MoCap によらない三次元空間における身振りの自動位置推定システムの提案

代表研究者　家　永　直　人　筑波大学　システム情報系　助教
共同研究者　関　根　和　生　早稲田大学　人間科学学術院　准教授

1 はじめに

ジェスチャー（発話中に生じる自発的な身振り）に関する研究は、ジェスチャーが人のコミュニケーションにおいてどのような機能や役割があるのかを明らかにし、言語学、心理学、認知科学など、多くの分野に貢献してきた。ジェスチャー研究にはさまざまな研究があるが、ジェスチャーが行われる空間（ジェスチャー空間）は重要な情報伝達媒体であり、ジェスチャー研究の中でも大きなトピックの 1 つとなっている。

ジェスチャー空間のデファクト・スタンダードは、McNeill によって提案されたものである。図 1 に示されているように、ジェスチャー空間はこれまで主に二次元平面上の運動として記録されてきた。しかし、身体運動は三次元空間の中での「動き」として実現されていることを考えると、二次元平面上での解析には限界がある。実際に、対話や学習において、ジェスチャーの前後軸（奥行き方向）の情報が重要だと指摘さている[2]。

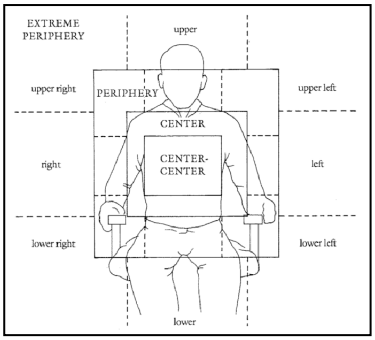

図 1. McNeill のジェスチャー空間[1]

従来の身体運動の三次元計測には、以下の 3 つの問題がある。(1) 一般的に光学式モーションキャプチャー（以下 MoCap）が利用されているが、MoCap は非常に高価で広い場所が必要であり、一部の研究機関でしか使えない。(2) MoCap での計測に必要な再帰性反射マーカーの装着が、話者の自然なジェスチャーの生成を妨害してしまう（我々の実験においても確認されたが、我々の研究はジェスチャーの質ではなく、計測精度の比較実験がメインであったため、問題はなかった）。(3) どの再帰性反射マーカーが、体のどの部位に対応するかのアノテーション（あるデータに対して、関連する情報を付与する作業のこと）に時間がかかる（MoCap のソフトウェアによっては、自動で各再帰性反射マーカーを追跡するものもある）。

従来研究では、三次元計測可能なデバイスを使用している場合でも、ジェスチャー空間を二次元平面として扱ってきた[3, 4, 5]。これは、ジェスチャー空間を三次元空間として定義することの難しさや、ジェスチャーの前後軸の使用の重要性が見逃されていることを反映している。前後軸情報が使用された研究もあり[6, 7]、聞き手もそれを利用する。Priesters と Mittelberg[8]は、MoCap を使いジェスチャーの三次元計測を行った。しかし、4 名の個人差を解析したにとどまっており、三次元ジェスチャー空間の提案には至っていない。Tillier ら[2]は、教師のジェスチャーに教育的価値があるという観点から、三次元空間による分析の重要性を提唱しているが、計測システムは提案していない。

現在の LLM（大規模言語モデル）や生成 AI の普及にも見られるように、機械学習や深層学習技術が急速に発展している。それらはさまざまな分野で利用され、精度も日々向上している。深層学習の技術の 1 つに、Human Pose Estimation（HPE）がある。HPE とは画像や動画から、キーポイントと呼ばれる人の体の関節などの各部位（足首、膝、腰、首、肘、手首、手や顔など）の位置を、二次元または三次元で推定する技術である。2014 年には、キーポイントの位置を回帰問題とすることで、HPE に深層学習を適用した研究が発表された[9]。その後、2017 年に OpenPose と呼ばれる手法が発表された[10]。同時に複数人のキーポイントを推定可能で、当時としては非常に高精度かつ高速に動作した。また、使いやすいように整備されたプログラムも公開されたため、当時のインパクトは非常に大きく、さまざまな課題に広く応用された。現在でも応用研究に利用されてはいるが、深層学習の発展の速さを考慮すると、より高精度な手法がある。

