卒 業 論 文 概 要 書

Summary of Bachelor's Thesis

Date of submission: 02/01/2022

学科名 Department	情報通信	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 1W182238-6		
研究題目 Title	一人称視点画像を用いた姿勢推定手法における位置情報の精度改善 Improving the Accuracy of Position Information in Egocentric Pose Estimation Methods				

1. まえがき

近年，医療や観光，スポーツ，ゲームなど，さまざまな分野で Virtual Reality (VR) が利用されている．VR の中でも特に，Head Mounted Display (HMD) の普及が目覚ましい[1]．HMD を用いた VR が利用される理由は，リアルな体験が，手軽にできることである．HMD 着用者の動きが VR 内に反映されることが，リアルな体験を生む一つの要因である．しかし，HMD やコントローラのセンサだけでは，着用者の動きの反映に限界がある．現在では，追加のデバイスを用いることにより，着用者の動きより精度よく VR 内に反映することができる．しかし，追加のデバイスが高価であることや，広い空間を必要とすることなど，手軽が損なわれるという問題がある．

HMD の手軽さを損なわずに，より精度よく着用者の動きを VR 内に反映するために，一人称視点画像を用いた姿勢推定を VR に活用することを考える．このような姿勢推定手法を用いた VR は存在せず，現状の姿勢推定手法の精度では，VR に用いるのに不十分である．

そこで，一人称視点画像を用いた姿勢推定の VR への利用に向け，既存推定手法の精度改善を本研究の目的とする．特に，位置情報の精度改善を図る．

2. Ego-Pose Estimation and Forecasting as Real-Time PD Control

Ego-Pose Estimation and Forecasting as Real-Time PD Control は Ye Yuan らが提案した一人称視点画像を用いた姿勢推定手法である[2]．図 1 のような，一人称視点画像を入力として，対象者の姿勢推定を行う．

図 1　一人称視点画像

3. 提案手法

本研究では，既存手法の Ego-Pose Estimation and Forecasting as Real-Time PD Control の報酬関数を変更することで，位置情報の精度を改善する手法を提案する．具体的には，既存の報酬関数に，ワールド座標報酬r_cを追加する．ワールド座標報酬r_cの式を次に示す．

$$r_c = \exp[-k_c((x_t - \hat{x}_t)^2 + (y_t - \hat{y}_t)^2)] \qquad (1)$$

ここで，(x_t, y_t)は時刻tにおける姿勢推定結果のxy座標，$(\hat{x}_t, \hat{y}_t)$は時刻tにおけるグランドトゥルースのxy座標を表し，k_cは関数の減少速度を表す．

4. 実験

4.1 環境

本実験では，既存手法と同様に，ヒューマノイドモデルを動かす物理制御ベースの仮想環境として mujoco[3]を用いる．実験で用いたヒューマノイドモデルは, 58 個の自由度と 21 個の剛体で構成され, Hips をルートジョイントとしている．ほとんどの非ルートジョイントでは，自由度を 3 つ持つが，膝と足首は 1 つの自由度のみを持つ．

4.2 データセット

仮想環境と同様に，既存手法と同じデータセットを用いる．データセットは，図 1 のような一人称視点画像と，カメラ着用者の姿勢情報が対となったデータで構成される．GoPro カメラを着用し，さまざまな動作（歩き，走り，回転，しゃがみなど）を実行し，得られた映像とそのときのポーズでデータセットが作られている．実験では，トレーニング用に 9 テイク，テスト用に 5 テイク使用した．

4.3 実験結果

既存手法と提案手法において，同じ一人称視点画像を入力としたときの姿勢推定結果を図 2 に示す．図 2 の左に並ぶ画像が既存手法の姿勢推定結果で，右に並ぶが画像が提案手法の姿勢推定結果である．また，図 2 において，赤色のヒューマノイドがグランドトゥルースで，橙色のヒューマノイドが推定結果を示す．得られた姿勢推定結果を 100 フレームごとに区切った画像の一部を載せている．この時$k_c = 0.3$としている．

図 2 より，提案手法の方が，ヒューマノイドモデルの位置をより正確に推定できていることがわかる．しかし，フレーム後半になるとグランドトゥルースと推定結果の位置の差が大きくなっている．

4. 4 評価実験

提案手法の定量的に評価するために，評価実験を行った． 既存手法で用いられている評価項目である Pose Error (Epose)， Velocity Error (Evel)， Average Acceleration (Aaccl) に加え，位置情報の精度改善を評価するために，Coordinate Error (Ecoor) という項目を追加する．

Epose は生成されたポーズシーケンスとグランドトゥルースのポーズシーケンスとの差を評価する． Evel は生成された速度シーケンスとグランドトゥルースの速度シーケンスとの差を評価する． Aaccl は生成されたポーズシーケンスの滑らかさを評価する． Ecoor は生成されたルートジョイントのxy座標とグランドトゥルースのルートジョイントのxy座標の差を評価する．

