

修士論文概要書

Master's Thesis Summary

Date of submission: 07/21/2025 (MM/DD/YYYY)

専攻名（専門分野） Department	情報理工・ 情報通信専攻	氏 名 Name	中島 聖	指 導 教 員 Advisor	渡辺 裕 印 Seal
研究指導名 Research guidance	オーディオビジュ アル情報処理研究	学籍番号 Student ID number	CD 5122F067-3-		
研究題目 Title	エゴセントリックな未来の姿勢予測手法の研究 Research on Egocentric Pose Forecasting Methods				

1. まえがき

近年、仮想現実(VR)や拡張現実(AR), ウェアラブルデバイスなどの技術が進化し, 人間の動作理解や姿勢推定の需要が高まっている. 中でも, 一人称視点(egocentric view)による映像を活用した姿勢推定は, 装着型センサを用いずに自然な形で動作を取得できる手法として注目されている.

従来の姿勢推定・予測手法の多くは, カメラ外部から取得した人物の全身画像や関節の時系列データをもとに処理を行うが, 一人称視点の場合, 撮影者自身の身体が映らないという特性上, 姿勢の再構成は極めて困難である. 一方で, こうした技術が実現すれば, スポーツトレーニングやリハビリ, 街中でのナビゲーション, 没入型ゲーム体験など幅広い応用が期待される.

本研究では, 一人称視点映像から未来の視覚コンテキストを生成し, それを活用して将来の姿勢を予測する新しい手法を提案する. 物理シミュレーションに基づいた予測モデルと, 映像生成技術とを統合することで, より自然で整合性のある未来動作予測を目指す.

2. 関連研究

一人称視点からの姿勢推定・予測に関する研究は近年注目されており, 身体装着型センサや特殊なカメラを用いない自然なインタラクションの実現に向けた技術的挑戦が続けられている.

従来の手法の多くは, 下向き魚眼カメラ映像や IMU (慣性センサ) を入力として用いることで, エゴセントリックなポーズ推定や予測を行っていた. しかし, これらはカメラやセンサの装着が必要であることから, 使用者の動作を妨げる, あるいは高コストとなるなど, 応用範囲の制限につながっていた. また, 物理シミュレーションと連携しないこれらの手法では, 足の滑りや地面へのめり込みといった物理的に不自然な動作が出力されることも少なくない.

そのような課題を踏まえ, 単眼カメラ映像のみを用い, かつ物理的整合性を重視した手法として「KinPoly」が提案されている[1]. KinPoly は, 視覚特徴と物理シミュレーションを統合し, 自然で整合的な3Dポーズ推定を実現する一方, 将来の動作予測(forecasting)には対応していない.

一方, エゴセントリックな姿勢予測(forecasting)手法は非常に限られており, 既存の代表的な研究としては

EgoCast[2]および EgoPose[3]が挙げられる. EgoCast は, IMUを必要とするキネマティクスベースの構成であり, センサなしでの完全なエゴセントリック映像ベースとは言い難い. EgoPose は, 深層強化学習を用いて物理的整合性を確保しながら姿勢予測を行う点で本研究と近いが, 視覚的未来情報を活用しておらず, オブジェクトとのインタラクションにも対応していない.

さらに, 近年では拡散モデル(diffusion model)を用いた高精度な未来フレーム予測手法として「SyncVP」[4]も提案されており, 映像ベースでの未来コンテキスト予測の可能性が広がっている.

本研究では, KinPoly の物理整合性重視の構造と, SyncVP の未来視覚予測能力を統合することで, 単眼カメラのみを用いながらも視覚的・物理的に整合性のある未来姿勢予測を実現することを目的とする.

3. 提案手法

本研究では, 単眼カメラ映像のみで動くダイナミクスベースかつオブジェクトとの相互作用を実現した新しい Egocentric Pose Forecasting 手法を提案する. 従来手法のダイナミクスベースのエゴセントリックな姿勢推定をベースに, これまでの姿勢予測で用いられる姿勢推定結果を用いるアプローチから, 現在までの映像から未来フレームを生成し, 姿勢予測をするというアプローチを取って, 従来では出来なかったオブジェクトとの相互作用を考慮したエゴセントリックな姿勢予測手法を提案する.

4. 実験

4.1 環境

物理シミュレーションの実行環境として, MuJoCo (Multi-Joint dynamics with Contact) [5]を採用している.

4.2 データセット

本研究では, Kinpoly で提供されているデータセットを用いて実験を行った. このデータセットは一人称視点映像(ビデオシーケンス)と, それと同期したモーションキャプチャ(Mocap)データからなる.

三名の異なる被験者から, 5種類のアクションを含む合計 266 のシーケンスからなる.

