アバタ媒介型見守りシステムにおける Kinect の姿勢推定エラー補正

Correcting Kinect's Pose Estimation Error in Avatar Mediated Distant-Care System

尾花謙伍 [1*] , 長谷川大 [2], 白川真一 [3], 金子直史 [4],
佐久田博司 [4], 鷲見和彦 [4]
Kengo Obana[1], Dai Hasegawa[2], Shinichi Shirakawa[3], Naoshi Kaneko[4],
Hiroshi Sakuta[4], and Kazuhiko Sumi[4]

[1] 青山学院大学大学院　理工学研究科
[1] Graduate School of Science and Engineering, Aoyama Gakuin University
[2] 東京工科大学　メディア学部
[2] School of Media Science, Tokyo University of Technology
[3] 横浜国立大学大学院　環境情報研究院
[3] Faculty of Environment and Information Science, Yokohama National University
[4] 青山学院大学　理工学部
[4] College of Science and Engineering, Aoyama Gakuin University

Abstract: プライバシに配慮した独居高齢者の見守りを実現するアバタ媒介型見守りシステムの提案が行われている．本稿では，アバタ媒介型見守りシステムで使用されている Kinect による姿勢推定誤差をディープラーニングを用いて補正することを目的とする． Kinect と高精度なモーションキャプチャで同時に取得したモーションデータをデータセットとし，時系列データを扱うことが可能な RNN を用いて Kinect による姿勢推定誤差を補正可能なネットワークを構築する．

1　背景

近年，少子高齢化の進行やライフスタイル・家族形態の変化にともなって，独居高齢者世帯や高齢者夫婦のみの世帯が増加しており，高齢者の社会的孤立が問題となっている．高齢者の社会的孤立により心身状態の変化に対する観取が遅れ，問題を悪化させるケースが存在し，自立した生活を困難にする間接的な要因となっている．

これまでに，上記の問題を解決すべく高齢者をカメラ映像で見守る手法が検討されている [1]．しかし，カメラ映像を利用した高齢者の見守り支援は介護者の負担を減らせることが期待できる反面，プライバシ侵害の懸念が大きいことや高齢者が監視されていることにより，精神的負担を感じることが指摘されている．

そこで長谷川ら [2] により，プライバシを保護しながら家族，友人，近隣住民による日常的な高齢者世帯の見守りを可能にするアバタ媒介見守りシステム (Avatar Mediated Distant-Care : AMD) が提案されている．

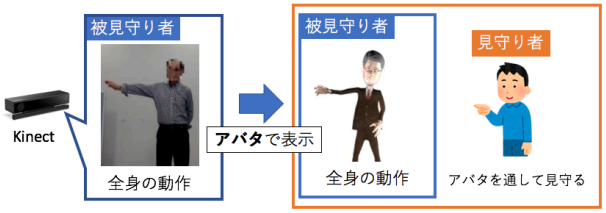

図 1: AMD のシステム概要

AMD のシステム概要を図 1 に示す．まず被見守り者の全身の動きを安価なマーカレスモーションキャプチャシステム (Kinect) によって取得し，得られた動きを 3D アバタに適用する．見守り者は被見守り者の動きが適用されたアバタを通して，被見守り者の見守りを行う．Kinect と高精度なマーカ付きモーションキャプチャシステムで同時に録画されたモーションを使用し，アバタが行っているモーションの把握率を調査したところ，「手を挙げる」や「伸びをする」などの大ま

*連絡先：青山学院大学大学院理工学研究科理工学専攻
〒 252-5258 神奈川県相模原市中央区淵野辺 5-10-1
E-mail: c5616146@aoyama.jp

かな動作であれば同程度の情報伝達を維持できることが明らかにされた [2]．これにより高価なカメラを複数台必要とするモーションキャプチャシステムを導入することや，日常的に身体にマーカを着けて生活をしなくても見守りが可能であることが考えられる．