既存手法と提案手法の各評価項目の値を表 1 に示す．提案手法に関しては，k_cの値を 0.3，3，30，300，3000 と変化させたときの 5 通りを示す． 表 1 の括弧の中は，各評価項目が既存手法に対してどれだけ増加したかをパーセントで表したものである．

表 1　既存手法と提案手法の評価項目値と増加率

評価項目	Epose	Evel	Aaccl	Ecoor
既存手法	0.5452	3.7862	4.3884	0.029184
$k_c = 0.3$	0.5633 (+3.3%)	3.8373 (+1.3%)	4.5086 (+2.7%)	0.018360 (-37.1%)
$k_c = 3$	0.5536 (+1.5%)	3.8289 (+1.1%)	4.4628 (+1.7%)	0.019357 (-33.7%)
$k_c = 30$	0.5385 (-1.2%)	3.7726 (-0.3%)	4.6030 (+4.9%)	0.025328 (-13.2%)
$k_c = 300$	0.5490 (+0.7%)	3.7778 (-0.2%)	4.3494 (-0.9%)	0.019827 (-32.1%)
$k_c = 3000$	0.5550 (+1.8%)	3.8028 (+0.4%)	4.4154 (+0.6%)	0.021508 (-26.3%)

表 1 より，どの提案手法においても，Ecoor を改善できた．また, その他の評価項目である Epose, Evel, Aaccl については, 精度の低下が見られる．しかし，図 2 からは，その他の評価項目の精度低下の影響は見られない．また，表 1 から，$k_c = 30$を除いて，k_cの値が小さいほど，Ecoor が大きく改善されていることがわかる．これは、k_cは関数の減少速度を決め，k_cの値が小さいほど，関数の減少速度が小さくなり，ワールド座標報酬の減少が滑らかになる．これにより，xy座標のエラーが大きくなってもある程度の報酬が与えられたことによるものだと考える．表 1 の Epose，Evel，Aaccl の値は既存手法と比べ増加しているものが多く見られるのは、ワールド座標報酬を追加したことにより，位置に対する強制力が強くなった結果, 生成されるポーズに対して，悪影響を及ぼした結果であると考る．

5. まとめ

本研究では，既存手法の Ego-Pose Estimation and Forecasting as Real-Time PD Control に新しくワールド座標報酬r_cを追加することで，位置情報の精度を改善する手法について提案した．

既存手法との比較の結果，位置をより正確に推定できていることがわかった．評価実験では，定量的にも，位置情報の精度が改善されていることが示せた．

今後の課題として，さらなる精度改善と，位置情報のエラーが蓄積によりフレーム後半になると位置情報の差が大きくなることがある．k_cの値や，各報酬の重みについて実験を行うことでさらなる精度改善が見込める．後者に関しては，姿勢推定時にグランドトゥルースにアクセスしていないため，座標エラーの蓄積は回避できるものではない． しかし，VR ゲームでは仮想環境内での座標にアバタの位置をリセットするなどして，座標エラーの蓄積の問題を回避することができる．この際，位置情報の精度が高いほど，リセット回数が少なくて済む．

図 2　右:既存手法の結果　左:提出手法の結果

参考文献

[1]ビジネス・ブレイクスルー大学大学院, “世界の VR 市場動向”, 19 10 2021. https://www.gii.co.jp/report/mama1003249-ar-vr-display-market-covid-impact-analysis-by.html [アクセス日: 26 1 2022].

[2] Y. Yuan and K. Kitani. “Ego-pose estimation and forecasting as real-time pd control”, In IEEE International Conference on Computer Vision (ICCV), pages 10082–10092, Oct. 2019.

[3] Mujoco, 2022, https://mujoco.org/ [アクセス日: 26 1 2022]

2021 年度

一人称視点画像を用いた姿勢推定手法における位置情報の精度改善

Improving the Accuracy of Position Information in Egocentric Pose Estimation Methods

指導教員　渡辺　裕　教授

提出日：2022 年 2 月 1 日

早稲田大学　基幹理工学部　情報通信学科

1W182238-6

中島　聖

目次

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 1
第 2 章 関連技術 3
2.1 まえがき 3
2.2 一人称視点画像 3
2.3 一人称視点画像を用いた姿勢推定 3
2.4 Ego-Pose Estimation and Forecasting as Real-Time PD Control 4
2.4.1 概要 4
2.4.2 深層強化学習 5
2.4.3 PWC-Net 5
2.4.4 ResNet-18 6
2.4.5 LSTM 6
2.5 むすび 6
第 3 章 提案手法 7
3.1 まえがき 7
3.2 提案手法 7
3.3 むすび 8
第 4 章 実験 9
4.1 まえがき 9
4.2 環境 9
4.3 データセット 9
4.4 評価実験 12
4.5 実験結果 12
4.6 考察 13
4.7 むすび 14
第 5 章 結論と今後の課題 15
5.1 結論 15
5.2 今後の課題 15
謝辞 16
参考文献 17
図一覧 19

