博士論文

話者中心の手話認識と翻訳に向けた全方位カメラを用いた装着型 MoCap システム

Wearable MoCap System with Omni-Camera towards Signer-Centered Sign Language Recognition and Translation

2023 年

三浦 哲平

論文要旨

手話は，ろう者がコミュニケーションで用いる主要な言語であり，他者との会話などの日常生活で使われている．しかし，健聴者で手話を習得している者は少なく，両者のコミュニケーションを支援する目的で手話の自動認識と翻訳に関する研究がおこなわれている．近年の，機械学習を用いた画像認識やパターン認識の技術発展に伴い，さまざまな手話認識・翻訳の手法が提案されているが，それらの多くは手話者の対面に設置したカメラで撮像した動画像を入力としている．対面カメラを用いた手話認識・翻訳では，一定の距離をおいたカメラの設置など，周囲の環境が整っていなければ利用できないという課題がある．手話は日常生活のコミュニケーションに使う対話言語であるため，認識・翻訳の研究とともに，常に利用できる仕組みも重要である．本研究は，ろう者自身が携帯し，常に利用できる手話の認識・翻訳システムの実現に向けた，全方位カメラを用いた装着型モーションキャプチャ (MoCap) システムに関する研究である．

手話は，手の形・位置・動きからなる手指動作，口形や眉の形といった非手指動作，そして話者の周囲の状況を組み合わせて表現する言語である．全周囲を撮像できる全方位カメラを身体に装着することで，持ち運び可能かつ常に手話表現の要素を捉えることが可能になる．本研究の主な貢献は，手話者の身体に装着した全方位カメラで撮像した画像から，装着者の身体動作（ 3 次元骨格）を推定する装着型 MoCap システムを提案することである．また，提案する装着型 MoCap システムを既存の手話認識モデルへ適用し，提案システムの有用性も確認する．本研究で提案する装着型 MoCap システムを用いた「話者中心の手話認識と翻訳」には，(1) 常時装着して利用することで膨大なデータを収集できる，(2) 利用者を限定することで手話認識・翻訳モデルを個人に最適化できる，(3) 対話相手を含めた周辺環境の情報を用いた手話認識・翻訳をできるといった利点があり，関連研究に大きく貢献できる可能性がある．

画像に撮像される人間の 3 次元骨格 (3D ポーズ) を取得する方法として，ニューラルネットワークを用いた 3D ポーズ推定がある．この手法は画像と 3D ポーズを一組とした学習用データセットを大量に必要とするが，本研究で提案する身体に装着した全方位カメラで撮像した画像と 3D ポーズのデータセットで一般に公開されているものは存在しない．そのため，第 3 章ではデータ収集の環境構築と，学習・評価用データの収集について述べる．さらに，既存の 3D ポーズ推定モデル VNect を用いて，収集したデータセットによる学習と評価をおこない，全方位カメラで撮像した歪みや切断を含む画像においても，データを収集することで装着者の自己 3D ポーズを推定できることを示す．

装着した全方位カメラで撮像した画像において，学習用データを収集することで自己 3D ポーズ推定が可能になることを明らかにしたものの，固有のカメラ設定における画像と 3D ポーズの収集は非常に負担の大きい作業である．第 4 章では，学習用データの収集とモデル学習の負担を軽減する手法について述べる．3D ポーズ推定において，

2D ポーズ推定部と 3D ポーズ拡張部を分離する 3D 単位ベクトル化を提案する．人工的に生成したデータセットを用いて学習と評価をおこない， 3D 単位ベクトル化を組み込んだ 3D ポーズ拡張モデルが学習データの収集とモデル学習の負担を軽減できることを示す．

上述の自己 3D ポーズ推定で得られた知見をもとに，手話の自動認識への適用に向けた全方位カメラを用いた装着型 MoCap システムのプロトタイプ開発をおこなう．第 5 章では，人工的に生成したデータセットを用いて， 3D 単位ベクトル化モジュールを組み込んだ 3D ポーズ推定モデルの学習と評価をおこない，提案する装着型 MoCap システムの自己 3D ポーズ推定における推定精度，頑健性，実行速度を評価する．さらに第 6 章では，装着した全方位カメラの摂動に対する自己 3D ポーズ推定モデルの頑健性について調査する．学習と評価に用いるカメラ摂動を含む合成データを生成し，いくつかの推定モデルついて比較・調査をおこない， 3D 単位ベクトル化を組み込んだ推定モデルがカメラ摂動に対する頑健性が高いことを示す．最後に，提案する装着型 MoCap システムを用いた既存の手話認識モデルへの適用について述べる．第 7 章では，装着した全方位カメラと通常の対面カメラで収録した手話動画から得られるそれぞれの画像を既存の手話認識モデルに適用し，対面カメラで撮像した画像を用いて学習した手話認識モデルと比較することで，全方位カメラで撮像した画像から自己 3D ポーズを用いた手話認識の有用性を検証する．