また Hasegawa ら [3] は，AMD は動画像を直接提示するシステムと比較して，プライバシの侵害のリスクが低いことや監視されていることに対する不快感が少ないことを明らかにした．これらのことから，被見守り者にとって AMD は動画像を使ったシステムと比較してより受容しやすい性質を持つことが示唆された．一方で，AMD では Kinect に含まれる関節位置角度の推定誤差が原因でアバタの動きに不自然さが生じてしまうため，見守り者と被見守り者の心理的近接性の増大の阻害や，細かな身体姿勢の識別を困難にしてしまうことが述べられた．そこで，AMD において関節位置角度の推定誤差を改善することで，システムの有用性の向上を試みる．

2　関連研究

カメラ画像から姿勢推定を行う研究も盛んに行われているが，本研究での姿勢推定エラー補正はアバタ媒介型見守りシステムへの応用を目的としているため，被見守り者のカメラ画像を取得されることへの抵抗感やプライバシの侵害などを考慮すると深度センサで取得された深度画像を使った姿勢推定手法の方が望ましい．

2.1　深度画像にもとづく姿勢推定

Plagemann ら [4] は，深度センサで取得された深度情報をもとに 3 次元メッシュを算出し，頭や手などの四肢の推定を通して姿勢推定を行った．この研究により，簡易的な深度センサを使って姿勢推定が可能であることが明らかにされたが，姿勢推定の精度が低いことや，手足の左右の識別が出来ないなどが指摘された．姿勢推定の精度を上げるために複数の角度から得られた画像をもとに姿勢推定を行う試みも行われているが [5][6]，事前に複数のカメラ間のキャリブレーションやフレームの同期などの設定を行う必要がある点や，あらかじめ用意されている 3D モデルを利用している点から，汎用性が低いことが示唆されている．

一方で，Shotton ら [7] は，現在 Kinect の姿勢推定アルゴリズムでは 1 台の深度センサのみでも高精度なマーカ付きモーションキャプチャシステムと比べて関節位置の平均二乗誤差が約 3 割以下の精度で予測可能であることを明らかにした．以上のことから導入コストを抑えつつ高精度な姿勢の推定手法として，Kinect

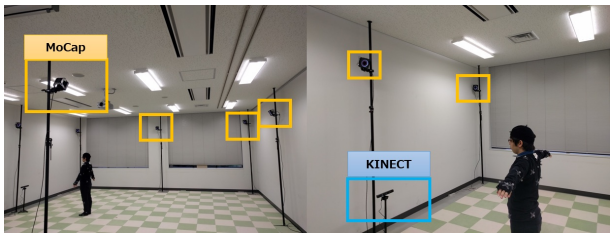

図 2: MoCap と KINECT を使ったデータセット作成の様子

によって取得された関節位置角度を利用することが望ましいと考えられる．

2.2　Kinect による姿勢推定結果の修正

Shen ら [8] は Kinect で得られるモーションを対象に Random Forest を用いて，Kinect で得られる関節位置を入力データ，高精度なマーカ付きモーションキャプチャシステムで取得した関節位置を正解データとして，Kinect に含まれる姿勢推定エラーの補正を行う学習コーパスの作成を行った．学習コーパスを用いて Kinect の姿勢推定エラーの補正前後での各関節位置の平均二乗誤差を比較したところ，約 3 割程度減少することが確認された．しかし，姿勢の推定エラーを補正することが出来るモーションが限定的である点や，モーションデータの連続性が考慮されていない点などが課題として挙げられた．そこで Park ら [9] は Shen らと同形式の入力データと正解データで，時系列データの学習に有効な Recurrent Neural Network (RNN) を用いた学習を行った．この結果，各関節の推定誤差の平均が約 5 割程度減少することが明らかにされ，Kinect によって取得される関節位置の修正には機械学習を用いた学習が有効であると考えられる．しかし，学習モデルを利用して得られた関節位置が人間にとって自然な動作になっているかについては考慮されていない．そこで本研究では学習モデルを使って得られた動作の自然さを検証するとともに，姿勢推定エラーの補正手法を見守りシステムに応用し，実証的に検証を行うことを目的とする．