4.3 評価実験

評価指標としては、姿勢推定でよく用いられる MPJPE (Mean Per Joint Position Error) という、各関節の 3D 位置の平均誤差を示す指標を用いる。表 1 から、提案手法は EgoPose に対して約 29%の精度向上を達成していることがわかる。

図 1 に EgoPose と提案手法の姿勢予測結果の可視化結果を示す。左の EgoPose のポーズよりも右の提案手法のポーズの方がオブジェクトとのインタラクションが適切に行えていることがわかる。

図 2 は、GT (ground truth)と提案手法の可視化結果を示す。青色のヒューマノイドが GT を表し、赤色のヒューマノイドが提案手法の測定結果を表す。

表1 EgoPose と提案手法の評価指標

評価指標	MPJPE[mm]↓
EgoPose	192.27
提案手法	135.85

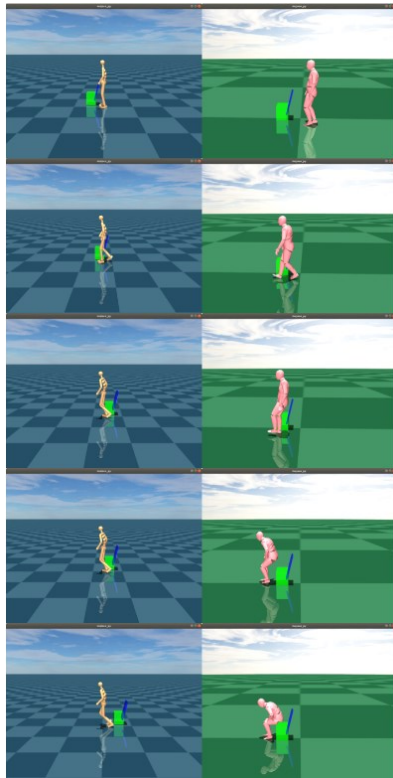


図 1 EgoPose と提案手法の可視化

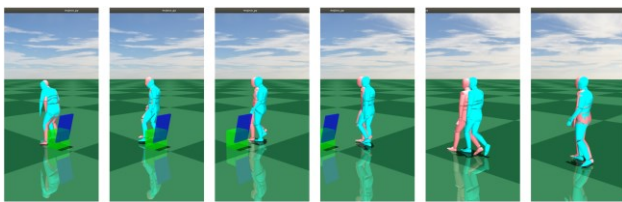


図 2 GT と提案手法の姿勢予測結果の可視化

5. 考察と今後の課題

定量的評価において提案手法は、MPJPE で約

29%の誤差低減(EgoPose: 192.27mm → 提案手法: 135.85mm)を達成しており、姿勢予測精度が大きく向上していることが確認された。この改善は、SyncVP による未来フレームの生成を通じて、視覚的な未来コンテキストを明示的に活用した点に起因していると考えられる。従来の手法では考慮されなかった未来映像中のオブジェクトや環境要素が、身体の運動に与える影響を反映できたことが、誤差低減に貢献したと推察される。

また、可視化結果からも提案手法は自然なオブジェクトとの相互作用を実現している。加えて、従来手法の問題点の1つである絶対位置のエラーについても、オブジェクトとの相互作用を考慮することで改善されているように見える。

提案手法ではフレーム生成の精度に依存する。また、SyncVP によって生成される未来フレームは視覚的に妥当であるが、特に物理的なインタラクションを考慮していない。解決策として、生成フレームを使った姿勢予測結果を生成フレームの品質改善にフィードバックするアプローチが考えられる。物理シミュレーションと未来フレーム生成の協調処理は未開拓領域であり、今後の研究価値がある。

6. むすび

本研究では、一人称視点の単眼カメラ映像から未来の身体姿勢を、物理的整合性を保ちながら予測する新しい手法を提案した。拡散モデルによる未来視覚情報の生成と、物理シミュレーションを用いた姿勢推定の統合により、従来手法では困難であったオブジェクトとの相互作用や自然な未来動作の予測を実現した。

参考文献

- [1] Yuan, Ye and Kitani, Kris. “Dynamics-Regulated Kinematic Policy for Egocentric Pose Estimation”, Advances in Neural Information Processing Systems, 2020
- [2] Escobar, Maria and Puentes, Juanita and Forigua, Cristhian and Pont-Tuset, Jordi and Maninis, Kevis-Kokitsi and Arbeláez, Pablo, “EgoCast: Forecasting Egocentric Human Pose in the Wild”, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2025
- [3] Yuan, Ye and Kitani, Kris. “Ego-Pose Estimation and Forecasting as Real-Time PD Control”, Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019, pp. 10082-10092
- [4] Pallotta, Enrico and Azar, Sina Mokhtarzadeh and Li, Shuai and Zatsarynna, Olga and Gall, Juergen, “SyncVP: Joint Diffusion for Synchronous Multi-Modal Video Prediction”, Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, June, pp. 13787-13797
- [5] Todorov, Emanuel and Erez, Tom and Tassa, Yuval, “MuJoCo: A physics engine for model-based control”, 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012, pp.5026-5033

2025 年度

早稲田大学大学院基幹理工学研究科情報理工・情報通信専攻修士論文

エゴセントリックな未来の姿勢予測手法の研究
Research on Egocentric Pose Forecasting Methods

中島 聖

(5122F067-3)

