胸部装着カメラを用いた間接的視線推定のための頭部姿勢推定

Head Pose Estimation Using a Chest-Mounted Camera for Indirect Gaze Estimation

山田泰盛[1]　山添大丈[1]
Taisei Yamada[1]　Hirotake Yamazoe[1]

[1] 兵庫県立大学
[1] University of Hyogo

Abstract: This study aims to estimate gaze direction indirectly by predicting head and chest poses using a chest-mounted camera. To achieve this, we propose a deep learning-based head pose estimation method utilizing images captured from a chest-mounted camera. Since deep learning requires a large amount of training data, which is labor-intensive to collect, we generate training data using a CG model. Experimental results demonstrate the effectiveness of the proposed method, and we also present examples of its application to real images.

1　はじめに

日常生活において，人は無意識に興味を引くものに対し，顔や視線を向ける．自然環境で頭部の姿勢や視線を追跡することで，人々の行動を記録するだけでなく，その関心を推測することが可能になる．人の関心・興味対象を推測できれば，パーソナライズ化された情報を提供することができる．また，屋外広告や公共交通機関の大画面ディスプレイなど，デジタルサイネージの急速な普及に伴い，この技術は広告への注目度を分析するために活用できる．

興味対象を推測する上で視線は重要である．そのため，これまでに視線推定に関し多くの研究が存在する．カメラの設置場所によって大きく 2 種類，装着型と据置型に分けられる．装着型では，計測対象人物の目前方にカメラを装着し視線方向を推定するもので，移動人物の視線も取得できる．一方で，目の前方にカメラを装着する必要があるため，視野の妨げになるという問題がある．据置型は，環境中にカメラを設置し，計測対象人物を観測する．計測対象人物が装置を装着する必要はないが，視線取得が装置周辺に限られ，移動する人物の視線推定は難しい．このように，従来手法では，視野の妨げにならずに，移動する人物の視線情報を取得することは難しかった．

これに対し我々は，視線推定のために目領域を直接観測するのではなく，間接的に視線を推定する方法を検討している．提案手法の基本的なアイデアは，視線方向変化により，他の身体部位の動作に変化が生じることを利用し，目以外の身体部位の観測から，間接的な視線方向の推定を取得するというものである．

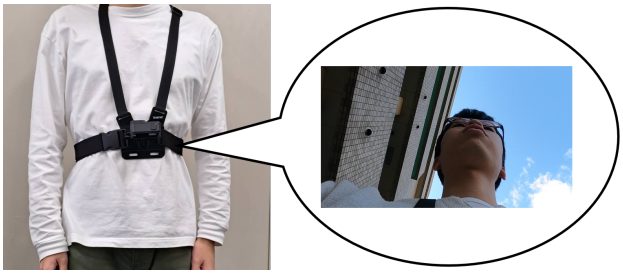

図 1: 胸部装着カメラを用いた頭部姿勢推定

人は目を動かすことで，周辺環境の視覚情報を取得しており，より広い範囲の情報を取得する際には，目だけでなく頭部や胸部（胴体）も動かしている．その際，目と頭部，胸部は協調的に運動するため，その関係性を調査した多くの研究が存在する [1, 2]．このような関係性を利用し，間接的に視線方向の推定を目指すのが本研究の基本アイデアである．本研究と同様に，視線・頭部方向の関係性を利用した視線推定を目指した手法として，Murakami らの手法 [3] も存在するが，ウェアラブルカメラの利用は想定していない．また，Nonaka らは，視線と頭部・身体方向の関係性を利用することで，遠隔からの視線推定を提案している [4]．

本研究では，図 1 に示すように，胸部にカメラを装着し，その画像から装着者の頭部・胸部姿勢を推定し，それらの姿勢変化から視線方向を間接的に推定することを目指しており，頭部・胸部姿勢変化に基づく視線推定の検討を進めてきた [5]．本稿では，この実現に向けた，胸部装着カメラによる頭部姿勢推定について検討する．