3　提案手法

本研究では図 2 で示すような環境で，Microsoft Kinect for Window v2(Kinect) の姿勢推定結果に対して高精度なマーカ付きモーションキャプチャ (Motive) の認識結果を正解データとするコーパスを作成し，時系列データの学習に有効な RNN を用いて Kinect の認識結果を修正する姿勢推定モデルを構築することで，これを実現

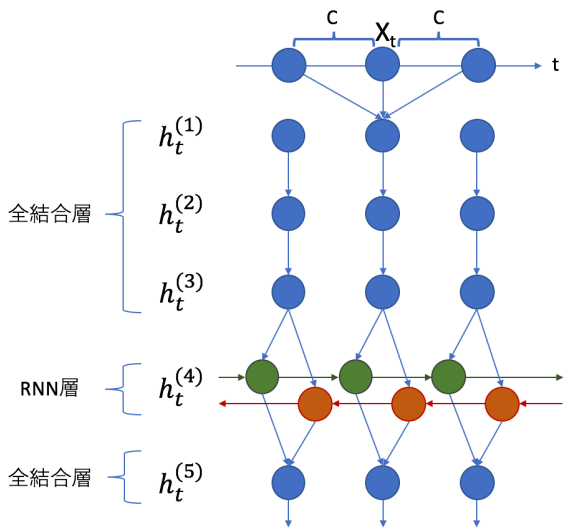

図 3: 学習に用いるネットワーク構造

する．RNN は CNN(Convolution Neural Network) のように各ステップにおいて入力を独立したものとして扱うのではなく，各学習データの前後データを含めて学習を行うため大量の学習データが必要であるが [9]，時系列データのような連続する要素の学習に向いている．

本研究で学習に用いるネットワーク構造を図 3 で示す．学習ネットワークは 5 層の隠れ層で構成され，1〜3 層目および 5 層目は全結合層，4 層目は Bidirectional RNN 層である．以下でデータセットの作成とネットワークの詳細を述べる．

3.1 データセットの作成

本研究では，訓練データとして Kinect で取得される全身 25 関節分の関節角度 $R_i^K(i:1...25)$ を利用する．これらは相対角度であり，各ボーンの親となるボーンの向きをベクトル $\vec{v}(x,y,z)=(0,1,0)$ とした時の回転角度を表したものである．正解データとして Motive で取得される全身 51 関節分の関節角度 (四元数) を利用する．Motive で取得される回転角度は，ボーンの初期状態 (T ポーズ) からの回転を示す点，四元数表現 (回転軸と回転角度) の軸表現が絶対座標系におけるものである点で Kinect と異なっている．Kinect の関節角度の修正を行う学習モデルを作成するため，Motive の関節角度を Kinect のフォーマットに合わせる必要がある．そのため前処理として，Kinect の関節角度に対応する Motive の関節角度を抽出する．その後，各関節角度ごとに補正用の四元数を用意し，親となるボーンの向きを $\vec{v}$ としたときの相対角度に変換し，正解データ R_i^M を作成する．こうして得られたデータセット $\mathrm{X}(R_i^K, R_i^M)_{i=1,...,25}$ を用いて学習を行う．

3.2 学習ネットワークの詳細

本研究で用いるネットワークでの隠れ層のユニット数は 256 とし，入力と出力のユニット数を 100 とした．これは Kinect で取得される関節数が 25 関節であり，各関節角度は (回転軸; 回転角度) $=(x,y,z;w)$ として表されるためである．またコンテキストは 10 とする．初めの 1〜3 層目および 5 層目の隠れ層では式 (1)，4 層目の RNN では式 (2)(3) のように学習を行う．また，1〜4 層目にはそれぞれ Batch Normalization と活性化関数として ReLU 関数，1〜5 層において 10%の Dropout を適用する．目的関数では，正解データと予測データの平均二乗誤差の計算を行う．

