CONFIDENTIAL EXTENDED ABSTRACT.
DO NOT DISTRIBUTE ANYWHERE.

低解像度遠赤外線画像を用いた人物の頭部向き推定

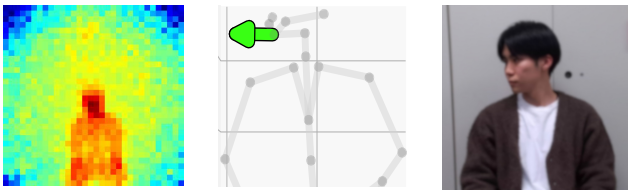
玉木 太耀[1,2,a)]　薗頭 元春[2,1]　北野 和哉[1]　藤村 友貴[1]　舩冨 卓哉[1]　向川 康博[1]　川西 康友[2,1,b)]

概要

本研究では 32 × 32 画素の遠赤外線センサアレイで撮影された低解像度遠赤外線 (LFIR) 画像から，人物の頭部の向きを推定する手法を提案する．LFIR 画像は，プライバシの保護や暗所での撮影に優れる一方，ノイズの混入と限られた解像度のために単純な手法で頭部の向きを推定することは困難である．また，手動でアノテーションすることも困難であり，現在適切なデータセットが存在しないという課題がある．これらの課題に対処するため，本研究では自動アノテーションシステムを用いて構築した LFIR2Head データセットを提案する．加えて，ノイズや情報欠損に対して正確かつ頑健な推定を実現するため，マルチモーダル情報および時系列変化を活用する推定モデル LFIR2Head モデルを提案する．提案手法の有効性を確認するため，構築したデータセットを用いて実験を行い，頭部向き推定性能を評価した．

1. はじめに

非接触型センサから人間の身体の動きを正確に把握するのは，高齢者ケア，健康のモニタリング，転倒予防などのアプリケーションで極めて重要である [1]．身体の姿勢，特に身体の部位の向きや微妙なねじれを詳細に分析することで，個人の身体状態や潜在的なリスクに関する重要な手がかりが得られる．従来の研究では，装着型ウェアラブルデバイスや，プライバシの問題を伴う可視光カメラに依存してきた [7]．近年では，低解像度遠赤外線（Low resolution Far-infrared:LFIR）センサアレイが注目されており [5], [8]，個人のプライバシーを保護しつつ，あらゆる照明条件下で継続的なモニタリングを可能にする．LFIR 画像に基づく研究は主に行動認識に焦点を当てているが，いくつかの研究 [2], [5] では，人間の姿勢に関する詳細な情報を取得するために身体の関節位置の推定に取り組んでいる．しかしながらこれらの研究では，関節の向き (回転角度) や微妙な体のひねりなど，人間の姿勢のより細かな側面についてはまだ検討されていない．身体の関節位置だけでなく，関節の向きを推定することで，人間の包括的な人体姿勢表現が得られ，動作意図の把握や関節状態の定量的評価に有効である．その中でも頭部は最も重要な身体部位であり，その向きは人物の意図を反映する重要な手がかりとなる．



(a) LFIR 画像 (入力)　(b) 頭部向き (出力)　(c) RGB 画像

図 1: LFIR 画像，頭部向き推定結果（緑の矢印），対応する RGB 画像の例．

本研究では，その予備的検討として頸部関節の向きの推定，すなわち頭部向き推定を扱う．具体的には，LFIR 画像を入力とし，人物の頭部が向いている方向を表すピッチとヨーを推定する問題を扱う．図 1 は，入力として用いた LFIR 画像，推定された頭部向きを示す．また，参考として対応する RGB 画像を (c) に示す．

LFIR 画像から頭部向きを推定する際，いくつかの課題が生じる：(i-a) LFIR 画像のノイズが安定した特徴抽出を妨げること，(i-b) 低解像度により幾何学的特徴が失われ，頭部の特徴が不明瞭になり，正確な推定が妨げられること，(ii) 頭部向き推定のための信頼できる教師信号が付与された LFIR 画像データセットが存在しないこと，が挙げられる．