本研究の目的は、前述の MoCap の問題解決のため、深層学習による HPE を活用することで、MoCap によらない簡便な装置でキーポイント位置を三次元計測することである。そのためにはさまざまなことを検証しなければならないが、本研究では 4 種類の HPE の精度比較を行った。将来的には、さらに多くのことを検証して提案するジェスチャーの自動位置推定システムの実用性を高めた後、広く公開することでジェスチャーの研究分野に貢献する。また、ジェスチャー空間計測などのジェスチャー研究への応用も行う。

公益財団法人電気通信普及財団
研究調査助成報告書　第39号　2024年度

2 方法

2-1 実験手順

実験は、筑波大学システム情報系研究倫理委員会の承認を得た後（審査承認番 2023R737)、早稲田大学所沢キャンパスで 2023 年 7 月 21-25 日に、次の手順で実施された。(1) 研究の目的、内容、プライバシー保護、倫理的配慮について事前に文書を確認していただくとともに、重要な事項については口頭でも説明した。また、実験前に実験参加者の体調について確認し、同意書・ビデオ画像の公開についての承諾書を記入していただいた。そして、本実験の手順や課題について説明した。(2)実験参加者に再帰性反射マーカーを貼付した。マーカーは、両手親指の爪、両手中指の爪、両手小指付け根の側面、両手首、両肘、両肩、両膝、両足親指の爪、両頬、顎の先、計 19 か所に貼付された。なお、本研究では上記から小指と下半身のマーカーを除いた 13 点を実験に利用した。また、単眼手法のスケール調整のため、右肘から右肩の距離をメジャーで計測した（詳細は後述)。(3) 実験参加者は着座した状態で、参加者から見て右側に設置されたディスプレイに表示された動画を 2 回視聴した。動画は次の 3 種類で構成されていた。(i) 20〜30 秒程度の白黒の実写映像（チャールズ・チャップリンなど）8 本。(ii) 1 分程度のカラーのアニメーション映像（トゥイーティー、ワーナー・ブラザース）8 本。(iii) 10 秒程度の、丸や三角の図形のキャラクターのアニメーション映像 10 本。なお、本研究では解析に必要な時間の制約上、(ii)のみ解析した。(4) 実験参加者は動画の視聴後、正面のカメラに向かって動画の内容を説明した。この時、なるべくジェスチャーを交えながら説明するように指示された。また、各装置の開始・終了フレームを一致させるために、実験参加者の説明の前後で再帰性反射マーカー付きのカチンコを打った。どの再帰性反射マーカーがどのキーポイントに対応するかのアノテーションと、各動画と MoCap データの開始・終了の時刻合わせは手作業で行った。なお、著者らがカチンコを打つために動画に映ってしまうため、全てのデータの最初と最後の 5 秒間は分析から除外した。

フライヤーなどを作成し、早稲田大学内で実験参加者を一般に募集した。実験参加者は 10 名（21〜24 歳、平均年齢 22.1 歳、標準偏差 0.83 歳、男性 4 名）であった。キーポイントをより正確に検出するために、なるべくタイトな服装（T シャツやロング T シャツなど）をするように伝えられた。

2-2 手法

図 2 に実装のフローチャートを示す。実験参加者を MoCap（OptiTrack Flex 3)、デプスカメラ（Microsoft 社製 Azure Kinect)、3 台のビデオカメラ(Panasonic 社製 HC-VX992MS)で撮影した。MoCap は正解データとして取得した。つまり、HPE が推定した各キーポインの三次元位置と MoCap の計測値を比較することで、HPE の精度を検証した。Kinect は深度センサーを搭載したカメラであり、それ単体でキーポイントの三次元位置を推定可能である。しかし、本研究では MoCap や HPE とフレーム単位で正確に同期させられなかったため、結果の解析をしなかった。一方、予備的に精度を HPE と比較した結果、単眼手法よりは高精度だが、ステレオ手法よりは大きく精度を落とすことがわかった（詳細は後述)。HPE は単眼、つまり 1 台のビデオカメラで撮影された動画から、そこに映る人のキーポイントの二次元（画像上の）位置を推定できるのはもちろんのこと、三次元位置を推定できる手法もある。だが、単眼では奥行情報が一切ないため、単眼での三次元推定（以下単眼手法）の精度は高くないことが予想された。そこで、2 台のビデオカメラの動画において HPE で二次元位置を推定した後、古典的なコンピュータ・ビジョンの三角測量の手法により、2 つの二次元位置から三次元位置を計算する手法（以下ステレオ手法）も実装した。3 台のビデオカメラは、実験参加者の正面と、左右斜め前に設置された（図 3)。正面のビデオカメラを単眼手法に、左右斜め前 2 台のビデオカメラをステレオ手法に使った。ステレオ手法用の 2 台のビデオカメラは、実験前にチェッカーパターンによりキャリブレーションされた[11]（図 4)。

