ①	4.91	6.78	3.39	7.45	4.96	5.50
	Comparative method ②	3.07	5.58	**3.17**	5.94	4.57	4.47
	Proposed method	**3.06**	**4.95**	3.29	**5.02**	**3.57**	**3.98**

実験条件として学習とテスト時のバッチサイズは 32，エポック数は 1,000，学習アルゴリズムには Adam [15] を用い，NVIDIA Quadro GV100（Volta）の GPU を 2 台搭載した PC で実験した．

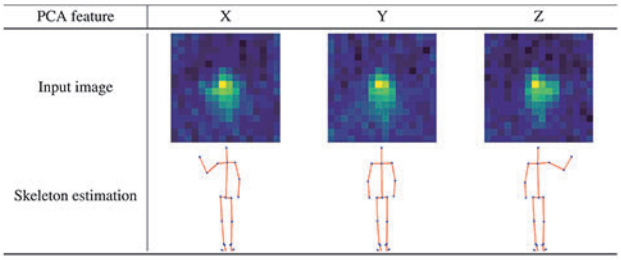
人物位置 3 において，提案手法による行動 A の骨格推定結果を図 8 に示す．また各行動における真値と各手法による推定結果の平方根平均 2 乗誤差（Root Mean Squared Error ; RMSE）を表 1 に示す．両方の行動において平均的に提案手法の方が高精度であった．比較手法②と提案手法と比較して，比較手法①は定量的にも定性的にも著しく精度が低かった．これは，各畳込み層の後に pooling 層を挿入することで，様々な位置のデータに対して，有用な情報を抽出できず，精度良く骨格推定をできなかったためと考えられる．一方，表 1 より，比較手法②及び提案手法では人物位置の違いに頑健になり，さらに提案手法では時系列情報を行動に合わせて有効活用できた．また推定に要する処理時間は 1 フレームあたり約 0.015 秒であったため，FIR 画像の標準的な入力に対して実時間処理が可能である．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，人に対する汎化性能は評価できていない．これはデータセットの拡充が必要なため，今後の課題とする．

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，ニューラル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，かつ計算量を削減できるという利点がある．しかし，本研究で対象としている超低解像度 FIR 画像では 1 画素が画像全体に対して相対的に非常に大きな割合を占めることから，比較手法①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．

Table 2 Example of RGB images and skeleton estimations corresponding to the representative PCA feature shown in Fig. 9 (iii)



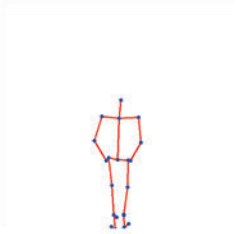


Fig. 10 Example of the skeleton of a standing moment in action A

Fig. 11 Example of a sequence of consecutive frames in action B

Table 3 RMSE of the ground truth and the estimation results ($\times 10^{-2}$) with different network structures

	Human position	2	4	6	8	Average
Action A	3D CNN method	3.47	4.05	4.05	4.40	3.93
	3D+2D CNN method	**3.08**	**3.44**	**3.53**	**4.05**	**3.52**
	Proposed（3D & 3D+2D）method	**3.08**	3.54	3.69	4.20	3.60
Action B	3D CNN method	3.29	5.24	5.14	**3.16**	4.32
	3D+2D CNN method	3.13	5.28	5.19	3.39	4.25
	Proposed（3D & 3D+2D）method	**3.06**	**4.95**	**5.02**	3.57	**3.98**

Table 4 RMSE of the ground truth and the estimation results ($\times 10^{-2}$) when the value of λ is changed

Human position		2	4	6	8	Average
Action A	$\lambda = 0.001$	3.16	4.15	4.07	4.14	3.88
	$\lambda = 0.01$	**3.08**	**3.54**	**3.69**	4.20	**3.63**
	$\lambda = 0.1$	3.15	4.08	4.38	**3.96**	3.89
Action B	$\lambda = 0.001$	**2.88**	5.24	**5.02**	3.51	4.16
	$\lambda = 0.01$	3.06	**4.95**	**5.02**	3.57	4.15
	$\lambda = 0.1$	2.91	5.23	5.06	**3.37**	**4.14**

それぞれの手法で，行動 B 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を主成分分析（Principal Component Analysis ; PCA）を用いて可視化したものを図 9 に示す．ここでは各手法において，最終層の直前の層の出力値を用いた．

比較手法①（図 9 (i)）と比較手法②（図 9 (ii)）を比較することで，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，提案手法では人物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入力画像と骨格推定の例を表 2 に示す．図 9 と表 2 から，骨格の変化をうまくとらえた特徴空間が得られていることを確認できる．以上のことから，global max-pooling の導入と，行動に適したネットワークを選択的に統合することで，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ることが分かった．