そこで本研究では，上記の課題を解決する手法を提案する．第一に，LFIR 画像系列から頭部の向きを推定するための手法である LFIR2Head モデルを提案する．本モデルは，LFIR 画像系列から時系列的変動を捉え，課題 (i-a) および (i-b) に対処するものである．まず，ノイズ (i-a) の影響を軽減するために，自己教師あり学習を採用し，同じ頭の向きを捉えた隣接フレーム間で類似した特徴表現が得られるよう学習を行う．次に，低解像度の問題 (i-b) に対

1　奈良先端科学技術大学院大学
2　理化学研究所
a)　taiyo.tamaki@riken.jp
b)　yasutomo.kawanishi@riken.jp

処するために，顔領域の局所的特徴を抽出し細かな手がかりを捉えるとともに，頭部と肩の空間的関係を捉えた上半身骨格情報を統合する．そして，これらの特徴量を，multimodal-temporal フレームワークの中で統合する．これにより，本手法はノイズや低解像度の悪影響を低減し，推定の安定性を向上させる．

第二に，(ii) に対しては，広範囲の頭部向きを網羅的に捉えた LFIR2Head データセットを提案する．このデータセットは，32 × 32 解像度の遠赤外線センサアレイを用いて撮影した．LFIR センサアレイの解像度は，プライバシ保護と正確な頭の向き推定のバランスを考慮して選択する．LFIR 画像における頭部向きの手動アノテーションは，ノイズや低解像度のために極めて困難である．そこで，各 LFIR 画像に関節の位置と向きを自動的にアノテーションする手法を導入する．これにより，頭部の向き，身体の関節位置，LFIR 画像データ間の信頼性の高い対応関係を確保することで，自動アノテーションを実現する．LFIR2Head データセットを用いた広範な評価により，提案する LFIR2Head モデルがノイズと低解像度の影響を抑えつつ，さまざまな環境条件や被験者に対しても正確かつ頑健に頭部の向きを推定できることを確認する．

2. 提案手法

本研究では，頭部向き推定のために LFIR 画像系列から 32 × 32 画素の画像 N フレームを入力し，その最終フレームにおける頭部向きとしてピッチとヨーを表す 2 次元ベクトルを出力する Transformer に基づくネットワーク LFIR2Head を提案する．図 2 に概要を示す．

2.1 LFIR2Head モデルの構造

LFIR 画像からの頭部向き推定は，ノイズの影響 (i-a) および低解像度 (i-b) により特に困難である．これらの課題に対処するため，本アーキテクチャでは，データの相補的な情報を活用すべく，3 種のエンコーダで構成されるマルチモーダル特徴抽出モジュールを導入した．Global Encoder は，ノイズ軽減 (i-a)，Face Encoder および Bone Encoder は，低解像度の問題 (i-b) に対処する．

Global Encoder Global Encoder は，ノイズの影響を低減することを目的とし，各 LFIR フレームから頑健な大域的画像特徴を抽出する．このエンコーダは，2.2 節に詳述するように，類似の頭部向きに対して類似の特徴を生成するよう事前学習し，ノイズに対する頑健性を確保する．各 LFIR フレームは，Global Encoder により特徴マップとして表現される．ここで，i 番目のフレームに対する特徴マップを $\mathbf{f}_i$ とし，その中の j 番目の位置に対応する特徴ベクトルを $\mathbf{f}_{ij}$ と定義する．

Face Encoder および Bone Encoders 低解像度という課題に対応するため，本研究では微細な局所的特徴を捉

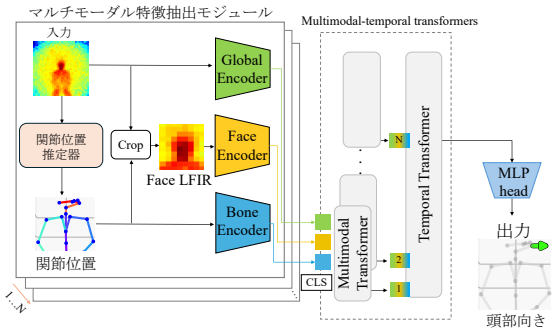


図 2: LFIR2Head モデルは，マルチモーダル特徴抽出モジュール，multimodal-temporal transformers から構成され，頑健な頭部向き推定を可能にする．