表一覧 20

第1章 序論

1.1 研究背景

近年，医療や観光，スポーツ，ゲームなど，さまざまな分野で Virtual Reality (VR) が利用されている。VR の中でも特に，Head Mounted Display (HMD) を用いた VR，いわゆる狭義の VR の活用が増加している[1]．さまざまな分野で VR が利用される理由は，リアルな体験が手軽にできるという点である．VR は，HMD と対応するコントローラさえあれば，利用可能であり，外出する必要もないため，とても手軽である．また，HMD による頭に合わせた映像の動きや，左右の目に別々の映像を見せることによる立体視，センサによる自身の動きの反映などにより，リアルな体験を生んでいる．VR には，もちろん短所もある．例えば，HMD 装着による蒸れや違和感，VR 酔い，現実には及ばないリアリティ性などがある．これらの短所は，現実との乖離が引き起こしていることであり，よりリアルな VR が求められる．VR の現実との乖離の原因の一つに，HMD やコントローラのセンサでのボディトラッキングの限界がある．HMD やコントローラのセンサでは，頭部や手の動きを感知できるが，その他の部分については情報が得られないためである．現在では，追加のデバイスでより良いボディトラッキングが可能であるが，追加のデバイスが高価であったり，大規模であったりして，VR の利点である手軽さを損なってしまう．

1.2 研究目的

VR の手軽さを損なわずに，より良いボディトラッキングを可能にするために，一人称視点画像を用いた姿勢推定を VR ゲームに利用することを考える．この姿勢推定を VR に用いた事例は，現状では存在せず，今の精度では不十分であることが言える．そこで，一人称視点映像を用いた姿勢推定の VR ゲームへの利用に向け，姿勢推定手法の精度改善を本研究の目的とする．特に，位置情報の精度改善を図る．

一人称視点画像を用いた姿勢推定は，画像内に身体の一部が映ることを利用し姿勢推定を行う手法と，身体が全く映らない手法の二つに分けることができる．前者は，多くの場合，一人称視点画像に身体が映るように下向き魚眼カメラが用いられる[2,3,4,5,6]．後者は身体が観測できないため，より困難な課題である[7,8,9,10]．VR ゲームへの利用において，前者の手法は，新たに魚眼カメラを追加する必要がある場合があることや，仮想世界の映像を用いられないことなどの問題がある．そこで，後者に分類される手法の中でも，目の高さからの一人称視点画像を用いた手法[10]が，VR ゲームの利用に適切であり，この手法の位置精度改善を図る．

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「Ego-Pose Estimation and Forecasting as Real-Time PD Control」[10]について述べる．また，一人称視点画像，姿勢推定，上記手法を構成する技術について述べる．

2.2 一人称視点画像

一人称視点とは，主に，ゲームで用いられる言葉で，プレイヤが操作するキャラクタの目に映るグラフィック視点のことを指す．この一人称視点を用いたガンシューティングゲームは First Person Shooter（FPS）と呼ばれ，現在多くのプレイヤ人口を持つゲームジャンルの一つとなっている．具体的に，FPS は図 1 左に示すような画像で構成される．一人称視点とは対照的に，三人称視点という言葉がある．三人称視点は，ゲームにおいて，プレイヤが操作するキャラクタの背後や斜め上等から見るカメラ視点のことで，図 2.1 右のような画像を指す．俯瞰視点とも呼ばれる．本研究では，図 2.1 左のような，一人称視点画像を取り扱う．

図 2.1　左: FPS の画像[11]　右: 三人称視点（minecraft）の画像[12]

2.3 一人称視点画像を用いた姿勢推定

姿勢推定は画像認識の一分野である．画像を入力とし，対象者の姿勢を推定する．一般的な姿勢推定が画像内に対象者が映っていることを利用し，姿勢を推定するのに対し，一人称視点画像を用いた姿勢推定は，1.2 節でも述べた通り，画像内に対象者の身体が全く映らない場合もある．また，本研究では，対象者の身体が全く映らない画像を用いる．図 2.2 に一人称視点画像を用いた姿勢推定の概略を示す．図 2.2 のように，一人称視点画像を入力とし，姿勢推定結果を出力として得る．