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，2 つのネットワークの出力の重み付き和において，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に各々特徴を重視するように重み α の学習が進んだ．行動クラス A では，立った状態，座った状態など 2 つの状態の切り替わりが存在する．この切り替わりをまたぐ画像列を入力する場合，時系列を重視すると，推定精度が低下してしまう．これに対し，時系列情報を重視しない 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に働いたと考えられる．また，図 10 のような立つ瞬間，座る瞬間などの状態の切り替わり中でも，ある程度推定ができることが確認できた．一方，行動 B では図 11 のように全体的に連続的な動作を行なっていたため，直近のフレームからの行動予測がしやすく，各畳み込み層において時系列方向の 5 フレームの情報を用いながら特徴抽出することで，比較手法①，②と比較して，定性的にも滑らかな推定ができた．

さらに，提案手法のネットワーク（図 2）のそれぞれの分岐である 3D 畳み込みネットワークと 3D+2D 畳み込みネットワークのみで，4 章と同条件で実験を行なった．このときの各行動における真値と各ネットワークにおける推定結果の平方根平均 2 乗誤差を表 3 に示す．ここで，3D 畳み込みネットワークのみを用いる手法を 3D CNN method，3D+2D 畳み込みネットワークのみを用いる手法を 3D+2D CNN method と表記する．3D 畳み込みネットワークと 3D+2D 畳み込みネットワークを定量的に評価すると，行動 A においては 3D+2D 畳み込みネットワークが優位であるが，両方のネットワークを用いる提案手法の精度が平均的に最も高精度であった．

また，3.4 節で示したネットワークの学習における式（1）内の λ について，4 章では $\lambda = 0.01$ のみで実験を行なったが，その他の値において，同様の実験を行なった結果を表 4 に示す．行動 B では大きな差は見られなかったが，平均的に $\lambda = 0.01$ が最も高精度であった．

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て実験を行なったが，被験者ごとの動きの違いや様々な行動，学習率の変更による極小値への影響などをふまえ，今後さらに大規模なデータセットを作成し，検証する必要がある．

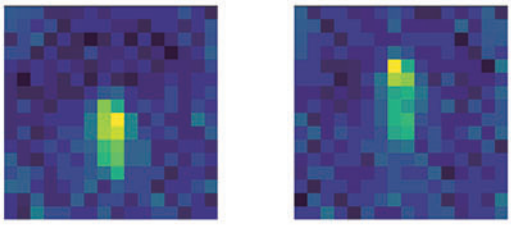
5.3 学習に用いるデータと精度差について

4 章の実験では図 5 中の人物位置 1, 3, 5, 7, 9 で撮影したデータを学習に用い，それらを内挿した位置でテストを行なった．実験結果より，撮影範囲内にある等間隔の 9 つの人物位置において，人物骨格が推定可能であったことから，その範囲での他

Fig. 12 RMSE sequence in each frame of skeleton estimation and ground truth for Action A at human position 2 by the proposed method

Table 5 RMSE of the ground truth and the estimation results by the proposed method ($\times 10^{-2}$)

Human position		2	4	6	8	Average
Action A	Training data a	8.99	5.41	6.00	5.02	6.35
	Training data b	**2.75**	5.30	5.10	7.19	5.09
	Training data c	3.67	**3.51**	**3.61**	**4.31**	**3.70**
Action B	Training data a	11.4	11.5	8.49	5.57	9.25
	Training data b	**2.97**	5.36	6.64	4.65	4.91
	Training data c	3.85	**4.74**	**5.09**	**4.29**	**4.32**



(i) Sitting pose at human position 2　　(ii) Standing pose at human position 5

Fig. 13 FIR image example at each human position

Table 6 RMSE of the ground truth and the estimation results for each joint point ($\times 10^{-2}$)

Joint point	Head	Hand	Knee
RMSE	3.64	3.90	3.86

Table 7 Variance of the RMSE of the ground truth of each part and the estimation results ($\times 10^{-6}$)