2　エゴセントリック法

人体に装着されたカメラを使用して 3D 姿勢を推定する手法を，エゴセントリック 3DHPE（3D Human Pose Estimation）と呼ぶ．外部カメラを必要としないため，広範囲での動作取得が可能で，スポーツ，アニメーション，ヘルスケアなど，様々な分野における実世界での活動の 3D 姿勢データ取得を目指し，研究が進められている．

Xu ら [6] は，帽子のつばに魚眼カメラを装着し，頭部を含む全身の姿勢を推定する手法を提案した．しかし，この手法では，カメラを装着した帽子をかぶる必要があるため，カメラの重量の問題があり，視覚的な影響も大きいといった問題もある．一方，Hwang ら [7] は，胸部に魚眼カメラを装着して，そのカメラの画像から全身の姿勢を推定する手法を提案している．この手法では，周辺環境に左右されず，またユーザーへの負担が軽減される．

本研究では，Hwang らの手法と同様に胸部に装着したカメラを用い，頭部姿勢のみに着目した，より高精度な姿勢推定を目指し，学習データ及び頭部姿勢推定モデルを構築し，その性能を評価した．

3　提案方法

3.1　学習データの作成

頭部姿勢推定モデルを学習するためには，大量の学習データが必要になるが，実際に人で画像データを撮影し，姿勢データを付与して学習データセットを作成すると膨大な時間と手間がかかる．そのため，本研究では CG モデルを用いて学習データを作成する．

Blender 内に CG モデルを配置し，モデルの胸部に上向きの仮想カメラに配置する．このとき，カメラの視点で CG モデルの頭部が収まるようにカメラ及びモデルの位置を調整する．仮想カメラは，実画像実験でも用いるカメラである「GoPro Hero 8 Black」を元にパラメータを調整した．背景画像は，より現実世界の撮影画像に再現するため,「Poly Haven」というサイト [8] で提供されている HDR(ハイダイナミックレンジ) 画像を使用した．

また，現実世界では，装着者の動きだけでなく，周辺環境，時間帯の変化など，様々な要因の影響を受け，撮影される画像が大きく変化する．これに対応するため，レンダリング毎に，仮想カメラの位置・姿勢を変化させ，さらに，仮想の光源の位置と色温度を変化させた．

図 2: 生成した学習データの例

3.2　頭部姿勢推定ニューラルネットワーク

頭部姿勢を推定する研究は近年盛んに行われているが，その手法は大きく分けて二つに区分される．一つ目は直接回帰法であり，これは画像から直接頭部姿勢を推定する手法である．画像を入力とし回転角を直接出力するため，処理がシンプルで他の手法に比べて推論速度が速い．二つ目は，キーポイントベース手法であり，これは顔の特徴点（鼻，目，口角，輪郭など）を検出し，そこから頭部姿勢を計算する手法である．2D の特徴点を 3D 空間のモデルと対応付けるため，物理的に一貫した結果が得られる．本研究では，胸部に装着したカメラから頭部を撮影するため，キーポイントベース手法による姿勢推定は難しい．そこで，本研究では直接回帰手法を用いる．

頭部姿勢を推定する際，一般的な回転表現としてオイラー角がある．しかし，オイラー角にはジンバルロックが生じる問題がある．ジンバルロックの状態では，一つの頭部姿勢に対して複数の回転パラメータが存在するため，ニューラルネットワークで正確な姿勢を学習が困難になる．一方，四元数表現はジンバルロックの問題は生じないが，対極対称性による問題が生じる．また，Zhou ら [9] は，3 次元回転について，4 次元以下の実ユークリッド空間ではすべての表現が不連続であることを示し，5 次元及び 6 次元の表現が連続的でネットワークの学習に適した表現であることを証明した．これを踏まえ，Hempel ら [10] は，9 パラメータの回転行列を 6 パラメータの回転行列に圧縮することで，頭部姿勢を推定する直接回帰型ネットワーク（6DRepNet）を提案した．