図 2.2　一人称視点画像を用いた姿勢推定の概略

2.4 Ego-Pose Estimation and Forecasting as Real-Time PD Control

2.4.1 概要

「Ego-Pose Estimation and Forecasting as Real-Time PD Control」は Ye Yuan らが提案した一人称視点画像を用いた姿勢推定手法である．姿勢推定と行動予測手法についての論文であるが，行動予測については，本論文では用いないため，姿勢推定についてのみ述べる．この手法は，一人称視点画像を用いた姿勢推定手法の中でも，推定対象者の身体が画像に映らないタイプの手法で，この手法の特徴ともなっている．姿勢推定の流れは図 2.3 に示す．また，図 2.3 の説明を以下に示す．

図 2.3　Ego-Pose Estimation and Forecasting as Real-Time PD Control の概略

まず，Ego Frames と言われる一人称視点画像の Optical flow を得る．Optical flow としては，PWC-Net[13]を用いる．次に，得られた Optical flow を ResNet-18[14]に通し，視覚的特徴を得る．そして，視覚的特徴を LSTM に通し，視覚的コンテキストを生成し，このコンテキストから，フレームごとのコンテキストϕ_tを得る．コンテキストϕ_tとヒューマノイドのポーズz_tから状態s_tを取得し，状態s_tのヒューマノイドエージェントはポリシーから得られるアクションa_tを実行する．そして，物理シミュレーション環境（mujoco[15]）を通して，状態s_{t+1}を生成し，エージェントに報酬関数によって決められた報酬を与える．

このようにして学習したエージェントに加え，Fail-Safe メカニズムと言われる機構がこの姿勢推定手法には，用いられている．Fail-Safe メカニズムは，ヒューマノイドのポーズの初

期状態の推定や，転倒防止機能に用いられているが，詳しくは既存手法[10]を参照。

既存手法は，深層強化学習を用いた物理制御ベースの姿勢推定手法であり，2.4.2 節から 2.4.5 節で，上記の要素技術について述べる．

2.4.2 深層強化学習

強化学習とは，エージェントが環境に対して行動をし，報酬関数に基づいて得られる報酬を最大化するような policy を求める機械学習アルゴリズムの一種である[16]．図 2.4 に強化学習の流れを示す．エージェントは，学習者という位置付けで，このエージェントが試行錯誤を繰り返し学習していく．環境とは，観測者という位置付けで，エージェントの行動に対し，状態の更新と報酬の付与を行う．報酬とは，エージェントの行動に対し，環境が与えるもので，報酬関数により，報酬の量が決められる．

図 2.4　強化学習の流れ

深層強化学習とは，上述の強化学習と深層学習を組み合わせた機械学習で，特徴量の抽出を人間の介入なしに行うことができる．強化学習に畳み込みニューラルネットワーク（CNN）が用いられているため，既存手法である「Ego-Pose Estimation and Forecasting as Real-Time PD Control」は深層強化学習と呼ばれる．

2.4.3 PWC-Net

PWC-Net は Optical flow の一種で，フレーム間の物体の動きを検出して速度をベクトルで表示する手法である．また，Optical flow は以下の過程に基づいている．「連続フレーム間で物体の画像上の明るさは変わらない」と，「隣接する画素は似たような動きをする」である．既存手法では PWC-Net という手法を用いている．PWC-Net を用いて得られる Optical flow を画像にしたものの例を図 2.5 に示す。

図 2.5　Optical flow の例(PWC-Net)

2.4.4 ResNet-18

ResNet-18 とは，深さが 18 層の畳み込みニューラルネットワーク（CNN）である．100 万枚を超えるイメージで学習させた ResNet-18 を ImageNet データベース[17]から読み込むことができる．この学習済みの ResNet-18 はイメージを 1000 個のオブジェクトカテゴリに分類でき，広範囲のイメージに対して，豊富な特徴表現を学習している．ResNet-18 は主に，画像の特徴抽出に用いられる．

2.4.5 LSTM

LSTM とは，Recurrent Neural Network(RNN)という，時系列を考慮するための層を改良してできたものである．RNN の概略を図 2.6 に示す．時刻tにおいて，A はx_1と時刻$t-1$からのインプットを受け取る．このように，前の時刻からインプットがあることで，時刻間の影響を考慮したネットワークを作っている．

図 2.6　RNN の概略

2.5 むすび

本章では，本研究で用いる技術の基礎である姿勢推定，およびその発展形である一人称視点画像を用いた姿勢推定について述べた．また，これらの技術を構成する深層強化学習，PWC-Net，ResNet-18，LSTM について述べた．

第3章 提案手法

3.1 まえがき

本章では，既存手法の「Ego-Pose Estimation and Forecasting as Real-Time PD Control」の報酬関数に，新しい報酬関数を追加することで，位置情報の精度を改善する手法について述べる．