提出日：2025.7.21

指導教員：渡辺 裕 教授

研究指導名：オーディオビジュアル情報処理研究

目次

目次	ii
第 1 章 序論.....	1
1.1 研究背景	1
1.2 研究目的	1
1.3 本論文の構成	2
第 2 章 関連技術.....	3
2.1 まえがき	3
2.2 Dynamics-Regulated Kinematic Policy for Egocentric Pose Estimation	3
2.3 SyncVP: Joint Diffusion for Synchronous Multi-Modal Video Prediction.....	4
2.4 むすび.....	4
第 3 章 提案手法.....	5
3.1 まえがき	5
3.2 提案手法	5
3.3 むすび.....	6
第 4 章 実験.....	7
4.1 まえがき	7
4.2 環境	7
4.3 データセット	7
4.4 評価実験	7
4.4.1 評価指標.....	7
4.4.2 実験の流れ.....	8
4.4.3 結果.....	8
4.5 考察	10
4.6 むすび.....	10
第 5 章 結論と今後の課題.....	11
5.1 結論	11
5.2 今後の課題.....	11
謝辞	13
参考文献	14
図一覧.....	15
表一覧.....	16
研究業績	17

第1章 序論

1.1 研究背景

近年、仮想現実（VR）や拡張現実（AR）、ウェアラブルデバイスの発展に伴い、一人称視点（egocentric view）から得られる映像を活用した身体動作理解の重要性が高まっている。中でも、カメラ装着者自身の身体が映像内に映らない状況において、動作や姿勢を推定・予測する技術は非常に挑戦的である一方、多くの応用可能性を秘めている。

この種の手法は、一人称視点映像のみを入力として用いるため、従来のボディトラッキングに必要とされていた外部センサやマーカなどの装着型デバイスを不要とし、低コストかつ簡便に姿勢推定・予測を行うことができる。その結果、運動やリハビリテーションの妨げにならず、スポーツや医療分野における活用にも適している。また、街中においては、将来の姿勢を予測することで事故の予防や適切なナビゲーション支援が可能となり、ゲーム領域では、ボディトラッキングなしに没入感や応答性の高い操作を実現できるといった利点がある。

このような背景から、近年では一人称視点に特化した姿勢推定・予測手法の研究が進展している。たとえば、Ye Yuan らによる Ego-Pose Estimation and Forecasting as Real-Time PD Control (EgoPose)では、深層強化学習と PD 制御を組み合わせることで、エゴセントリック映像から物理的に整合性のある姿勢を推定・予測する手法が提案されている[1]。

また、従来の姿勢予測手法は、現在までの関節角度や速度などの時系列情報のみをもとに将来の姿勢を予測しており、視覚情報がどのように人の行動に影響を与えるかという点には十分に対応できていない。現実の人間の動作は、関節の状態だけでなく視界に映る環境要素にも強く依存する。たとえば、信号機が赤に変われば歩行を停止するなど、視覚に基づく意思決定が動作に直結する場面は数多い。

このように、一人称視点を用いた姿勢推定・予測は、VR/AR、医療、スポーツ、日常生活といった幅広い分野において非常に有望な技術であると同時に、視覚的コンテキストを含めた将来予測の実現や、物理的に妥当なポーズ生成の観点から、いまだ多くの技術的課題を抱える未成熟な研究分野でもある。

1.2 研究目的

本研究の目的は、一人称視点映像から、将来の身体姿勢を予測する新たな手法を提案することである。特に、映像から得られる視覚的コンテキストと物理シミュレーションに基づいた予測を融合させることで、直感的かつ物理的に妥当な未来動作の生成を目指す。

エゴセントリックな姿勢推定手法の多くは、下向き魚眼カメラによるポーズ推定[3, 4]や、IMU を活用したマルチモーダルなポーズ推定[5]など、単眼前面カメラ映像のみからポーズを推定する手法は数少ない[1,2]。姿勢予測に関しては、そもそも数が少なく2つの手法のみ[1,6]。中でも、単眼前面カメラ映像のみを入力とする手法は現在1つ[1]。

既存の姿勢予測手法は、過去の関節の時系列情報のみを用いて未来の姿勢を推定する設計となっており、映像に含まれる視覚的情報が予測に十分活用されていない。これにより、視界の中にある物体や環境との関係を踏まえた行動選択が再現できず、人間の自然な動きとの乖離が生じる可能性がある。

また、物理的整合性の観点では、Dynamics-Regulated Kinematic Policy for Egocentric Pose Estimation (Kinpoly) のような物理シミュレータ内での姿勢再現技術が注目されているが、現時点では将来の動作予測 (forecasting) を扱っておらず、現在のポーズ推定 (estimation) に限定されている [2]。

そこで本研究では、Kinpoly を基盤とし、将来フレームの映像を生成し、それを視覚コンテキストとして活用することで、未来の姿勢を予測する機構を統合する。これにより、視覚的な未来予測と物理的制約の双方を考慮した、新たなエゴセントリック姿勢予測モデルを実現することが本研究の主たる目的である。