Abstract

Sign language is the primary language for people with hearing impairment to communicate with others in daily lives. However, most people with no hearing impairment do not learn sign language. Sign language recognition and translation research has been conducted to support people with hearing impairment and others. In recent years, machine learning technology has developed to propose sign language recognition and translation based on advanced image processing and pattern recognition. However, most methods place a camera in a front of the signer to capture the input image or movie. The method works for people with hearing impairment only in environments where the camera is placed far away. The method is important not only to research recognition and translation but also how to incorporate in daily situations because they use sign language in daily communications. We propose a wearable motion capture (MoCap) system equipped with omnidirectional camera towards sign language recognition and translation that people can wear and use in their daily situations.

Sign language consists of manual expressions (shape, position, movement), non-manual (mouth, eyebrow, etc.) expressions, and the surrounding environment. An omnidirectional camera mounted on the user's body can capture the elements of sign language expression in everyday situations. Our main contribution is to proposes a 3D human pose estimation using a wearable omnidirectional camera. Additionally, we apply the wearable MoCap system to existing sign language recognition models. "Signer-Centered Sign Language Recognition and Translation" using a wearable MoCap system has 3 advantages (1) Enables constant data collection with use, (2) Optimizes the sign language recognition and translation models for each user, and (3) Comprehensively processes environment information. These advantages have the potential to contribute to research field.

Deep neural network is a mainstream method for estimating 3D human pose on images. The method requires a vast training dataset consisting of images and 3D poses. However, training datasets for our unique camera settings is not publicly available. We build a data collection system and collect training and validation datasets in Section 3. Additionally, we use an existing model (VNect) to validate the performance of self-3D pose estimation with a wearable omnidirectional camera.

We verified the feasibility of self-3D pose estimation with a wearable omnidirectional camera using training dataset. However, collecting a dataset consisting of images and 3D poses with our unique camera setup is a time-consuming task. We propose to reduce the burden of data collection and model training in Section 4. The method decouples the model into 2D pose estimation and 3D lift-up using 3D unit vectorization. We validate the effectiveness of the model for data collection and training burden using a

synthetic dataset.

We develop a prototype wearable MoCap system with an omnidirectional camera for application in sign language recognition and translation. In Section 5, we first generate a synthetic dataset for training and validation. We evaluate the 3D pose estimation model with 3D unit vectorization for accuracy, robustness, and run time. Additionally, we evaluate the robustness of 3D pose estimation model for omnidirectional camera perturbations in Section 6. We generate a synthetic dataset with camera perturbations that increase in several steps. We indicate that the 3D pose estimation model with 3D unit vectorization has high robustness for camera perturbations since training and evaluation using the dataset. Finally, we apply the wearable MoCap system to an existing sign language recognition model in Section 7. We acquire sign language movies using the wearable omnidirectional camera and normal camera (third viewpoint), respectively. We compare existing models trained on each camera dataset. We validate the availability of wearable MoCap system to sign language recognition.

目 次

図 目 次

表 目 次

第1章

序論

1.1 研究の背景

2006 年, 国連総会において「障がい者の権利に関する条約」が採択され, 日本は 2014 年に同条約に批准した．これにより，障がい者の人権や自由を守るために，公的機関や会社などが「合理的な配慮」をすることが義務付けられた．こういった制度の変化の中で，誰もが活躍できる社会の実現を目指して，障がい者が抱えるさまざまな課題を解決するために情報技術が利用されている．

手話は，ろう者がコミュニケーションで用いる主要な言語であり，他者との会話などの日常生活で使われている．しかし，健聴者で手話を習得している者は少ないため，ろう者と健聴者のコミュニケーションは筆談や手話通訳者を介さなければならない．

ろう者と健聴者のコミュニケーションを支援することを目的として，手話の自動認識と翻訳に関する研究は以前からおこなわれている．近年では，ディープラーニングを用いた画像認識やパターン認識技術の発展にともなって，動画像を入力としたさまざまな手法 [1, 3, 4, 5, 6, 7] が提案されている．これらの手法は，手話者の対面に設置したカメラで撮像した画像を入力として，手話の自動認識や翻訳をおこなっている．