3.2 提案手法

既存手法で用いられる，ポーズ報酬r_{p}，エンドエフェクタ報酬r_{e}，ルートポーズ報酬r_{rp}，ルート速度報酬r_{rv}の四つの報酬関数に新しくワールド座標報酬r_{c}を追加する．ポーズ報酬r_{p}は，推定結果のポーズp_tとグランドトゥルースのポーズ$\hat{p}_t$の差を評価する関数で，次の式で表される．ここで，q^jはジョイントjのローカル方向のクォータニオンを表す．

$$r_{\mathrm{p}} = \exp\left[-2\left(\sum_j \left\| q_t^j \ominus \hat{q}_t^j \right\|^2\right)\right] \tag{3.1}$$

エンドエフェクタ報酬r_eは，推定結果のエンドエフェクタベクトルe_tとグランドトゥルースのエンドエフェクタベクトル$\hat{e}_t$の差を評価する関数で，次の式で表される．ここで，エンドエフェクタベクトルe_tはルートから各エンドエフェクタ（足，手，頭）までのベクトルを表す．

$$r_{\mathrm{e}} = \exp\left[-20\left(\sum_e \| e_t - \hat{e}_t \|^2\right)\right] \tag{3.2}$$

ルートポーズ報酬r_{rp}はヒューマノイドのルートジョイント(Hips)がグランドトゥルースの高さ$\hat{h}_t$とジョイントrのクォータニオン$\hat{q}_t^{\mathrm{r}}$を持つように促す関数で，次の式で表される．

$$r_{\mathrm{rp}} = \exp\left[-300\left(\left(h_t - \hat{h}_t\right)^2 + \| q_t^{\mathrm{r}} \ominus \hat{q}_t^{\mathrm{r}} \|^2\right)\right] \tag{3.3}$$

ルート速度報酬r_{rv}はルートの線形速度l_tと角速度ω_tとグランドトゥルースの線形速度$\hat{l}_t$と角速度$\hat{\omega}_t$からの偏差を評価する関数で，次の式で表される．

$$r_{\mathrm{rv}} = \exp\left[-\left\| l_t - \hat{l}_t \right\|^2 - 0.1 \| \omega_t^{\mathrm{r}} - \hat{\omega}_t^{\mathrm{r}} \|^2\right] \tag{3.4}$$

これら四つの報酬それぞれに重みを掛け，足し合わせたものを重みの合計で割ったものが既存手法の最終的な報酬関数r_tなる。この式を次に示す．

$$r_t = \frac{\mathrm{w_p} r_{\mathrm{p}} + \mathrm{w_e} r_{\mathrm{e}} + \mathrm{w_{rp}} r_{\mathrm{rp}} + \mathrm{w_{rv}} r_{\mathrm{rv}}}{\mathrm{w_p} + \mathrm{w_e} + \mathrm{w_{rp}} + \mathrm{w_{rv}}} \tag{3.5}$$

ここで，$\mathrm{w_p}$，$\mathrm{w_e}$，$\mathrm{w_{rp}}$，$\mathrm{w_{rv}}$は重み係数を表す。報酬の重み$(\mathrm{w_p}, \mathrm{w_e}, \mathrm{w_{rp}}, \mathrm{w_{rv}})$は$(0.5, 0.3, 0.1, 0.1)$に設定されている．

上述の通り，提案手法では，上記の報酬関数に加え，新たにワールド座標報酬r_{c}を追加する。

ワールド座標報酬r_cは，推定結果のルートのxy座標とグランドトゥルースのルートのxy座標の差を評価する関数で，次の式で表される．

$$r_c = \exp[-\mathrm{k_c}((x_t - \hat{x}_t)^2 + (y_t - \hat{y}_t)^2)] \tag{3.6}$$

ここで，$\mathrm{k_c}$は関数の減少速度を表す。つまり，提案手法の最終的な報酬関数は，次の式のようになる．

$$r_t = \frac{\mathrm{w_p} r_\mathrm{p} + \mathrm{w_e} r_\mathrm{e} + \mathrm{w_{rp}} r_\mathrm{rp} + \mathrm{w_{rv}} r_\mathrm{rv} + w_\mathrm{c} r_\mathrm{c}}{\mathrm{w_p} + \mathrm{w_e} + \mathrm{w_{rp}} + \mathrm{w_{rv}} + \mathrm{w_c}} \tag{3.7}$$

3.3 むすび

本章では，既存手法に新しくワールド座標報酬を追加することで，位置情報の精度改善する手法について述べた．

第4章 実験

4.1 まえがき

本章では，提案手法において，$k_c = 0.3, 3, 30, 300, 3000$の五つのパターンと既存手法との比較した結果を示す．また，実験に用いた環境，データセット，結果，評価実験，考察についても述べる．