本手法により、スポーツ、医療、ナビゲーション、VR/AR などの分野において、より応答性が高く、自然で安定した身体動作の予測・制御が可能になることが期待される。

1.3 本論文の構成

本論文の構成を以下に示す。

第 1 章は本章であり、本研究の背景、目的について述べる。

第 2 章では、本研究に関連する技術について述べる。

第 3 章では、本研究で提案する手法について述べる。

第 4 章では、提案手法を用いた実験、結果及び考察について述べる。

第 5 章では、本論文の結論と今後の課題について述べる。

第2章 関連技術

2.1 まえがき

本章では，本研究で用いる一人称視点画像を用いた姿勢推定手法の「Dynamics-Regulated Kinematic Policy for Egocentric Pose Estimation」[2]，未来フレーム予測手法の「SyncVP: Joint Diffusion for Synchronous Multi-Modal Video Prediction」[7]，従来の Egocentric Pose Estimaion 手法について説明する．

2.2 Dynamics-Regulated Kinematic Policy for Egocentric Pose Estimation

「Dynamics-Regulated Kinematic Policy for Egocentric Pose Estimation」は，Zhengyi Luo らによって提案された一人称視点（egocentric）の 3D ポーズ推定手法であり，キネマティックスペースのポーズ予測と，物理ベースのダイナミクスシミュレーションを統合したフレームワークである．この手法は，カメラ装着者の全身ポーズを，一人称視点の映像のみから物理的整合性を保ちながら高精度に推定することを目的としている．この手法の概略を図 2.1 に示す．

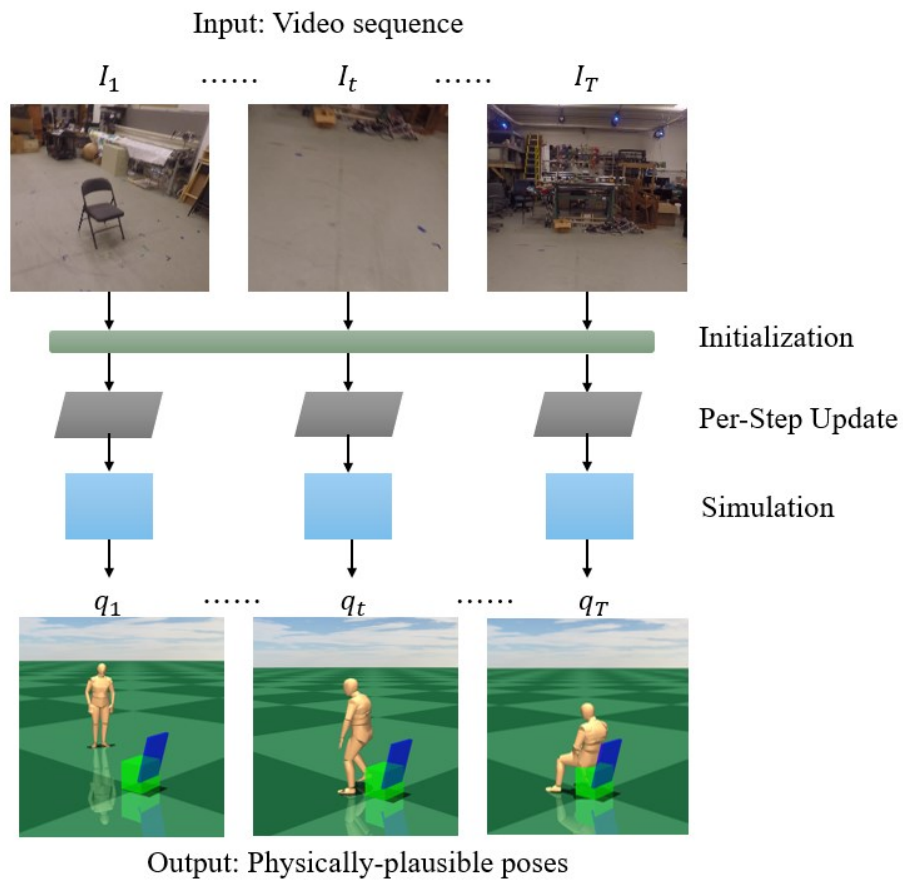


図 2.1 一人称視点画像を用いた姿勢推定の概略

Kinpoly では、図 2.1 のように、一人称視点画像の連続であるビデオシーケンスを入力として、主に Initialization, Simulation の 2 つの層を通して、物理的に妥当な姿勢を出力として得る。Initialization では、人間の初期ポーズと、オブジェクトの初期状態、画像特徴量を抽出する。Simulation では、抽出した画像特徴量を入力として、時間ステップごとに、物理シミュレーションを繰り返し、最終的に物理的に妥当な姿勢を出力として得る。

キネマティクスとダイナミクスの両方を活用して、自然かつ整合的な身体動作を再構成できる。ダイナミクスとオブジェクトの状態を組み込むことで、オブジェクトの相互作用も考慮できる。

2.3 SyncVP: Joint Diffusion for Synchronous Multi-Modal Video Prediction

SyncVP: Joint Diffusion for Synchronous Multi-Modal Video Prediction (SyncVP) は、画像・映像生成分野で注目されている拡散モデル (diffusion model) に基づく、マルチモーダル同期ビデオ予測手法である。従来のビデオ予測モデルと異なり、SyncVP は映像だけでなく、自然言語や音声といった他のモダリティと時間的に整合性のあるフレーム予測を行うことが可能であり、各モダリティの潜在情報を効果的に統合することに特徴がある。RGB 画像のみ (単独モダリティ) においても、高い性能を達成している。