える Face Encoder と，頭と肩の相対的な関係性などの構造的情報を捉える Bone Encoder という，2 つの相補的なエンコーダを採用する．これらの特徴抽出には関節位置の特定が不可欠であり，当該特徴は，頭部および上半身を表現する 12 点の関節位置を検出する関節位置推定器によって抽出する．LFIR 画像における関節位置推定には，Iwata ら [5] が提案した 3 次元畳み込みニューラルネットワーク (CNN) を用いる．

Face Encoder は特定した頭部領域を入力として，微細な熱パターンや形状特徴を抽出するエンコーダである．頭部領域は，12 点の関節位置の検出された主要な顔ランドマーク（鼻，目，耳）を用いて頭部中心点を算出し，この中心点を基準に 12 × 12 画素の固定領域を切り出すことで得られる，このように領域を限定することで，低解像度という制約化においても，正確な頭部向き推定に不可欠な微細な熱パターンおよび形状詳細に焦点を当てることが可能となる．

Bone Encoder は，グラフニューラルネットワーク（GNN）を用いて関節位置を処理し，肩と頭部の相対位置といった空間的関係性をモデル化することにより構造的特徴を抽出する．これにより，熱パターンの変動に対して影響を受けにくい，頑健な構造情報が得られる．

Multimodal-Temporal Transformers 上記のエンコーダから得られる 3 つのモダリティ特徴 (Global, Face, Bone) は，各モダリティ間の関係性を捉えるために Multimodal Transformer により統合される．各モダリティから抽出された特徴ベクトルは，同一次元の特徴空間に変換され，各フレームごとにまとめられ，学習可能なクラストークンとともに Multimodal Transformer に入力される．ここで自己注意機構によってモダリティ間の特徴が統合され，得られたクラストークンが各フレームの統合表現として抽出される [10]．これらを時系列順に並べ，Temporal Transformer に入力することで，モデルは頭部運動の時間的変動を学習し，各フレームに対する時系列処理後の特徴表現を出力する．これらの出力トークンは MLP ヘッドに入力され，各

フレームにおける 2 次元頭部向きベクトル（ピッチ，ヨー）を回帰する．推定段階では最終フレームのみを使用し，学習時には全フレームの出力に対して誤差を算出する．

2.2 2 段階学習

Global Encoder の事前学習 Global Encoder の事前学習では，LFIR 画像から頑健な特徴表現を学習するため，対照学習 [3] を採用する．ここでは，あるフレームを基準 (アンカー) として，その頭部向きに類似したフレームを正例，異なるフレームを負例として定義する．正例・負例は，同一バッチ内の他フレームとの頭部角度差に基づいて構成する．アンカーとなるフレームのインデックスを $a \in \{1, \ldots, N\}$ としたとき，頭部向き差 $l(a,p) < \tau_{\text{pos}}$ を満たすフレームインデックスの集合を正例集合 $\mathcal{P}_a$ と定義する．

$$\mathcal{P}_a = \{\, p \mid l(a,p) < \tau_{\text{pos}},\ p \neq a \,\}, \tag{1}$$

ここで $l(\cdot,\cdot)$ は，与えられた 2 フレーム間の真値頭部姿勢ベクトルの絶対角度差を返す関数を示す．同様に，頭部向き差 $l(a,n) > \tau_{\text{neg}}$ を満たすフレームインデックスの集合を負例集合 $\mathcal{N}_a$ と定義する．

$$\mathcal{N}_a = \{\, n \mid l(a,n) > \tau_{\text{neg}},\ n \neq a \}. \tag{2}$$