手話者を対面から撮像する手話の自動認識・翻訳システムは，窓口にカメラを設置することや，スマートグラスを装着することで，健聴者による手話の理解を補助することができる．一方で，ろう者にとっては周囲の環境が整っている状況でなければ利用できないといった課題がある．手話は日常生活のコミュニケーションに使う対話言語であるため，自動認識・翻訳の研究とともに，常に利用できる仕組みの研究も重要である．しかし，自動認識・翻訳の研究に比べて，そのような仕組みの検討は十分になされていない．

手話の自動認識・翻訳システムを，ろう者自身が携帯して常に利用できるシステムとすることで，周囲の環境に影響されずに，いつでも自発的，能動的に健聴者と対話することが可能になる．さらに，ろう者の主体的な発話を補助することにより，結果として積極的な社会参加を促すことも期待できる．

1.2 研究の目的

ろう者と健聴者が対話的に，直接に意思を伝え合うコミュニケーションを日常の生活でおこなうためには，携帯して常に利用できる手話の自動認識・翻訳システムが求められる．そのようなシステムを実現するためには，以下の技術について研究が必要である．

1. 手話の自動認識と翻訳に関する研究.

2. 携帯して常時利用できる身体動作情報の取得方法に関する研究.

本研究の主な貢献は (2) に関して，全方位カメラを用いた装着型モーションキャプチャ (MoCap) システムを提案することである．また，提案する装着型 MoCap システムを既存の手話認識モデルへ適用して，提案システムの有用性を確認する．

手話は，主に手指動作と非手指動作で表現する言語である．手指動作とは，手と指で表される手形，手の位置，動きのことである．また，非手指動作とは，口形や眉の形，視線などの手指以外の動きのことである．さらに，手話は周囲の状況を利用して表現することもある．たとえば，ひとさし指を出して何かを指し示す手指動作をする場合を考える．指の先が手話者自身を指し示すと「わたし」を表し，他者を指し示すと「あなた」やその個人を表す．また，物や場所を指し示すと，その物や場所自体を表す．このように，手話は手指動作，非手指動作を用いながら，周辺の空間を使って表現される．

既存の手話の自動認識・翻訳に関する研究 [8] では，まず入力データから手話者の動作を取得して，次に入力動作を処理することで手話の自動認識・翻訳をおこなう．手話者の動作を取得する方法は主に以下の 2 つに分類できる．

- ジャイロセンサ等のついた手袋や衣服を身体に装着して，機器の値の計測，処理によって骨格の位置を取得する方法.

- RGB カメラを用いて画像を取得する方法.

手袋や衣服を装着する方法は，携帯性が高く，動作を常に取得できるといった利点がある．しかし，ジャイロセンサが磁気の影響を受けやすく，長時間安定したデータを取得することは難しい．また，非手指動作まで取得しようとするとセンサ数が増えるため，すべてを装着して計測することは不可能である．

RGB カメラを用いた方法は，カメラ 1 台で動作を取得できるため，導入が容易であり，身体に装着する煩わしさがないといった利点がある．一方で，手話者から離れた

位置に設置して，身体が収まるように画像を取得しなければならないため，携帯性が低く，常に身につけて使うことができないといった課題がある．

本研究では，携帯性と装着の負荷を考慮して，小型の全方位カメラを用いた装着型 MoCap システムを提案する．全方位カメラはふたつの超広角レンズによって，その全周囲を撮像できるため，身体に近接したカメラ位置からでも，手話者の上半身の身体動作情報を取得できる．さらに，ディープニューラルネットワークを用いた 3 次元上の骨格推定モデルを使って，身体に装着した全方位カメラの画像から自己 3D ポーズを推定することで，携帯して常に利用できる装着型 MoCap システムとなる．

1.3 話者中心の手話認識と翻訳の利点

本研究で提案する全方位カメラを用いた装着型 MoCap システムは，利用者の身体動作情報と共に，手話表現に必要な周辺空間の画像情報を取得できる．既存研究で多く見られる手話者の対面に設置したカメラで撮像した動画像を用いた手話認識と翻訳に対して，提案システムを装着者した手話者の身体動作情報と周辺画像を用いる方法を「話者中心の手話認識と翻訳」と定義する．提案システムを用いた話者中心の手話認識・翻訳には以下の利点があり，関連研究に大きく貢献できる可能性がある．