Training data	a	b	c
Variance	26.1	1.85	1.23

の人物位置における推定も可能であると考えられる．

そこで，さらに追加実験として，学習に用いるデータを

- 学習データ a：人物位置 5
- 学習データ b：人物位置 1, 3, 5
- 学習データ c：人物位置 1, 3, 5, 7, 9

で観測したものに限定し，提案手法の外挿性を評価した．4 章における実験と同様，真値と推定結果の RMSE を表 5 に示す．ただし，それぞれの行動における，学習データ c を用いた結果は，表 1 における提案手法と同じである．学習データ a のとき，前列である位置 2 における誤差が非常に大きくなった．後列である人物位置 8 では FIR 画像における人物位置の占める割合が小さいため，全体的に動きが小さくなり，前列よりも誤差が小さくなったと考えられる．学習データ b を用いた場合は，学習データにおける前列の割合が大きくなるため，人物位置 2 による誤差が最も小さくなった．平均誤差評価としては，より多くの位置におけるデータを用いた学習データ c が最も高精度であった．

さらに，各学習データを用いて提案手法により人物位置 2 における行動 A で骨格推定した結果と真値との各フレームにおける RMSE 系列を可視化したものを図 12 に示す．図より，学習データ a を用いた場合，人物が座っているときの RMSE（図中の黄色い枠）が特に大きいことが分かる．これは図 13 に示すよう学習データ a（人物位置 5）における起立時の FIR 画像の見えと，テストデータ（人物位置 2）における着席時の FIR 画像の見えが似ていて区別できていないことが原因であると考えられる．一方，学習データ b，c で学習した時は行動毎の RMSE の分散も小さくなり，定性的におおむね正しく骨格推定ができた．

5.4　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，関節点位置ごとの真値と提案手法による推定結果の RMSE の平均を表 6 に示す．ここで，頭と膝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であり，左右の手の部位における推定値の平均とする．定量的に見ると，行動の中でよく動いている手の誤差が一番大きくなっているが，誤差の差は大きくはない．さらに，上記の考察の追加実験で用いた，学習データ a，b，c の

それぞれを用いて検証した，頭，手，膝の関節点の RMSE 誤差の分散を表 7 に示す．より多くの人物位置でのデータを学習に用いることによって，それぞれの関節点誤差の分散が減少し，平均的に精度が向上したことが分かる．

6. 結　　言

本論文では，超低解像度 FIR 画像内での人物位置と行動の違いに着目した人物骨格推定手法を提案した．

人物位置の違いに頑健にするため，global max-pooling を導入し，さらに行動の違いに頑健にするため，空間的情報と時系列情報を選択的に活用するニューラルネットワーク構造を提案した．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，未学習の人物位置での骨格推定精度を評価した．その結果，人物の行動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．

今後の課題として，撮影対象者や行動，被験者の向き，服装の種類，撮影環境条件などの違いに対処するため，各条件を増やした実験を行ない，詳細な分析をすることが挙げられる．また行動に応じて適切な特徴を抽出できるネットワークの提案や骨格推定精度を定性評価に近い基準で評価できる定量化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（2020/5/12 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,（2019）.
3) MS. Ryoo, B. Rothrock, C. Fleming and HJ. Yang: Privacy-preserving human activity recognition from extreme low resolution, Proc. Thirty-First AAAI Conf. on Artificial Intelligence, (2017) 4255.
4) Y. Bai, G. Dai and L. Chen: Extreme low resolution activity recognition with spatial-temporal attention transfer, Computing Research Repository arXiv preprint, 1909.03580 (2019).
5) 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 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, 動的画像処理実利用化ワークショップ（DIA）2020, OS5-1,（2020）.
6) 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, (2018) 7103.
7) B. Xiao, H. Wu and Y. Wei: Simple baselines for human pose estimation and tracking, Proc. Fifteenth European Conf. on Computer Vision, 6, (2018) 466.
8) A. Newell, K. Yang and J. Deng: Stacked hourglass networks for human pose estimation, Proc. Fifteenth European Conf. on Computer Vision, 8, (2016) 483.
9) Z. Cao, T. Simon, S. Wei and Y. Sheikh: Realtime multi-person 2D pose estimation using part affinity fields, Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition, (2017) 7291.
10) G. Papandreou, T. Zhu, L. Chen, S. Gidaris, J. Tompson and K. Murphy: PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model, Proc. Fifteenth European Conf. on Computer Vision, 14, (2018) 269.
11) 岡田遼太郎, 矢入郁子：プライバシーに配慮した屋内行動モニタリングシステムの提案, 第 27 回人工知能学会全国大会, 1D-3,（2013）.
12) 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,（2015）.
13) H. Fujita and S. Otsuka: Posture detection for elderly using infrared array sensor and fine tuning, Proc. 2018 IEEE Visual Communications and Image Processing Conf., (2018) 1.
14) 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, **84**, 12,（2018）1025.
15) D. Kingma and J. Ba: A method for stochastic optimization, Proc. Third Int. Conf. on Learning Representations, (2015) 4.