上記の正例集合 $\mathcal{P}_a$ および負例集合 $\mathcal{N}_a$ の定義を用いて，アンカーフレーム a の j 番目のピクセルに対するピクセル単位の InfoNCE 損失 [3] を次のように定義する：

$$\ell_{aj} = -\log \frac{\displaystyle\sum_{p \in \mathcal{P}_a} \exp\left(\frac{\mathrm{s}(\mathbf{f}_{aj}, \mathbf{f}_{pj})}{\tau}\right)}{\displaystyle\sum_{\mathbf{f}_i \in \mathcal{P}_a \cup \mathcal{N}_a} \exp\left(\frac{\mathrm{s}(\mathbf{f}_{aj}, \mathbf{f}_{ij})}{\tau}\right)}, \tag{3}$$

ここで $\mathrm{s}(\cdot,\cdot)$ はピクセルにおける特徴間の類似度関数を示し，τ は温度パラメータである．この損失は，モデルがアンカーフレームの特徴 $\mathbf{f}_{aj}$ とその正例との類似度を最大化し，同時に負例との類似度を最小化することで，画素ごとにおける頑健な特徴学習を促す．

モデルの学習 Global Encoder および関節位置推定器は事前に学習済みとし，本訓練中はパラメータを更新しないものとする．LFIR2Head モデルの最適化には，学習の安定性と外れ値に対するロバスト性を考慮し，Huber 損失を採用する [4]．訓練時には，各フレームにおける予測頭部向きと真値頭部向きの差分（ピッチおよびヨーからなる 2 次元ベクトル間のユークリッド距離）に基づいて損失を定義し，全フレームの損失を合計した値を最小化するようにモデルを学習する．学習の最適化には Adam [6] を使用する．

3. 実験および考察

3.1 LFIR2Head データセット

LFIR 画像中の頭部向きを手動で直接アノテーションすることは，視覚的特徴の欠如により困難である．一方，高解像度の RGB-D データからは，顔のランドマークや骨格キーポイント推定が可能であり，許容可能な精度で頭部向きが自動的に得られる．これらの特性を活かし，本研究では LFIR センサと同期させた RGB-D センサを用いて頭部向きのアノテーションを取得する自動アノテーションシステムを構築した．データ収集には，32 × 32 解像度の FIR センサアレイ（Heimann Sensor 社製サーモパイル赤外線アレイ）と RGB-D センサ（Microsoft Azure Kinect）を使用した．Kinect は，その高精度な 3 次元関節追跡能力により，照明条件や被験者の個人差に影響を受けず頭部向き推定を可能とする．センサ間の同期は，Iwata ら [5] の手法を基に，本研究の設定に適合させる形で実装した．これにより，センサ間の整合性を確保し，頭部姿勢に対する正確かつ一貫性のあるアノテーションが可能となった．また，包括的な頭部向き推定を実現するため，撮影時には被験者に日常的な動作に基づく自然な頭部運動を実施させた．また，環境条件に対する頑健性を確保するため，複数の撮影シーンにおいて温度および照明条件を変化させて収録を行った．同時に，不要な熱源による影響を排除するため，被験者の周囲には本人以外の熱源が存在しないようにした．環境温度は，21.0 °C, 26.0 °C の ±0.5 °C に調整し，被験者 14 名に対しデータを収集した．データは合計 50,000 フレーム取得し，ピッチが ±40°，ヨーが ±60° の範囲をカバーしている．

3.2 実験概要

本研究では，平均誤差と安定性の 2 つの観点からモデルの性能を評価した．平均誤差は，平均絶対角度誤差（MAAE）を用いる [9]．MAAE は，推定角度と正解角度の差の絶対値の平均として定義され，ここではピッチとヨーの 2 次元に分けてそれぞれの推定誤差の傾向を詳細に評価した．一方，モデルがある誤差許容範囲内でどれだけ安定して正しく推定できるかを評価するために，さまざまな角度誤差閾値における真陽性率（TPR）と偽陽性率（FPR）をプロットした ROC 曲線の下面積（AUC）を指標として用いた．AUC についても，ピッチとヨーに分けて評価し，各方向での推定安定性を把握した．ここで陽性とは，推定誤差が閾値未満である推定を指し，陰性は誤差が閾値以上である推定と定義する．これら二つの指標を併用することで，推定結果の平均的なずれの大きさと，誤差許容領域での推定結果の安定性を同時に把握し，モデル性能を総合的に評価できるようにした．また，本研究は新規タスクに取り組んでおり，LFIR 画像から人物の頭部向きを推定する既存手法は存在しない．そのため，提案手法の有効性を評価するために，Iwata ら [5] のモデルに基づく 3 つの手法をベースラインとして比較実験を行った．2D-CNN 手法では，LFIR 画像専用に設計された CNN モデルを採用し，入力 LFIR 画像