4.2 環境

Mujoco という接触のある多関節の動力学シミュレーションに適した物理エンジンを仮想環境として使用している．ヒューマノイドモデルは 58 個の Degree of Freedom (DoF) と 21 個の剛体で構成され，Hips をルートジョイントとしている．ほとんどの非ルートジョイントでは，DoF を三つ持つが，膝と足首は一つの DoF のみを持つ．DoF は自由度を表す．安定した PD コントローラ[18]を使用して，関節トルクを計算する．

4.3 データセット

既存手法で用いられているデータセットを提案手法でも用いる．データセットの中身は，図4.1のような一人称視点画像と，カメラ着用者のポーズが対となったデータで構成される．GoPro カメラを着用し，さまざまな動作（歩き，走り，回転，しゃがみなど）を実行し，得られた映像とそのときのポーズでデータセットが作られている．各テイクの長さは，約 1 分でモーションのセグメント化やラベル付けは行われていない．また，各テイクは約 1800 フレームからなる．画像サイズは 224×224.

4.4 実験結果

既存手法と提案手法において，同じテイクでトレーニングを行い，同じテイクを入力としたときの得られた姿勢推定結果を図 4.2 に示す．トレーニング用のテイクとテスト用のテイクは異なるものを用いている．トレーニング用に 9 テイク，テスト用に 5 テイク用いているが，図 4.2 はその中の 1 テイクの姿勢推定結果を載せている．図 4.2 の左に並ぶ画像が既存手法の推定結果を示し，右に並ぶ画像が提案手法の推定を示す．また，赤色のヒューマノイドモデルは，グランドトゥルースの姿勢推定結果を示し，橙色のヒューマノイドモデルは推定結果を示す．既存手法と提案手法ともに，得られた推定結果を 50 フレームごとに分割し，その一部を図 4.2 に載せている．提案手法において，$k_c = 0.3$を用いている．

図 4.2 より，提案手法の方が，グランドトゥルースのヒューマノイドモデルと推定結果のヒューマノイドモデルの位置の差が小さく，提案手法の方がより正確に位置を推定できていることがわかる．また，フレーム後半になると，グランドトゥルースと推定結果の位置の差が大きくなっていることもわかる．

図 4.1　データセットの一人称視点画像例

図 4.2　右: 既存手法の推定結果　左: 提案手法の推定結果

4.5 評価実験

4.5.1 評価指標

既存手法で用いられている評価指標に加えて，推定結果とグランドトゥルースの座標の差を評価するために，新しく座標エラーに関する評価項目を追加する．既存の評価項目としては，Pose Error(Epose)，Velocity Error(Evel)，Average Acceleration(Aaccl)を用いる．Eposeは生成されたポーズシーケンス$p_{1:T}$とグランドトゥルースのポーズシーケンス$\hat{p}_{1:\mathrm{T}}$の間のユークリッド距離を測定する．次の式で計算される．Tはフレーム数を表す．

$$\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\|p_t-\hat{p}_t\|_2 \tag{4.1}$$

Evel は，生成された速度シーケンス$v_{1:T}$とグランドトゥルースの速度シーケンス$\hat{v}_{1:T}$の間のユークリッド距離を測定する．次の式で計算される．v_tと$\hat{v}_t$は有限差分法で計算される．

$$\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\|v_t-\hat{v}_t\|_2 \tag{4.2}$$

Aaccl は関節の加速度の平均の大きさを使用し，生成されたポーズシーケンスの滑らかさを測定する．次の式で計算される．ここで，$\dot{v}_t$は関節加速度を，G は動作した DoF の数を表す．

$$\frac{1}{\mathrm{TG}}\sum_{t=1}^{\mathrm{T}}\|\dot{v}_t\|_1 \tag{4.3}$$

新しい評価項目として，Coordinate Error(Ecoor)というxy座標に関する評価項目を追加する．Coordinate Error は生成されたポーズシーケンスのxy座標(x,y)とグランドトゥルースのののxy座標$(\hat{x},\hat{y})$の差を測定する．次の式で表される．

$$\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}|(|x_t-\hat{x}_t|+|y_t-\hat{y}_t|)-(|x_{t-1}-\hat{x}_{t-1}|+|y_{t-1}-\hat{y}_{t-1}|)| \tag{4.4}$$

4.5.2 結果

提案手法と既存手法での各評価項目の値を表 4.1 に示す．提案手法に関しては，式(6)の$\mathrm{k_c}$の値を 0.3，3，30，300，3000 と変化させたときの 5 通りを評価する．この時の既存手法に対しての各提案手法の各評価項目の増加率を括弧の中に示す．どの提案手法においても，Ecoor を改善できていることがわかる．特に，$\mathrm{k_c}=0.3$の時の Ecoor が最も大きく改善できていることがわかる。