そこで，本研究では 6DRepNet を利用した頭部姿勢推定ネットワークを構築し，作成した学習データを用いてネットワークの学習を行い，その性能評価を検証した．

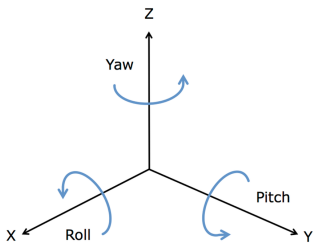

図 3: Rool 角，Pitch 角，Yaw 角の正回転方向

4　生成画像を用いた性能評価

性能評価を行うため，以下の通りの実験を行った．学習用データとして 53 名分の CG モデルを用い，Roll 角で-20〜20 度，Pitch 角で-30〜30 度，Yaw 角で-40〜40 度の範囲で 105,417 枚の画像を生成した．性能評価をするために，K-分割交差検証を用いた．K=5 として，CG モデルの人物ごとに 11 名分ずつを 1 グループとした 5 グループに分割し交差検証を行った．ヒストグラムに関しては，特定の角度のみ抽出し図示した．

まず，Roll，Pitch，Yaw について，各軸の 0 度と姿勢範囲の最小・最大値における姿勢推定結果の分布を図 6，6，6 に示す．横軸は推定値 (predicted)，縦軸は頻度 (Frequency) であり，3 方向回転の正負の方向は図 3 に示した通りである．ヒストグラムの結果より，0 ° に近づく傾向，つまり姿勢が小さめに推定される傾向が見られたものの，単峰性で正規分布に近い形状の分布になっていることがわかる．また，平均絶対値誤差（MAE）は，Roll，Pitch，Yaw の順に 2.544，5.169，3.156 度となった．

次に，姿勢推定誤差の分布を図 6 に示す．横軸が真値，縦軸が推定値であり，点線は y=x の直線であり，成果位置を示している． Yaw は，全体的に高精度で姿勢推定できていることがわかる．Roll は，MAE こそは低かったものの，全体的に推定誤差が大きく，ヒストグラムの結果から見られたように，推定結果が 0 度に近づく傾向が確認できる．特に± 10 ° 周辺の推定値にはばらつきが大きく出ている．Pitch も同様に，Pitch も同様に，全体的に推定誤差が大きく，推定結果が 0 度に近づく傾向が確認できる．

5　実画像への適用

胸部にカメラを装着するとともに，頭部に VR 用トラッカを装着し頭部姿勢を計測し，画像による推定値とトラッカによる計測値を比較した．結果を図 6 に示す．黒い波形は，VR 用トラッカで計測した実測値の軌跡であり，橙色，青色，緑色の波形はそれぞれ，提案手法の Roll，Pitch，Yaw の推定値である．多少の誤差はあるものの，推定値の軌跡は実測値とほぼ同じ軌跡を描いているのがわかる．

6　まとめ

本研究では，胸部装着カメラを用いた間接的視線推定を目指し，胸部装着カメラを用いた頭部姿勢推定手法を提案し，CG モデルに基づく学習用画像データの作成と，作成した画像を用いた頭部姿勢推定モデルの学習と評価を行った．さらに，実際に胸部にカメラを装着し，撮影した動画から頭部姿勢を推定できることを確認した．交差検証の結果，以前よりもネットワークの推定精度，汎化性能が大きく向上していることが分かった．また，実画像を用いた評価より，極端に頭部姿勢を変化させた場合を除き，安定して頭部姿勢を推定できることを確認した．一方で，頭部やカメラの姿勢が大きく変化した場合には，推定誤差が大きくなる傾向が確認された．これらの問題は，学習範囲の拡大で改善できる可能性はある．