表 1: 提案手法とベースライン手法との比較

	Angular errors ↓			AUC ↑	
	Pitch	Yaw	MAAE	Pitch	Yaw
2D-CNN	11.23°	18.34°	14.79°	0.84	0.79
3D-CNN	11.14°	16.05°	13.60°	0.81	0.78
2D-CNN + transformer	**9.85°**	14.53°	12.19°	0.82	0.80
LFIR2Head(proposed)	9.87°	**11.05°**	**10.46°**	**0.89**	**0.89**

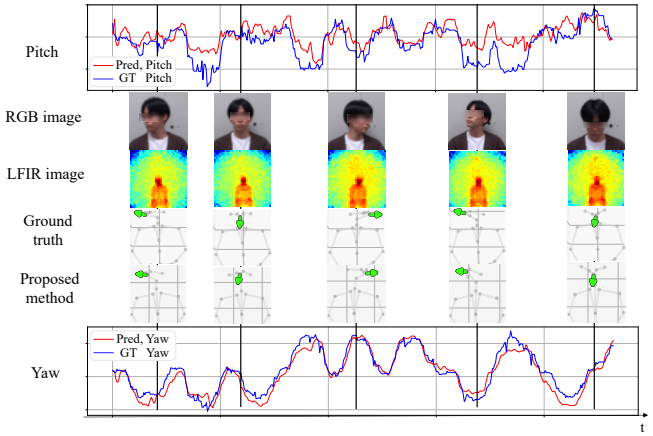


図 3: 頭部の向き推定結果の例．緑色の矢印は推定された頭部の向きを示し，参考のため骨格構造も重ねて表示している．

N を 1 として，モデル内の max pooling の代わりに，3 つの畳み込み層の後に global max pooling 層を組み込んだ．3D-CNN 手法では，Iwata らの 3D-CNN モデルを使用し，2D-CNN+transformer 手法では，2D CNN と transformer の組み合わせを適用し，2D CNN で抽出された空間的特徴と transformer による時間的情報を活用する．3D-CNN 手法と 2D-CNN+transformer 手法ともに，入力 LFIR 画像 N は 16 とした．これらの比較実験により，提案手法の有効性を分析し，性能を評価した．

3.3 実験結果

提案手法の有効性を検証するため，LFIR2Head データセットを用いて提案手法およびベースライン手法を学習し，それらの推定精度を比較した．定量的な結果を 表 1 に，頭部向き推定の一例を 図 3 に示す．提案手法は，頭部向き推定に関する両評価指標において，概ねベースライン手法を上回っていることが分かる．ピッチおよびヨーの AUC スコアは，それぞれベースライン手法よりも高く，角度誤差の低減と共に推定精度の改善が示唆された．2D-CNN ベースの手法と他の手法の結果から，マルチモーダル情報の活用や，複数フレーム使用して時間変動を考慮することが，ノイズや低解像度の影響を抑制し，より正確な頭部向き推定に寄与することが示された．

表 2: マルチモーダル特徴抽出モジュールの有効性の確認

	Angular errors ↓			AUC ↑	
	Pitch	Yaw	MAAE	Pitch	Yaw
No Encoder	**9.85°**	14.53°	12.19°	0.82	0.80
(a) Global Encoder	10.39°	14.74°	12.62°	**0.89**	0.84
(b) Face and Bone Encoders	10.51°	13.11°	11.81°	0.88	0.86
LFIR2Head(proposed)	9.87°	**11.05°**	**10.46°**	**0.89**	**0.89**