- 常時装着して，利用することで膨大なデータを収集できる．
- 利用者を限定することで手話認識・翻訳モデルを個人に最適化できる．
- 対話相手を含めた周辺環境の情報を用いた手話認識・翻訳をできる．

近年の手話認識・翻訳は，機械学習，特にニューラルネットワークを用いたモデルによる研究が主流である．このような手法では，モデルの設計とともに，学習に用いるデータセットの拡充が非常に重要である．しかし，学習用データとして利用できるアノテーション付きの手話データは非常に少ない．日本国内最大規模の手話データベースである KoSign [1] でも，男女 1 名ずつのおよそ 6,000 単語と，限定された数対話のみの収録となっており，精度と汎化性をもった手話認識・翻訳モデルを構築するためにはデータが不足している．装着型 MoCap システムを利用して手話認識・翻訳をおこなうことで，同時に膨大な手話データを収集することが可能になる．

日本語などの音声言語に方言や，個人によるイントネーションや言い回しの特徴があるように，手話においても同じ文意・単語で個人によって表現の特徴がある．特に手話においては個人による差異が大きく，これは音声言語と比較して限られたコミュ

[1] KoSign: https://www.nii.ac.jp/dsc/idr/rdata/KoSign/

ニティ内で手話の習得，対話をしていることが要因のひとつと言われている．一般に，手話者によって表現の差異がある場合には，すべての人に対応した認識・翻訳モデルを構築するよりも，ある個人の表現のみに対応するモデルを構築する方が容易である．提案システムを用いた話者中心の手話認識・翻訳では利用者を限定できるため，モデルを個人に最適化することで精度の高い手話の認識・翻訳が可能になる．

手話は手指動作，非手指動作を用いながら，話者の周辺空間を使って表現される．たとえば，人差し指を伸ばして指し示す表現では，何を指し示すかによって手話表現の意味が変化する．また，音声言語がそうであるように，手話も対話相手や発話前の文脈によって認識・翻訳の内容が変化する．つまり，手話の認識・翻訳をおこなうためには，対話相手を含めた周辺環境の情報を扱う必要があり，提案する装着型 MoCap システムを用いることで，それらの情報を含んだ手話認識・翻訳が可能になる．

1.4 本論文の構成

第 2 章では，関連研究として「手話の自動認識と翻訳」と「身体に装着した魚眼カメラを用いた自己 3D ポーズ推定」に関する研究について述べる．手話の自動認識と翻訳については，近年の主流な手法となっている画像と機械学習を用いた手法について述べる．ポーズ推定については，画像認識を用いた 2D / 3D ポーズ推定について述べた後に，本研究に特に関係する魚眼カメラを用いた自己 3D ポーズ推定について述べる．

第 3 章では，身体に装着した全方位カメラを用いた上半身の自己 3D ポーズ推定について述べる．本研究で提案する手法に適合するデータセットで一般に公開されているものは存在しないため，データ収集の環境を構築し，学習・評価用データの収集をおこなう．既存の 3D ポーズ推定モデル VNect [9] を用いて学習と評価をおこない，歪みや切断を含む全方位カメラの画像においてもデータを集めることで，自己 3D ポーズ推定をおこなえることを示す．

第 4 章では，全方位カメラを用いた自己 3D ポーズ推定において， 2D ポーズ推定部と 3D ポーズ拡張部を分離可能な 3D 単位ベクトル化について述べる．3D 単位ベクトル化を用いることで， 3D ポーズ推定モデルの学習だけではなく，学習データの収集においても 2D ポーズ推定部と 3D ポーズ拡張部を分離できることを示す．これによって，自己 3D ポーズ推定において特に大きい負担となる学習データ収集の負荷を軽減できることを示す．

第 5 章では，手話の自動認識に適用するための全方位カメラを用いた装着型 MoCap システムのプロトタイプ開発について述べる．人工的に生成したデータを用いて 3D

ポーズ推定モデルの学習と評価をおこない，提案する装着型 MoCap システムの自己 3D ポーズ推定における精度と頑健性を評価する．

第 6 章では，装着した全方位カメラの摂動に対する自己 3D ポーズ推定モデルの頑健性について述べる．学習と評価に用いるカメラ摂動を含む合成データを生成し，いくつかの推定モデルについて学習と評価をおこない，3D 単位ベクトル化を組み込んだ推定モデルがカメラ摂動に対する頑健性が高いことを示す．