さらに，現在の学習データでは，カメラと被写体の間にオクルージョンがある場合は想定しておらず，そういった場合に対応できていない．

そのため，近年提案されている ViT(Vision Transformer) に基づくオクルージョンに強いネットワーク（TokenHPE など）[11, 12] を参考に，オクルージョンにも対応できる姿勢推定を目指す．また，本研究では広角カメラを用いたが，Hwang らの手法のように魚眼カメラを利用することで，頭部だけでなく前方の環境も撮影できるため，頭部姿勢結果と前方環境から人の興味対象も取得できるシステムを目指していきたい．さらに，提案手法で推定された頭部方向に基づく間接的視線推定への拡張についても検討していく．

謝辞

本研究は、JSPS 科研費 23K26082, 21K11968 の助成を受けて実施した．

参考文献

[1] Fang, Y., Nakashima, R., Matsumiya, K., Kuriki, I. and Shioiri, S.: Eye-head coordination for visual cognitive processing, *PloS one*, Vol. 10, No. 3, p. e0121035 (2015).

[2] Yamazoe, H., Mitsugami, I., Okada, T. and Yagi, Y.: Analysis of head and chest movements that

correspond to gaze directions during walking, *Experimental Brain Research*, Vol. 237, pp. 3047–3058 (2019).

[3] Murakami, J. and Mitsugami, I.: Gaze from Head: Gaze Estimation Without Observing Eye, *5th Asian Conference on Pattern Recognition (ACPR2019)*, Springer, pp. 254–267 (2020).

[4] Nonaka, S., Nobuhara, S. and Nishino, K.: Dynamic 3D Gaze From Afar: Deep Gaze Estimation From Temporal Eye-Head-Body Coordination, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2192–2201 (2022).

[5] Yamazoe, H.: Indirect gaze estimation from body movements based on relationship between gaze and body movements, *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, pp. 1–2 (2023).

[6] Xu, W., Chatterjee, A., Zollhoefer, M., Rhodin, H., Fua, P., Seidel, H.-P. and Theobalt, C.: Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera, *IEEE transactions on visualization and computer graphics*, Vol. 25, No. 5, pp. 2093–2101 (2019).

[7] Hwang, D.-H., Aso, K., Yuan, Y., Kitani, K. and Koike, H.: Monoeye: Multimodal human motion capture system using a single ultra-wide fisheye camera, *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pp. 98–111 (2020).

[8] Haven, P.: High-quality 3D assets, `https://polyhaven.com`. Accessed: Feb. 8, 2025.

[9] Zhou, Y., Barnes, C., Lu, J., Yang, J. and Li, H.: On the continuity of rotation representations in neural networks, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5745–5753 (2019).

[10] Hempel, T., Abdelrahman, A. A. and Al-Hamadi, A.: 6d rotation representation for unconstrained head pose estimation, *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, pp. 2496–2500 (2022).

[11] Zhang, C., Liu, H., Deng, Y., Xie, B. and Li, Y.: Tokenhpe: Learning orientation tokens for efficient head pose estimation via transformers, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8897–8906 (2023).

[12] Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C. and Ding, Z.: 3d human pose estimation with spatial and temporal transformers, *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11656–11665 (2021).