表 4.1 と図 4.2 より，定量的にも定性的にも，位置情報の精度が改善されていることがわかった。表 4.1 より，Epose，Evel，Aaccl の値が増加していることがわかる。しかし，図 4.2 より，Epose，Evel，Aaccl の増加による，姿勢推定結果への視覚的な影響は見られない．

表 4.1　既存手法と提案手法の評価項目値

評価項目	Epose	Evel	Aaccl	Ecoor
既存手法	0.5452	3.7862	4.3884	0.029184
$k_c = 0.3$	0.5633 (+3.3%)	3.8373 (+1.3%)	4.5086 (+2.7%)	0.018360 (-37.1%)
$k_c = 3$	0.5536 (+1.5%)	3.8289 (+1.1%)	4.4628 (+1.7%)	0.019357 (-33.7%)
$k_c = 30$	0.5385 (-1.2%)	3.7726 (-0.3%)	4.6030 (+4.9%)	0.025328 (-13.2%)
$k_c = 300$	0.5490 (+0.7%)	3.7778 (-0.2%)	4.3494 (-0.9%)	0.019827 (-32.1%)
$k_c = 3000$	0.5550 (+1.8%)	3.8028 (+0.4%)	4.4154 (+0.6%)	0.021508 (-26.3%)

4.6 考察

本研究において, Ecoor が最も重視する評価項目であり, その他の評価項目の Epose, Evel, Aaccl に関しては，同程度の重要性とする．また，K_cの値を変化させたときの Ecoor とその他の評価項目の関係を調べるために，その他の三つの項目の増加率の合計と Ecoor の増加率を表 4.2 に示す．このとき，上述の通り，その他の評価項目は，同程度の重要度と考えているため，1:1:1 の割合で，増加率を加算し，Ecoor の増加率とその他の評価項目の増加率の合計に分けている．この時のグラフを図 4.3 に示す．

表 4.2　その他の三つの項目（Epose，Evel，Aaccl）の増加率の合計と Ecoor の増加率

増加率[%]	三つの項目の増加率の合計	Ecoor の増加率
$k_c = 0.3$	7.5	-37.1
$k_c = 3$	4.4	-33.7
$k_c = 30$	3.3	-13.2
$k_c = 300$	-0.4	-32.1
$k_c = 3000$	2.9	-26.3

図 4.3　その他の三つの項目（Epose，Evel，Aaccl）の増加率の合計と Ecoor の増加率の関係

図 4.3 において，プロットが左下に行くほど，その他の三つの項目と Ecoor がより大きく改善されたということを示すため，一番左下にプロットされている$k_c = 300$の時がその他の三つの項目と Ecoor をともに改善できていることがわかる．

図 4.3 から$k_c = 30$を除いて，k_cの値が小さいほど，Ecoor が大きく改善されていることがわかる．k_cは関数の減少速度を決め，k_cの値が小さいほど，関数の減少速度が小さくなり，ワールド座標報酬の減少が滑らかになる．これにより，xy座標のエラーが大きくなってもある程度の報酬が与えられたことによるものである．また，$k_c = 300$を除いて，k_cの値が小さいほど，その他の三つの項目の増加率が大きくなっていることもわかる．これも，上述の通り，xy座標エラーが大きくなっても，報酬がある程度与えられることで，位置に対する強制力が強くなった結果である．

研究の根幹の目的となる VR ゲームへの利用を考えると，プレイヤはコントローラによりアバタを動かすため，推定結果とプレイヤの操作にギャプがないように，提案手法のような一人称視点画像を用いた姿勢推定を用いるとき，座標エラーが小さいことが望まれる．しかし，フレーム後半になると座標エラーの蓄積により，位置情報の精度は低くなってしまうという問題があるが，実際に VR ゲームに利用する時は，仮想環境内での座標にアバタの位置をリセットするなどして，座標エラーの蓄積の問題を回避することができる．この際，位置情報の精度が高いほど，リセット回数が少なくて済む．そのため，提案手法による位置情報の精度を改善は，一人称視点画像を用いた姿勢推定の VR ゲームへ利用に関して，有益な結果である．

4.7 むすび

本章では，提案手法の実験結果と既存手法との比較，及び考察について述べた．提案手法において，位置情報の精度改善が定量的にも定性的にも見ることができた．

第5章 結論と今後の課題

5.1 結論

本研究では，既存手法の「Ego-Pose Estimation and Forecasting as Real-Time PD Control」に新しくワールド座標報酬を追加することで，位置情報の精度を改善する手法について提案した．実験結果より，ワールド座標報酬を追加することで，位置情報の精度を改善できることがわかった．

5.2 今後の課題

提案手法では，ワールド座標報酬の重みを 0.1 として，実験を行ったが，各報酬の重みについては，検討の余地がある．また，ワールド座標報酬内のk_cの値についても，より多くのパターンで実験することにより，よりよい精度が得られる可能性がある．