第 7 章では，全方位カメラを用いた自己 3D ポーズ推定を用いた既存の手話認識モデルへの適用について述べる．提案システムと通常の対面カメラでそれぞれに収録した手話の画像を既存の手話認識モデルに適用し，提案システムの有用性を確認する．

最後に第 8 章では，本研究のまとめとして，本研究で得られた知見と今後の課題について述べる．

第2章

関連研究

2.1 序言

本章では関連研究について述べる．まず 2.2 節では，本研究の応用先である手話の自動認識と翻訳に関する研究について述べる．次に 2.3 節では，画像認識を用いた 2D / 3D ポーズ推定について述べた後に，本研究に特に関係する魚眼カメラを用いた自己 3D ポーズ推定について述べる．

2.2 手話の自動認識と翻訳に関する研究

手話は主に手指動作と非手指動作で表現する言語である．手指動作とは，手と指で表される手形，手の位置，動きのことである．また，非手指動作とは，口形や眉の形，視線などの手指以外の動きのことである．さらに，手話は周囲の状況を利用して表現することもある．たとえば，ひとさし指を出して何かを指し示す手指動作をする場合を考える．指の先が手話者自身を指し示すと「わたし」を表し，他者を指し示すと「あなた」やその個人を表す．また，物や場所を指し示すと，その物や場所自体を表す．このように，手話は手指動作，非手指動作を用いながら，周辺の空間を使って表現される．

実際の環境下で手話の認識と翻訳をおこなうためには，単語をつなげた文として連続的に表される手話の認識が求められる．しかし，以下の理由から，独立した手話単語を認識することに比べて，文としての連続的な手話を認識することは非常に難しい．

- ある単語の動きの中に，別の単語の動きの特徴が現れることがある．
- 前の単語の終わりの手形や位置が，次の単語の始まりに影響を与える．
- スラングや抑揚など，個人によって手話の特徴が異なる．

手話の自動認識と翻訳をおこなうには，まずはじめに手話者の身体動作情報を取得する必要がある．身体動作情報を取得するには大きく分けて 2 つの方法がある．

- ジャイロセンサ等のついた手袋や衣服を身体に装着して，機器の値の計測，処理によって骨格の位置を取得する方法.

- RGB カメラを用いて画像や深度を取得する方法.

デバイスを装着して情報を取得する手話認識の方法として，1983 年にワイヤと回路で作成された電子グローブ [10] を用いたアメリカ手話 (ASL) の指文字認識が提案された．その後も，さまざまなセンサを手袋や衣服などに装着して，手話認識をおこなう手法が提案されてきた．手袋や衣服を装着する方法は，携帯性が高く，動作を常に取得できるといった利点がある．しかし，これらの方法はセンサが磁気などの影響を受けやすく，長時間安定したデータを取得することが難しい．また，手話者にとって装着する負担があることや，非手指動作まで取得しようとするとセンサ数が増えるため，全てを装着して計測することは不可能といった課題がある．

近年の主流の手法は動画像を用いた手話の自動認識と翻訳である．1988 年に Tamura ら [11] は，画像認識を用いて日本手話の独立した 10 単語を認識するシステムを提案した．その後も，画像認識を用いたさまざまな手話認識システムが提案されている．動画像を用いる方法は，手話者が身体動作情報を取得するためのデバイスを装着する煩わしさがないことや，非手指動作も使った手話認識と翻訳ができるといった利点がある．一方で，手話者から離れた位置に設置して，身体が収まるように画像を取得しなければならないため，携帯性が低く，常に身につけて使うことができないといった課題がある．

近年のディープニューラルネットワークや畳み込みニューラルネットワークといった機械学習の発展にともなって，動画像を用いたさまざまな手話の自動認識と翻訳の方法 [1, 3, 4, 5, 6, 7] が提案されている．しかし，これらの機械学習を用いた方法では，膨大な量の学習用データセットが必要となるため，実際の環境下での動作はいまだ限定的である．Cui [3] らの手話認識モデルは，連続的な手話動画のベンチマークとなっているデータセット RWTH Phoenix [12] において，同じ人物を学習とテストに用いた場合に WER 22.9% (Word Error Rate：単語の誤認識率)，異なる人物を学習とテストに用いた場合に WER 39.6% であった．RWTH Phoenix は，連続的な手話動画の認識で用いられるデータセットのうちでは充実したものであるが，含まれている単語の種類はおよそ 1000 語程度である．