3.4 Ablation study

マルチモーダル特徴抽出モジュールの有効性を検証するため，ablation study を実施した．対象とした構成は，まず，提案したエンコーダを用いない 2D-CNN+transformer を用いたベースライン手法 (No Encoder)，次に，(a) Global Encoder のみを有効化し，Face Encoder および Bone Encoder を無効化した構成，さらに (b) Face Encoder および Bone Encoder のみを有効化し，Global Encoder を無効化した構成である．これらの結果を，提案手法 (Proposed method) と比較した．その結果を 表 2 に示す．結果として，ノイズ低減機能，ならびに顔特徴および骨格の抽出機能の両方を組み込んだ提案手法が概ね最も高い精度を達成し，これらのモジュールの有効性が示唆された．(a) の Global Encoder と (b) の Face Encoder および Bone Encoder の性能を比較すると，Global Encoder はピッチの推定においてわずかに優れた性能を示す一方，Face および Bone Encoder は頭部周辺のより微細な局所的詳細を捉えるためヨーの推定において優れていることが分かる．

4. おわりに

本研究では、32 × 32 画素の FIR 画像から人物の頭部の向きを推定する手法を提案した．提案手法による頭部向き推定結果の定量・定性的な評価により，32 × 32 画素の FIR 画像系列から人間の頭の向きを推定可能であることを示した．また，自動アノテーションシステムを用いて LFIR2Head データセットを構築し，この領域における適切なデータセットの不足という課題に対処した．本データセットを用いた広範な実験を通じて，本タスクに対する提案手法の有効性と実現可能性を確認した．しかしながら，本実験は被験者以外の熱源が混在する状況を十分には扱っていない．今後の課題として，実環境下で周囲の人物や暖房機器など多様な外部熱源が混在していても頑健に推定できる工夫が必要であると考えられる．

参考文献

[1] Agrawal, D. K., Usaha, W., Pojprapai, S. and Wattanapan, P.: Fall risk prediction using wireless sensor insoles with machine learning, *IEEE Access*, Vol. 11, pp. 23119–23126 (2023).
[2] Arai, T., Isogawa, M., Sakurada, K. and Sugimoto, M.:

3D Human Pose Estimation Using Ultra-low Resolution Thermal Images, *SIGGRAPH Asia 2024 Posters*, SA '24, Association for Computing Machinery (2024).

[3] Chen, T., Kornblith, S., Norouzi, M. and Hinton, G.: A simple framework for contrastive learning of visual representations, *International Conference on Machine Learning*, Vol. 149, pp. 1597–1607 (2020).

[4] Huber, P. J.: Robust Estimation of a Location Parameter, *The Annals of Mathematical Statistics*, Vol. 35, No. 1, pp. 73–101 (1964).

[5] Iwata, S., Kawanishi, Y., Deguchi, D., Ide, I., Murase, H. and Aizawa, T.: LFIR2Pose: Pose Estimation from an Extremely Low-resolution FIR image Sequence, *25th International Conference on Pattern Recognition*, pp. 2597–2603 (2021).

[6] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).

[7] Liu, J., Liu, H., Chen, Y., Wang, Y. and Wang, C.: Wireless Sensing for Human Activity: A Survey, *IEEE Communications Surveys & Tutorials*, Vol. 22, No. 3, pp. 1629–1645 (2020).

[8] Morawski, I., Lie, W.-N., Aing, L., Chiang, J.-C. and Chen, K.-T.: Deep-Learning Technique for Risk-Based Action Prediction Using Extremely Low-Resolution Thermopile Sensor Array, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 33, No. 6, pp. 2852–2863 (2023).

[9] Ruiz, A. and Chona, E.: Fine-Grained Head Pose Estimation Without Keypoints, *Proc. 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2074–2083 (2018).

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I.: Attention is All you Need, *Advances in Neural Information Processing Systems*, Vol. 30, pp. 6000–6010 (2017).