2.4 むすび

本章では、本研究に関連する姿勢推定手法の Kinpoly、および未来のフレーム生成手法である SyncVP、従来の Egocentric Pose Estimation and Forecasting について述べた。

第3章 提案手法

3.1 まえがき

本章では、本研究で新たに提案する一人称視点映像に基づく未来ポーズ予測手法について説明する。本手法は、KinPoly の「物理整合性を重視したポーズ推定構造」と、SyncVP の「映像ベース未来予測能力」を組み合わせることで、未来の視覚コンテキストを活用した自然で安定した身体動作の予測を実現するものである。

従来の Egocentric Pose Forecasting 手法では、視覚的な未来情報を用いないため、物理整合性や、画像内のオブジェクトの影響を考慮できないという問題があった。本研究では、未来のフレーム生成を用いて、これらの課題の解決のアプローチを示す。

3.2 提案手法

本研究では、単眼カメラ映像のみで動くダイナミクスベースかつオブジェクトとの相互作用を実現した新しい Egocentric Pose Forecasting 手法を提案する。従来手法のダイナミクスベースのエゴセントリックな姿勢推定をベースに、これまでの姿勢予測で用いられる姿勢推定結果を用いるアプローチから、現在までの映像から未来フレームを生成し、姿勢予測をするというアプローチを取って、従来では出来なかったオブジェクトとの相互作用を考慮したエゴセントリックな姿勢予測手法を提案する。図 3.1 に提案手法の概略を示す。

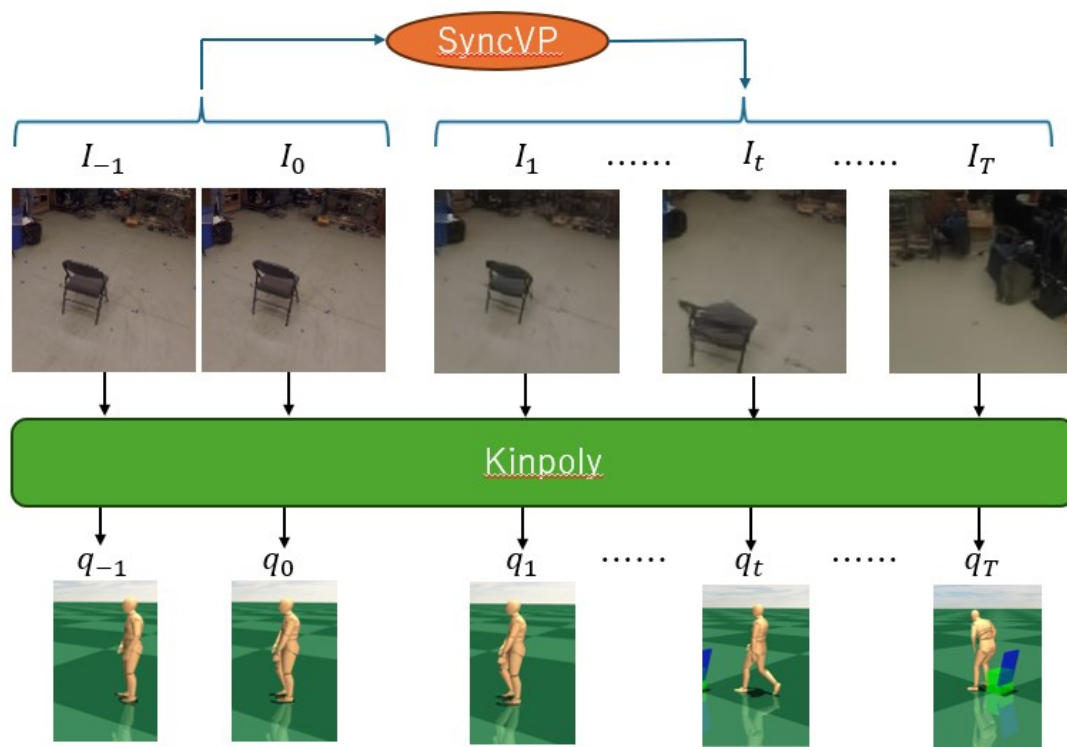


図 3.1 提案手法の概略

図 3.1 は現在までの視覚映像から SyncVP を用いて、未来フレーム(I_t)を得る。得られた未来フレームと現在までのフレームから、Kinpoly のネットワークを通して、姿勢推定する。このとき、得られた未来フレームからの姿勢推定結果は、未来の姿勢の予測結果となっている。このようにして、単眼カメラ映像のみで動くダイナミクスベースかつオブジェクトとの相互作用を実現した新しい Egocentric Pose Forecasting 手法を実現する。

3.3 むすび

本章では、姿勢推定結果を用いて姿勢予測を行う従来のアプローチから、未来フレーム生成を用いる新たなアプローチを提案し、単眼カメラ映像のみで動くダイナミクスベースかつオブジェクトとの相互作用を実現した新しい Egocentric Pose Forecasting 手法について述べた。