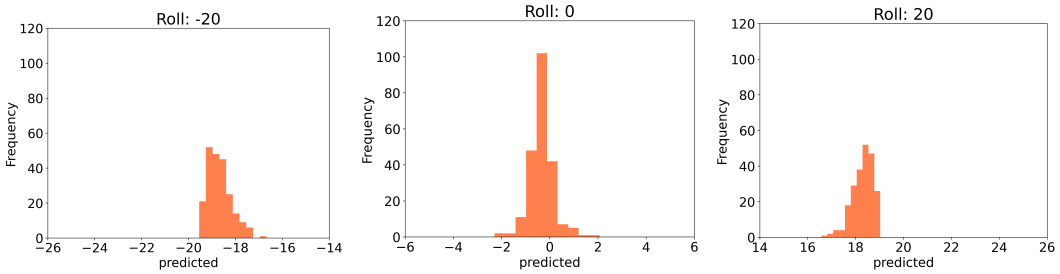

図 4: Roll のヒストグラム

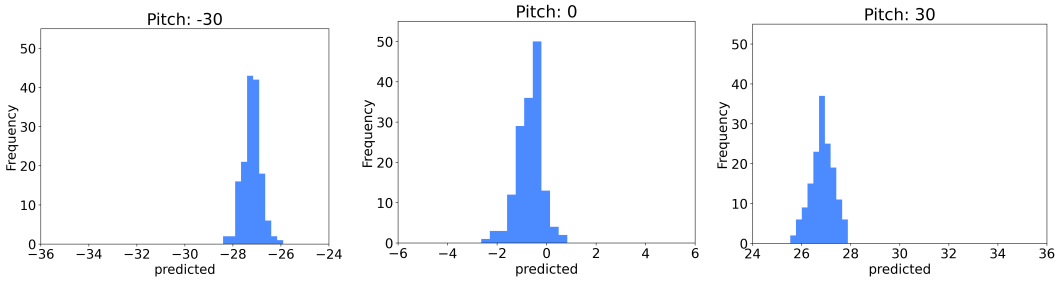

図 5: Pitch のヒストグラム

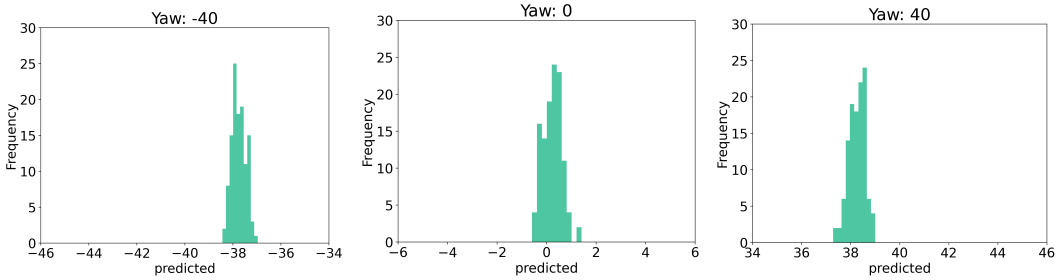

図 6: Yaw のヒストグラム

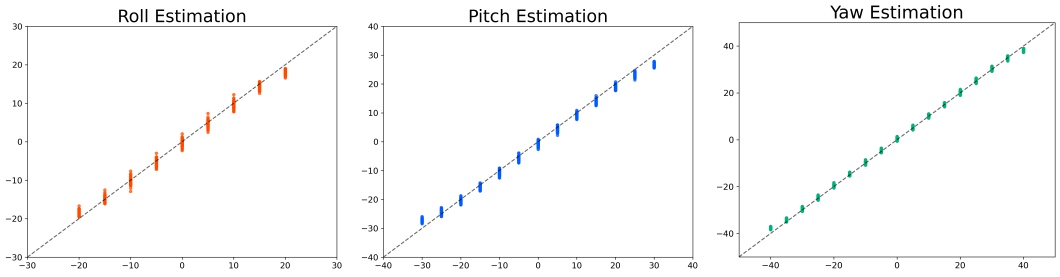

図 7: 姿勢推定誤差の分布

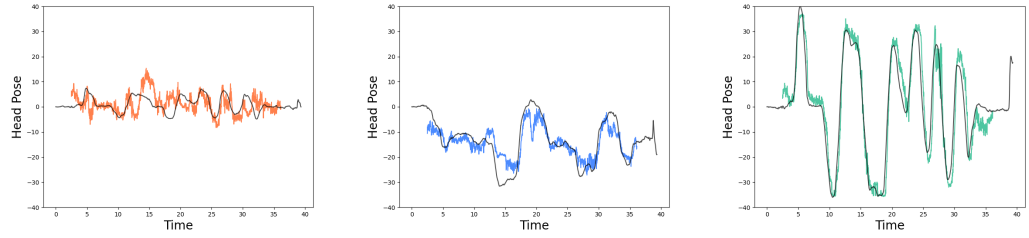

図 8: 実画像に適用した場合の推定値と実測値の比較