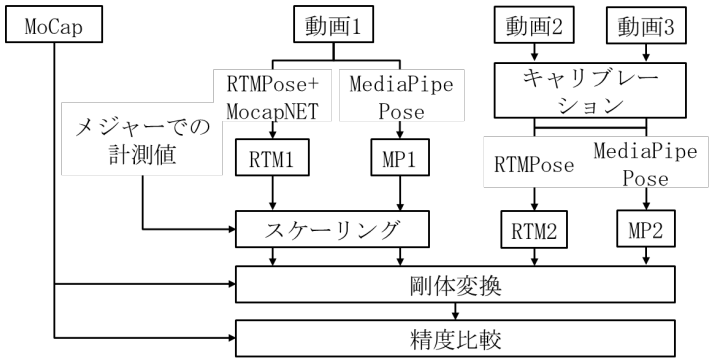

図 2. 実装のフローチャート

HPE は次の 2 つの手法であった。1 つ目は、実装時に state-of-the-art の性能であった RTMPose である[12]。

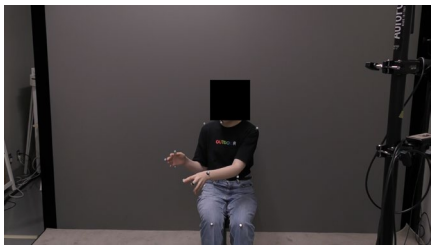

図 3. 正面のビデオカメラで撮影している様子

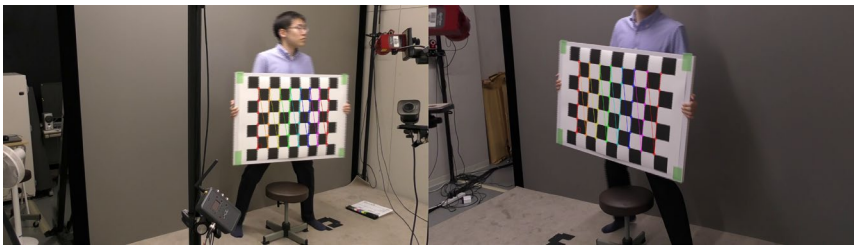

図 4. キャリブレーションの様子

RTMPose は、SimCC[13]の活用や学習方法の改良により、安価な計算資源でも高速に動作するように設計された手法である。SimCC とは、画像を水平軸と垂直軸のビンに分割し、連続座標を整数ビンラベルに離散化することで、キーポイント位置の推定を分類問題として解く手法である。精度の面でも、OpenPose を含む多くの既存手法よりも正確であることが報告された。RTMPose 単独ではキーポイントの二次元位置のみを推定する。そこで MocapNET[14]を使用した。MocapNET は、キーポイントの二次元位置を三次元位置に変換させる手法である。RTMPose と MocapNET を組み合わせることで、単眼で三次元位置を推定できる（以下 RTM1)。また、RTMPose を使ったステレオ手法も実装した（RTM2)。2 つ目は MediaPipe Pose である[15]。MediaPipe Pose はライブラリとして提供されており、実装が容易であるため、システムの保守や更新において大きな利点がある。さらに、GPU 不要で高速に動作するため、実用性の観点から選択された。MediaPipe Pose の利点は、作業療法分野での研究でも強調された[16]。MediaPipe Pose 単体で、キーポイントの三次元位置の推定も可能である。RTMPose と同様に、単眼手法とステレオ手法が実装された（以下 MP1、MP2)。

RTM は、133 個のキーポイント（体に 17 個、足に 6 個、顔に 68 個、両手に 21 個)の位置を推定する。姿勢推定モデルは、もっとも複雑なモデル（正確だが遅い）を選択した。我々の実装では、RTMPose は約 2.1FPS で実行した。MocapNET は、133 個の RTMPose のキーポイントを 212 個のキーポイント（体に 19 個、足に 40 個、顔に 99 個、両手に 27 個）に変換した。MocapNET を実行するのには、1 つの動画に対して数十秒程度しか、かからなかった。MediaPipe Pose は、543 個のキーポイント（体に 33 個、顔に 468 個、両手に 21 個）の位置を推定する。使用可能な 3 つのモデルのうち、もっとも複雑なモデルが使用された。我々の実装では、MediaPipe Pose は約 9.7FPS で実行した。