提案手法により，位置情報の精度を改善することができたが，座標エラーの蓄積により，フレームの後半になると位置情報の精度は低くなってしまう．しかし，姿勢推定時に位置情報のグランドトゥルースにアクセスしていないため，座標エラーの蓄積は回避できるものではない．提案手法段階では，座標エラーの蓄積をなくすことはできないが，上記のように，より位置情報の精度を高めることや，アプリケーションに利用する時に，4.6 節で述べたような利用の仕方をすることで，問題を回避できる．

本実験では，現在世界の一人称視点画像を用いて評価を行っているため，VR ゲームへの利用に対して，仮想世界の一人称視点画像を用いて評価を行うことで，新たな問題が見えてくるかもしれない．

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や，適切なアドバイスをくださった早稲田大学国際情報通信センターの石川孝明様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.

[1] ビジネス・ブレイクスルー大学大学院, “世界の VR 市場動向”, 19 10 2021. https://www.gii.co.jp/report/mama1003249-ar-vr-display-market-covid-impact-analysis-by.html [アクセス日: 26 1 2022].

[2] D. Tome, P. Peluse, L. Agapito, and H. Badino. ”xR-EgoPose: Egocentric 3D Human Pose from an HMD Camera”, In IEEE International Conference on Computer Vision (ICCV), Jul. 2019

[3] H. Rhodin, C. Richardt, D. Casas, E. Insafutdinov, M. Shafiei, H. P. Seidel, B. Schiele, and C. Theobalt. “Egocap: Egocentric marker-less motion capture with two fisheye cameras”, ACM Trans. Graph., 35(6), Nov. 2016.

[4] W. Xu, A. Chatterjee, M. Zollhoefer, H. Rhodin, P. Fua, H. P. Seidel, and C. Theobalt. “Mo2Cap2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera” IEEE Transactions on Visualization and Computer Graphics, pages 1–1, 2019.

[5] D. Tome, T. Alldieck, P. Peluse, G. P. Moll, L. Agapito, H. Badino, and F. Torre. “SelfPose: 3D Egocentric Pose Estimation from a Headset Mounted Camera”, Nov. 2020

[6] H. Jiang, and V. K. Ithapu. “Egocentric Pose Estimation from Human Vision Span”, In IEEE International Conference on Computer Vision (ICCV), Apr. 2021

[7] . Jiang and K. Grauman. “Seeing invisible poses: Estimating 3d body pose from egocentric video”, In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3501–3509, Jun. 2016.

[8] E. Ng, D. Xiang, H. Joo, and K. Grauman. “You2me: Inferring body pose in egocentric video via first and second person interactions”, In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Mar. 2020

[9] Y. Yuan and K. Kitani. “3d ego-pose estimation via imitation learning”, In The European Conference on Computer Vision (ECCV), Sep. 2018.

[10] Y. Yuan and K. Kitani. “Ego-pose estimation and forecasting as real-time pd control”, In IEEE International Conference on Computer Vision (ICCV), pages 10082–10092, Oct. 2019.

[11] pixabay, “高品質なフリー画像素材”, https://pixabay.com/ja/ [アクセス日: 27 1 2022]

[12] Minecraft, 2022, https://www.minecraft.net/ja-jp [アクセス日: 27 1 2022]

[13] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8934–8943, 2018.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016

[15] Mujoco, 2022, https://mujoco.org/ [クセス日: 27 1 2022]

[16] AVINTON, “深層強化学習の基礎”, https://avinton.com/blog/2017/10/basic-deep-reinforcement-learning/ [アクセス日: 26 1 2022]

[17] Imagenet, 2021, https://www.image-net.org/ [アクセス日: 26 1 2022]

[18] J. Tan, K. Liu, and G. Turk. Stable proportional-derivative controllers. IEEE Computer Graphics and Applications, 31(4):34–44, 2011

図一覧

図 2.1　左: FPS の画像[11]　右: 三人称視点（minecraft）の画像[12] 3
図 2.2　一人称視点画像を用いた姿勢推定の概略 4
図 2.3　Ego-Pose Estimation and Forecasting as Real-Time PD Control の概略 4
図 2.4　強化学習の流れ 5
図 2.5　Optical flow の例(PWC-Net) 6
図 2.6　RNN の概略 6
図 4.1　データセットの一人称視点画像例 10
図 4.2　右: 既存手法の推定結果　左: 提案手法の推定結果 11
図 4.3　その他の三つの項目（Epose，Evel，Aaccl）の増加率の合計と Ecoor の増加率の関係 14

表一覧

表 4.1　既存手法と提案手法の評価項目値 13
表 4.2　その他の三つの項目（Epose，Evel，Aaccl）の増加率の合計と Ecoor の増加率 ... 13