第4章 実験

4.1 まえがき

本章では、提案手法の有効性を検証するために実施した実験の概要、結果、および考察について述べる。

まず、評価環境および使用したデータセットの詳細について説明した後、従来手法との比較とともに、提案手法の性能を定量的に評価する。加えて、定性的な結果も示し、視覚的に生成される未来ポーズの自然さや妥当性についても考察する。

4.2 環境

本研究では、物理シミュレーションの実行環境として、MuJoCo (Multi-Joint dynamics with Contact) [8]を採用している。MuJoCo は高精度かつ高速な物理シミュレーションが可能であり、複雑な関節構造や接触計算をリアルタイムで処理できる点から、EgoPose や KinPoly といった先行研究でも広く使用されている。

4.3 データセット

本研究では、Kinpoly で提供されているデータセットを用いて実験を行った。このデータセットは一人称視点映像(ビデオシーケンス)と、それと同期したモーションキャプチャ(Mocap)データからなる。

このデータセットは、モーションキャプチャスタジオ内で収集され、3名の異なる被験者によって撮影された。被験者の頭部にはGoProカメラが装着され、一人称視点の映像が取得された。また、被験者自身とオブジェクトの双方にモーションキャプチャ用のマークが取り付けられ、3D全身ポーズおよび6DoF(3次元位置+3次元回転)によるオブジェクトポーズのグラウンドトゥルースが得られている。

データセットには椅子に座る／立ち上がる、障害物の回避、箱の上に登る、箱を押す、一般的な歩行(歩く、走る、かがむ)の5種類のアクションが含まれる。合計266のシーケンス(約14万8千フレーム)が含まれており、各アクションには約50のシーケンスが存在する。

4.4 評価実験

本章では、提案手法の有効性を評価するために行った実験の設定および方法について説明する。

4.4.1 評価指標

以下の評価指標を用いて、各手法の性能を測定した。

- MPJPE (Mean Per Joint Position Error) [mm]: 各関節の3D位置誤差の平均。姿勢

予測の精度を測定する代表的指標.

4.4.2 実験の流れ

本節では, 提案手法に基づくエゴセントリックな姿勢予測の実験プロセスについて説明する.

- データ準備

Kinpoly で提供される学習用・評価用の MoCap データセットを用いて, ビデオシーケンスに対応 GT ポーズデータを取得する. 各シーケンスは, 一人称視点映像 (RGB フレーム列) と, 対応する全身のポーズ系列から構成させる.

- 未来フレーム生成 (SyncVP)

現在までの 2 フレームを入力として, SyncVP を用いて将来の視覚フレーム (次の 90 フレーム) を生成する.

- 姿勢予測

生成された未来フレームと, これまでの視覚フレームの組を KinPoly モデルの入力とし, 未来の姿勢を推定する.

- 評価指標の算出

得られたポーズと GT を比較し, MPJPE を算出する.

- 従来手法との比較

提案手法の有効性を明らかにするために, EgoPose の姿勢予測と比較実験を実施した.

4.4.3 結果

本節では, 提案手法および比較手法について, 評価指標に基づく定量的な実験結果を示す. 評価には, KinPoly データセットのテストシーケンス (全 51 シーケンス) を使用した. EgoPose と提案手法の MPJPE を表 4.1 に示す.

表 4.1 EgoPose と提案手法の評価指標

評価指標	MPJPE[mm] ↓
EgoPose	192.27
提案手法	135.85

この結果から, 提案手法は EgoPose に対して約 29%の精度向上を達成しており, 未来フレームを活用した視覚的コンテキストの導入が, 姿勢予測精度の改善に有効であることが示された.

次に各手法における姿勢予測データを可視化した結果を図 4.1 に示す.