ビデオカメラのフレームレートは 30FPS で 100FPS の MoCap よりも低かったため、開始フレームと終了フレームを一致させた後、均等にフレームが配置しているという前提の下で、MoCap のフレーム数をビデオカメラのフレーム数と一致させた。単眼手法ではスケールが不定である。RTM1 と MP1 で推定された各実験参加者の右肘と肩の距離が、メジャーで測定された距離と一致するようにスケーリングした。肘と肩の距離は、これらのキーポイントの両方が同時に有効と判断された場合にのみ計算され、その平均値がスケーリングに使用された。各手法で推定されたキーポイントは異なる座標系を持っていたため、剛体変換[17]により一致させた。アフィン変換はキーポイントの不自然な変形を引き起こし、誤って高い精度が算出される可能性があったため利用しなかった。剛体変換は回転と移動のみから構成されるため、先にスケーリングした。剛体変換でも、変換に利用される点の数が少な過ぎる場合は、精度が不当に高くなる。剛体変換とスケーリングのパラメーターは、各参加者の 8 本の動画のすべての有効なキーポイントから計算された。有効なキーポイントとは、すべての手法で有効と判断されたキーポイント、つまり MoCap で検出され、各 HPE の手法で特定のしきい値以上の信頼値を持って推定されたキーポイントである。信頼値とは、モデル自身が推定精度の「自信」を表したような値であり、0〜1 である（1 に近いほど信頼できる)。この研究では、信頼値のしきい値は 0.3 に設定された。本実験では、71,777 点〜270,526 点が剛体変換に利用されていたため、精度が不当に高くなっていたことはなかったと言える。スケーリングと剛体変換は実験参加者ごとに実行された。すべてのコードは Python で実装された。

3 結果

4 つの HPE の手法が推定したキーポイントの三次元位置と、MoCap で計測された三次元位置のユークリッド距離を誤差として計算した。結果は表 1 の通りである。なお、両肩と顔の 3 点のキーポイントは、それぞれ

表 1. 平均ユークリッド距離（mm）。カッコ内は標準誤差。

	MP1	MP2	RTM1	RTM2
親指（右）	367.8（34.5）	43.6（3.5）	354.3（18.7）	38.7（3.2）
親指（左）	400.9（35.5）	59.6（11.7）	427.1（44.8）	56.7（10.6）
中指（右）	404.0（35.3）	44.4（9.2）	396.1（23.2）	47.6（8.3）
中指（左）	426.4（35.1）	51.1（10.4）	461.0（45.3）	56.3（10.5）
手首（右）	365.0（17.0）	48.7（2.9）	285.2（13.9）	45.3（2.4）
手首（左）	334.6（19.9）	44.8（2.7）	312.9（21.5）	43.5（2.7）
肘（右）	174.8（8.7）	58.7（2.0）	197.1（12.3）	59.2（2.0）
肘（左）	179.1（8.4）	54.9（2.0）	234.7（14.0）	54.7（1.8）
両肩	128.6（4.2）	48.3（3.3）	180.4（9.4）	53.0（3.3）
両頬＋顎	153.4（11.8）	39.6（1.7）	185.6（15.2）	42.6（2.3）
平均	293.5	49.4	303.4	49.8

1 つにまとめて誤差を計算した。表から、単眼手法の誤差がステレオ手法に比べて非常に大きいことがわかる。定性的に、単眼手法は奥行方向の推定精度が特に低いことがわかった。単眼カメラでは、奥行方向の情報を取得できないからだと考えられる。一方、ステレオ手法の誤差はどちらも平均 5cm 未満であり、MoCap を使用するコストを考慮すると、目的によっては MoCap の代替となり得ると考えられる。4 つの HPE の手法間で有意にユークリッド距離が異なるかどうかを確認するために、4 つの手法を独立変数（参加者間要因）とし、各キーポイントの平均ユークリッド距離を従属変数として、一元配置分散分析（ANOVA）を実施した。分散分析の結果、すべてのキーポイントで手法の種類が主効果を示した（$p < .001$）。Bonferroni t 検定（$p < .05$）では、すべてのキーポイントで 2 つのステレオ手法の誤差が、2 つの単眼手法よりも有意に小さいことが示された。一方、RTMPose と MediaPipe Pose の精度に大きな差はなかった。左中指と両肩では MP2 は RTM2 よりも有意に誤差が小さく、左肘と両肩では MP1 は RTM1 よりも有意に小さかった。右手首では RTM2 と RTM1 はそれぞれ、MP2 と MP1 よりも有意に誤差が小さかった。