$$h_t^{(l)} = g(W^{(l)}h_t^{(l-1)} + b^{(l)}) \tag{1}$$

$$h_t^{(f)} = g(W^{(4)}h_t^{(3)} + W_r^{(f)}h_{t-1}^{(f)} + b^{(4)}) \tag{2}$$

$$h_t^{(b)} = g(W^{(4)}h_t^{(3)} + W_r^{(b)}h_{t+1}^{(b)} + b^{(4)}) \tag{3}$$

4 学習

4.1 学習セットアップ

本研究ではディープラーニングフレームワークの TensorFlow のラッパーライブラリである Keras を用いて実装を行った．データセットは図 2 に示す環境において，Microsoft Kinect SDK 2.0 を用いて取得された 30fps のデータと，OptiTrack 社の Motive Body 1.10 で取得された 120fps のデータを 30fps に間引いたものを用意した．2 つのモーションキャプチャシステム間の同期は，モーションの録画開始後に「手を挙げるモーション」など特徴的なモーションを基に手動で行った．また，モーションの取得範囲は Kinect v2 センサの取得範囲内 (0.5〜4.5m) に限定し，データセットの作成を行った．本研究では 353,799frame(約 196 分) のデータセットを作成し，学習用に 318,419frame，評価用に 35,379frame 用いて学習を行った．

4.2 学習結果および評価

学習の結果得られた学習モデルを用いて，Kinect による姿勢推定エラーの補正を行う．その結果を関連研究と比較することで評価を行う．Shen らの研究 [8] では式 (4) を用いて，Kinect の姿勢推定エラーの補正前後で関節位置誤差の総和の平均 (Average Position Error : APE) を算出している．本研究では入力データ，推定エラー補正後のデータ，正解データともに関節角度であ

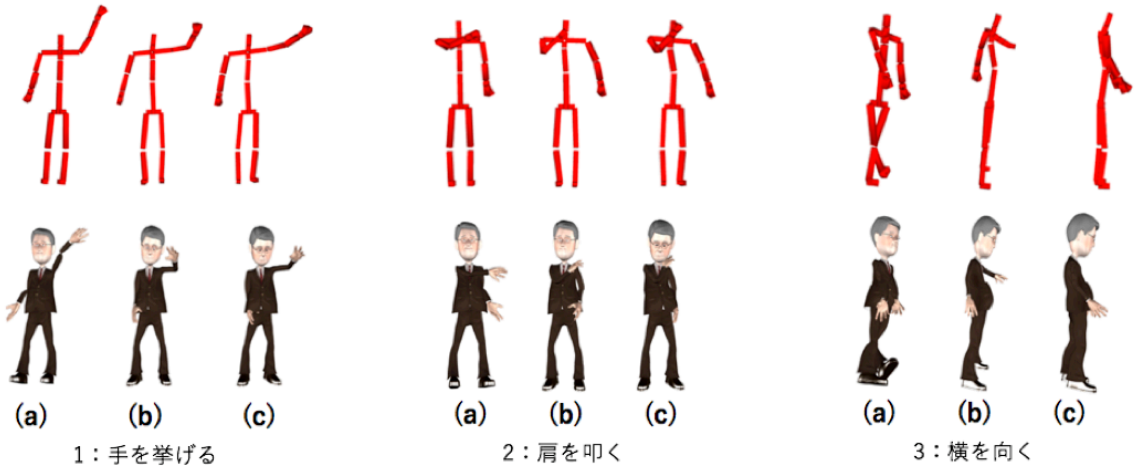

図 4: 姿勢の比較例（a：Kinect Skeleton，b：提案手法，c：Ground Truth）

表 1: 推定エラーの補正結果

比較	APE(meter)
Kinect Skeleton(25 関節)	3.186
提案手法 (25 関節)	2.854
Kinect Skeleton(20 関節)	1.883
提案手法 (20 関節)	0.662
Shen et al.[7](20 関節)	1.577