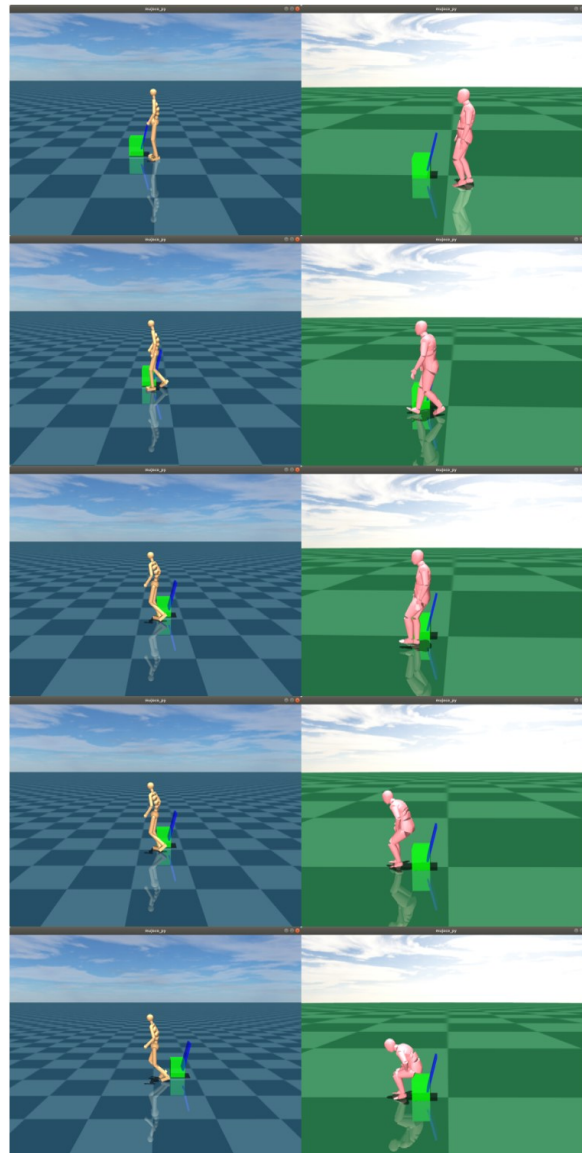


図 4.1 EgoPose と提案手法の姿勢予測結果の可視化

図 4.1 において，左の画像列が EgoPose で右が提案手法の姿勢予測を可視化したものである．左の画像列では橙色のヒューマノイドが姿勢予測結果を表し，右の画像列では赤色のヒューマノイドが姿勢予測結果を表している．図 4.1 からわかる通り，従来手法ではオブジェクトとの相互作用を考慮できない．提案手法では，自然にオブジェクトとの相互作用が実現できていることがわかる．

次に，GT と提案手法の姿勢予測結果の比較を図 4.2 に示す．

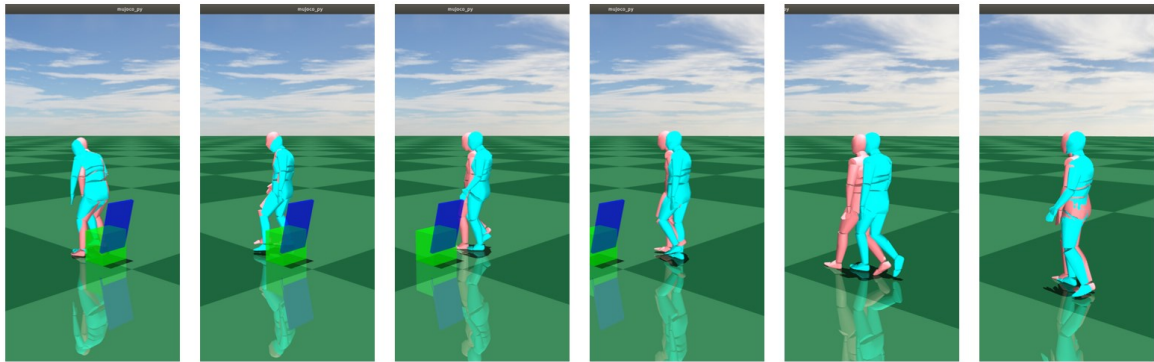


図 4.2 GT と提案手法の姿勢予測結果の可視化比較

図 4.2 の青色のヒューマノイドが GT を表し，赤色のヒューマノイドが提案手法の予測結果を表す。

4.5 考察

定量的評価において提案手法は，MPJPE で約 29%の誤差低減（EgoPose: 192.27mm → 提案手法: 135.85mm）を達成しており，姿勢予測精度が大きく向上していることが確認された。この改善は，SyncVP による未来フレームの生成を通じて，視覚的な未来コンテキストを明示的に活用した点に起因していると考えられる。従来の手法では考慮されなかった未来映像中のオブジェクトや環境要素が，身体の運動に与える影響を反映できたことが，誤差低減に貢献したと推察される。

また，可視化結果からも提案手法は自然なオブジェクトとの相互作用を実現している。加えて，従来手法の問題点の 1 つである絶対位置のエラーについても，オブジェクトとの相互作用を考慮することで改善されているように見える。

一方で，本手法ではいくつかの問題も存在する。まず，姿勢予測に未来フレーム生成を用いるため，その精度に依存する。

さらに，計算コストの面でも，映像生成と物理シミュレーションを組み合わせる本手法は，従来手法よりも推論時間が長くなる傾向にある。リアルタイム応用を視野に入れる場合は，今後さらなる高速化や軽量化が求められるだろう。

4.6 むすび

本章では，提案手法の定量的および定量的な評価実験を行った。定量的評価の結果から，提案手法は従来手法（EgoPose）と比較して MPJPE の指標において大幅な誤差低減を達成し，姿勢予測精度の向上が確認された。また，定量的評価においても，従来手法との比較では，オブジェクトとの自然な相互作用を実現し，GT との比較では，予測結果の妥当性を確認できた。

第5章 結論と今後の課題

5.1 結論

本研究では、一人称視点映像のみを入力として未来の身体姿勢を、物理的整合性を保ちながら予測する新たな Egocentric Pose Forecasting 手法を提案した。提案手法は、物理ベースの姿勢推定手法である KinPoly を基盤とし、映像から将来の視覚コンテキストを生成する SyncVP を統合することで、映像に映るオブジェクトとのインタラクションを考慮した自然な未来姿勢の生成を実現した。

実験では、KinPoly データセットを用いて提案手法の有効性を評価し、従来手法(EgoPose)と比較して MPJPE の指標において高い精度を達成した。さらに、定性的な可視化結果からも、物理的に一貫性のある姿勢を自然に予測できることが確認された。