表 1 からは、RTMPose と MediaPipe Pose は同程度の精度に見え、ステレオ手法は高精度に見えるが、精度の他に重要な評価指標がある。それはキーポイントの推定に失敗した割合（本実験では、信頼値が 0.3 未満の場合）である。表 1 は信頼値が 0.3 以上、つまり精度よく推定できる場合に限った比較になっている。RTM2 と MP2 で推定に失敗した割合を比較すると、MP2 の方が非常に大きかった。MP2 は、できる限り推定しようとするのではなく、困難な場合は推定を放棄している印象であった。これは、MediaPipe Pose が GPU なしでも高速に動作するための工夫かもしれない。一方、RTM2 では偽陽性が時々観察された（人がいない場所に顔を検出する、など）。ただし、本研究で使用した動画では確認されなかった。結論として、本研究結果の印象では RTM2 の方が優れている。ただし、RTM2 において推定に失敗する割合がどの程度なのか、話者のジェスチャーの速さや複雑さと関連付けて慎重に調査する必要があるだろう（本実験では RTM2 が推定に失敗したフレーム数は、問題になるほどではなかったと定性的には判断した）。

MoCap と RTM2 で計測したジェスチャー空間を可視化して比較し、HPE の実用性を評価した。図 5 は、ある実験参加者 1 名のすべてのデータを使用し作成された。空間を 1 辺 33.3mm のボクセルに分割し、各ボクセ

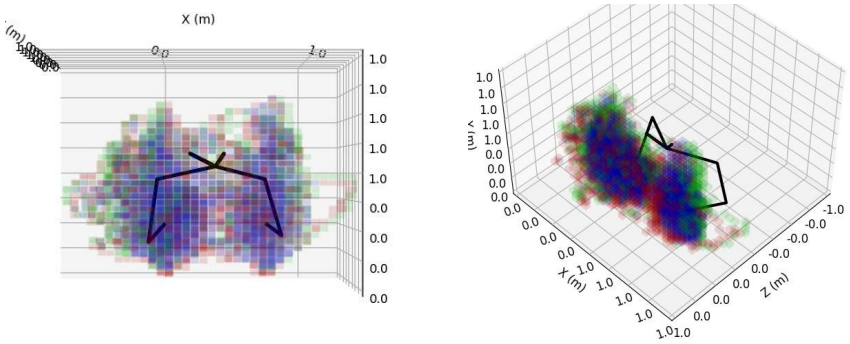

図 5. ジェスチャー空間の可視化例

ルに左手首または右手首のキーポイントが少なくとも 1 回含まれたかどうかを表している（つまり、手首の左右と時間情報は考慮されていない)。青色のボクセルは MoCap と RTM2 の両方を含むボクセル、緑色は MoCap のみ、赤色は RTM2 の手首のキーポイントのみを含んだボクセルである。また、MoCap によって測定された参加者のキーポイントの平均位置が実線でプロットされている。ボクセルサイズによって変化するが、図 5 より、ジェスチャー空間の概形は、おおむね MoCap と RTM2 で一致したと言える。

4 まとめ

本研究では、近年著しい進歩を遂げている深層学習に基づく HPE が、ジェスチャー研究において MoCap の代替となるかどうかを、精度の観点から調査した。実験の結果、単眼手法の精度は不十分だが、ステレオ手法は、mm 単位の非常に高い精度が求められる場合などを除いた特定の目的においては、MoCap の代替となり得ると考えられる。HPE の精度はまだ MoCap には及ばないが、深層学習は日々進歩しており、今後も改善が期待できる。また、専用のスペースも不要であるし、必要な機材は一般的なビデオカメラだけで良い（場合によっては高額な GPU が実行に必要になる場合もある）ため、MoCap に比べて安価である。再帰性反射マーカーの貼付が不要で、手や顔を含めて非常に多くのキーポイントを推定できる。

本研究では精度の観点からいくつかの HPE を比較したが、検証すべき点はまだある。例えば、服装の影響である。本研究ではなるべくオーバーサイズな服を着ないように参加者に伝えたが、そのような服装の場合、HPE の精度が落ちると予想される。また、話者とカメラとの距離の影響や、カメラの高さや解像度の影響、HPE の誤差の傾向（奥行はいつも実際よりも大きく（奥に）推定される傾向がある、など）なども検証する必要があるだろう。これらの検証を通じて本提案システムの実用性を向上させた後、システムを広く公開したいと考えている。