るため，式 (4) を用いて評価を行うためには関節位置に変換を行う必要がある．そこで Kinect で取得される関節位置をもとにアバタを作成し，関節角度をアバタに適用することにより関節位置を求めた．この結果得られた関節位置を使って，Shen らの研究と比較した結果を表 1 に示す．姿勢推定エラー補正の結果，Kinect v2 で取得される 25 関節分の APE，また 25 関節から関節位置の推定結果の信頼度が低い両手足・頭の 5 関節を除いた計 20 関節分の APE においてともに減少していることが確認された．また，Shen らの手法と比較しても高い精度で関節位置の補正が行えることが明らかにされた．

$$APE = \sum_{i=1}^{N} (\|R_i^M - R_i^K\|_2)/N \qquad (4)$$

5　考察

関連研究と比較した結果，本研究の学習モデルを用いた Kinect の姿勢推定エラー補正は高い精度で補正が可能であることが明らかにされた．これは時系列データを考慮に入れて学習を行ったことにより，自己遮蔽が生じた際でも前後のフレームをもとにエラーの補正が行えたと考えられる．また表 1 より，Kinect v2 で取得される 25 関節から関節位置の推定結果の信頼度が低い両手両足と頭の 5 関節を除いた計 20 関節分の APE が約 7 割減少している．これは大きい推定エラーが起きやすい関節 (両手足・頭などの末端に近い関節) を除外したことによる影響と考えられる．

また図 4 のようにアバタに補正前後の姿勢をそれぞれ適用し表示してみたところ，右肘の向きの修正が行えているのが見受けられる．今後，この学習モデルを用いて補正されたモーションの動画を作成し，モーションの自然さの主観評価を行く．また，学習に利用しているモーションには大きな動作が多いため，見守りシステムに導入するためには細かな動作についても学習を行う必要がある．加えて関連研究では Kinect v1 を利用しており，本研究では Kinect v2 を用いているため基本性能の差が補正結果に影響したことも要因の 1 つとして考えられる．そのため関連研究の手法を Kinect v2 で再現し比較する必要がある．

6　結論および今後の予定

本研究では安価なマーカレスモーションキャプチャシステムの Kinect に含まれる推定エラーを，人間工学的に自然な関節位置角度に修正する手法の開発を目的として研究を行った．Kinect の姿勢推定結果に対して高精度なマーカ付きモーションキャプチャシステムの認識結果を正解データとするコーパスを作成し，時系列データの学習に有効な RNN を用いて Kinect の認識結果を修正を行った結果，1 割程度誤差の修正を行うことが出来た．今後はより高い精度で誤差を修正することが出来るネットワークの検討や，修正されたモーションの主観評価を行っていきたい．

参考文献

[1] 杉原太郎, 藤波努, 高塚亮三. グループホームにおける認知症高齢者の見守りを支援するカメラシステム開発および導入に伴う問題. 社会技術研究論文集, Vol. 7, pp. 54–65, 2010.

[2] 長谷川大, 小林裕, 白川真一, 佐久田博司, 安彦智史, 安達栄治郎, 中山栄純. アバタ媒介型見守りシステムの開発. 知能と情報, Vol. 28, No. 6, pp. 974–985, 2016.

[3] Dai Hasegawa, Naoki Yokoyama, Hiroyuki Morikawa, Eijun Nakayama, and Hiroshi Sakuta. P-30 evaluation of avatar mediated distant-care system by the elderly. *The Japanese Journal of Ergonomics*, Vol. 53, No. Supplement2, pp. S754–S755, 2017.

[4] Christian Plagemann, Varun Ganapathi, Daphne Koller, and Sebastian Thrun. Real-time identification and localization of body parts from depth images. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3108–3113. IEEE, 2010.

[5] Carsten Stoll, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 951–958. IEEE, 2011.

[6] Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Tompson, Leonid Pishchulin, Micha Andriluka, Chris Bregler, Bernt Schiele, and Christian Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3810–3818, 2015.

[7] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, Vol. 56, No. 1, pp. 116–124, 2013.

[8] Wei Shen, Ke Deng, Xiang Bai, Tommer Leyvand, Baining Guo, and Zhuowen Tu. Exemplar-based human action pose correction. *IEEE transactions on cybernetics*, Vol. 44, No. 7, pp. 1053–1066, 2014.

[9] Youngbin Park, Sungphill Moon, and Il Hong Suh. Tracking human-like natural motion using deep recurrent neural networks. CoRR，abs/1604.04528，2016. http://arxiv.org/abs/1604.04528.