本研究は、視覚情報と物理的制約を融合した新たな姿勢予測アーキテクチャの提案とその有効性の実証という観点において、Egocentric Pose Forecasting 分野に貢献するものである。

5.2 今後の課題

本研究では、未来フレーム生成と物理ベースの姿勢予測を組み合わせた新たな Egocentric Pose Forecasting 手法を提案したが、今後の発展に向けて以下の課題が残されている。

- 未来フレーム生成の品質向上

SyncVP によって生成される未来フレームは視覚的に妥当であるが、特にオブジェクトの関係において、現実の視覚情報と乖離が生じることがある。

- マルチモーダル情報の活用

本研究では RGB 映像のみに基づく予測を行ったが、音声やテキスト、加速度センサなどの他のモダリティと統合することで、より正確で意味のある行動予測が可能になると期待される。

- リアルタイム性

映像生成を追加していることによる計算コストの増加により、推論時間が長くなる傾向にある。

一つ目の課題においては、生成フレーム自体が物理的な制約を受けた上で生成される機構の実現が考えられる。具体的には、生成フレームを使った姿勢予測結果を生成フレームの品質改善にフィードバックされるシステムで、物理的な制約の影響を受けたフレーム生成が可能。物理エンジンからのフィードバックを受けるフレーム生成は、現在存在せず、新規性がある。

二つ目の課題においては、IMU からの頭部の情報を活用した手法 (EgoCast [6]) があるが、ダイナミクススペースの姿勢予測手法に関しては存在しない。人の行動は視覚情報だけでなく、身体感覚からの影響もある。例えば、背中にボールが当たれば振り向くなど。オブジェクトと身体感覚も考慮できれば、より精度の高い予測が可能。

三つ目の課題に関しては, 未来フレーム生成を利用した手法は, 本研究が初めてのため, 内外の研究機関では, まだ未着手である.

謝辞

本研究の実験環境を整えてくださり，そして，丁寧なご指導をいただいた渡辺裕教授に心より感謝申し上げます。

また，日頃より貴重な意見や，相談に乗ってくださった研究室の皆さんに，深く感謝申し上げます。

最後に，生活面や学費など，すべての面で支えてくれた家族に感謝致します。

参考文献

- [1] Yuan, Ye and Kitani, Kris. “Ego-Pose Estimation and Forecasting as Real-Time PD Control”, Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019, pp. 10082-10092
- [2] Yuan, Ye and Kitani, Kris. “Dynamics-Regulated Kinematic Policy for Egocentric Pose Estimation”, Advances in Neural Information Processing Systems, 2020
- [3] Tome, Denis and Peluse, Patrick and Agapito, Lourdes and Badino, Hernan, “xR-EgoPose: Egocentric 3D Human Pose from an HMD Camera”, Proceedings of the IEEE International Conference on Computer Vision, 2019, p. 7728-7738
- [4] Wang, Jian and Luvizon, Diogo and Xu, Weipeng and Liu, Lingjie and Sarkar, Kripasindhu and Theobalt, Christian, “Scene-aware Egocentric 3D Human Pose Estimation”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13031-13040
- [5] Jiang, Jiaxi and Streli, Paul and Meier, Manuel and Holz, Christian, “EgoPoser: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere”, European Conference on Computer Vision, 2022, Springer
- [6] Escobar, Maria and Puentes, Juanita and Forigua, Cristhian and Pont-Tuset, Jordi and Maninis, Kevis-Kokitsi and Arbeláez, Pablo, “EgoCast: Forecasting Egocentric Human Pose in the Wild”, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2025
- [7] Pallotta, Enrico and Azar, Sina Mokhtarzadeh and Li, Shuai and Zatsarynna, Olga and Gall, Juergen, “SyncVP: Joint Diffusion for Synchronous Multi-Modal Video Prediction”, Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, June, pp. 13787-13797
- [8] Todorov, Emanuel and Erez, Tom and Tassa, Yuval, “MuJoCo: A physics engine for model-based control”, 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012, pp.5026-5033

図一覧

図 2.1 一人称視点画像を用いた姿勢推定の概略.....	3
図 3.1 提案手法の概略.....	6
図 4.1 EgoPose と提案手法の姿勢予測結果の可視化.....	9
図 4.2 GT と提案手法の姿勢予測結果の可視化比較.....	10

表一覧

表 4.2 EgoPose と提案手法の評価指標	8
--------------------------------	---

研究業績

- [1] 福田大翔, 中島聖, 渡辺裕, “骨格絶対長を考慮した三次元姿勢推定手法の検討”, 2023 年電子情報通信学会総合大会, D-12-67, pp.103, Mar. 2023.
- [2] 中島聖, 福田大翔, 渡辺裕“骨格の左右対称性を考慮した , 3 次元姿勢推定法の検討”, 情報処理学会第 120 回オーディオビジュアル複合情報処理研究会, Vol.2023-AVM-120, No.4, pp.1-2, Mar. 2023
- [3] S. Nakashima and H. Watanabe, “Improving Accuracy of Position Information in Egocentric Pose Estimation Methods”, IEEE 11th Global Conference on Consumer Electronics(GCCE2022), pp. 683-687, Oct. 2022