将来展望として、主に三つの方向で、本システムを使用した研究の発展や応用的利用を計画している。第一に、新しいジェスチャー空間を用いて、これまで二次元平面で行われてきたジェスチャー空間の研究を再検討する。例えば、ジェスチャーの文化比較や発達的変化、個人差などに関する研究を新しいシステムとともに追試する。第二に、下半身に注目したコミュニケーション研究への拡張である。提案した三次元計測手法は、手や腕に限らず、頭や下半身(膝やつま先など)のキーポイント位置も推定可能である。上半身と比較してあまり注目されてこなかった、下肢の位置や動きについても研究を進めることで、コミュニケーションの土台となる身体活動がより明らかになるだろう。第三は、本システムの他領域への応用的利用の検討である。本研究成果は身体活動が深く関わる分野、例えば、スポーツや演劇、教育現場、遠隔通話など、広く適用できるため、それぞれの分野での応用を検討する。

【参考文献】

[1] McNeill, D. (1992). Hand and mind: What gestures reveal about thought. Chicago: Chicago University Press.

[2] Tellier, M., Stam, G., & Ghio, A. (2021). Handling language: How future language teachers adapt their gestures to their interlocutor. Gesture, 20(1), 30-62.

[3] Trujillo, J. P., Vaitonyte, J., Simanova, I., & Özyürek, A. (2019). Toward the markerless and automatic analysis of kinematic features: A toolkit for gesture and movement research. Behavior Research Methods, 51, 769-777.

[4] Kipp, M., von Hollen, L., Hrstka, M.C., Zamponi, F. (2014). Single-person and multi-party 3D visualizations for nonverbal communication analysis. International Conference on Language Resources and Evaluation.

[5] Prové, V. & Oben, B. (2021). Social attraction and the adaptive use of hand gestures in native-non-native dyadic interactions. International Conference of Asia-Pacific LSP & Professional Communication Association.

[6] Özyürek, A. (2002). Do speakers design their cospeech gestures for their addressees? The effects of addressee location on representational gestures. Journal of Memory and Language, 46(4), 688-704.
[7] Núñez, R. E., Motz, B. A., & Teuscher, U. (2006). Time after time: The psychological reality of the ego-and time-reference-point distinction in metaphorical construals of time. Metaphor and Symbol, 21(3), 133-146.
[8] Priesters, M. A., & Mittelberg, I. (2013). Individual differences in speakers' gesture spaces: Multi-angle views from a motion-capture study. Tilburg Gesture Research Meeting, 19-21.
[9] Toshev, A., & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. Computer Vision and Pattern Recognition, 1653-1660.
[10] Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. Computer Vision and Pattern Recognition, 7291-7299.
[11] Zhang, Z. (2000). A flexible new technique for camera calibration. Pattern Analysis and Machine Intelligence, 22(11), 1330-1334.
[12] Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., & Chen, K. (2023). RTMPose: Real-time multi-person pose estimation based on mmpose. arXiv Preprint, arXiv:2303.07399.
[13] Li, Y., Yang, S., Liu, P., Zhang, S., Wang, Y., Wang, Z., Yang, W., & Xia, S. T. (2022). SimCC: A simple coordinate classification perspective for human pose estimation. European Conference on Computer Vision, 89-106.
[14] Qammaz, A., & Argyros, A. A. (2019). MocapNET: Ensemble of SNN encoders for 3D human pose estimation in RGB images. British Machine Vision Conference.
[15] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., Chang, W.T., Hua, W., Georg, M., & Grundmann, M. (2019). Mediapipe: A framework for building perception pipelines. arXiv preprint, arXiv:1906.08172.
[16] Ienaga, N., Takahata, S., Terayama, K., Enomoto, D., Ishihara, H., Noda, H., Hagihara, H. (2022). Development and verification of postural control assessment using deep-learning-based pose estimators: Towards clinical applications. Occupational Therapy International.
[17] Arun, K.S., Huang, T.S., Blostein, S.D. (1987). Least-squares fitting of two 3-d point sets. Pattern Analysis and Machine Intelligence, 698-700.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
身振りの三次元計測に向けた簡便な手法の比較検討	電子情報通信学会総合大会	2024 年 3 月
Comparison of deep learning-based three-dimensional human pose estimation methods with motion capture for gesture research	Computer Vision and Image Understanding	査読中（2024 年 6 月現在）