画像認識を用いた手話の自動認識と翻訳において，画像から特徴量を抽出する方法について，いくつかのモデルが提案されている．近年，多く用いられているのは画像全体をそのまま入力として End-to-End で学習をおこなうモデルである [3, 4, 5]．一方で，まず画像から上半身の 3D ポーズ，手形，口形の特徴量を抽出し，それらを入力

とするマルチチャネルモデル [6] も提案されている．このマルチチャネルモデルでは，手話言語の定義に応じて入力画像を 3D ポーズ，手形，口形に分類してからモデルへの入力とするため，End-to-End モデルに比べて，手話の言語学的な知見が含まれたモデルと言える．また，マルチチャネルモデルは手話認識を言語的に説明可能ないくつかの問題に分割でき，それぞれの課題に個別に取り組むことが可能といった利点がある．Li ら [1] は，独自に収集した大量の ASL 手話単語動画を用いて， End-to-End モデルと 2D ポーズを入力とするモデルの比較評価をおこない，そこでは End-to-End モデルを用いた手話認識が高い推定精度を示している．

2.3 魚眼カメラを用いた自己 3D ポーズ推定に関する研究

カメラの設置方法に着目して，単眼カメラを用いた人間の 3D ポーズ推定の関連研究について述べる．まず，もっとも一般的な状況である離れた位置にカメラを設置して撮像した画像を用いた 3D ポーズ推定について述べる．次に，身体に小型カメラを装着して撮像した一人称視点画像を用いた自己 3D ポーズ推定について述べる．最後に，本研究に最も関連する，広角な画像を撮像できる魚眼カメラを装着しておこなう自己 3D ポーズ推定について述べる．

近年, 畳み込みニューラルネットワークと大規模な 2D / 3D ポーズのアノテーション付きデータセットの充実によって，単眼カメラ画像を用いた 3D ポーズ推定に関する研究は大きく発展してきた．3D ポーズ推定には主にふたつのアプローチがある．ひとつは，画像から直接に 3D ポーズを推定する直接回帰アプローチ [9, 13, 14, 15, 16, 17, 18] である．もうひとつは，画像から 2D ポーズを推定する部分 (2D ポーズ推定部) と，画像平面上の 2D ポーズから 3D ポーズを推定する部分 (3D ポーズ拡張部) にタスクを分割するパイプラインアプローチ [19, 20] である．

直接回帰アプローチの精度と汎化性は，実環境下での 3D ポーズのアノテーション付きの画像が，どの程度学習データとして利用できるかに大きな影響をうける．推定を 2 ステップに分割するパイプラインアプローチにはふたつの利点がある．ひとつめは， 2D ポーズ推定部に既存の高精度な 2D ポーズ推定モデルを利用可能なことである．このとき，学習に必要な 2D ポーズのアノテーション付き画像は， 3D ポーズと比較して非常に収集が容易なことも注目すべき利点である．ふたつめは，3D ポーズ拡張部の推定モデルを学習するために入力画像を必要とせず，平面上の 2D ポーズ (関節の位置) と 3D ポーズのデータセットで学習可能なことである．Martinez ら [19] はパイプラインアプローチにおいて， 3D ポーズ拡張部はシンプルで，軽量なモデルであっても高精度に推定できることを示した．

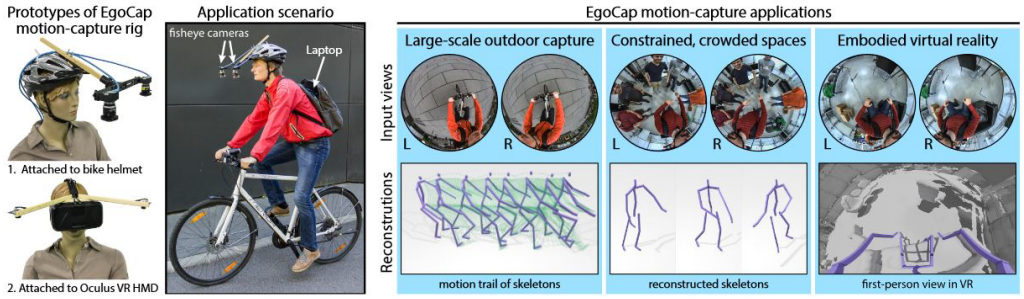

図 2.1　Rhodin らの提案した魚眼カメラを用いた 3D ポーズ推定.

身体に装着した魚眼カメラで撮像した画像から，全身のポーズ推定をおこなう最初のアプローチは Rhodin ら [21] によって提案された．ヘルメットに伸縮可能な棒を取り付けることで，装着者の頭部からおよそ 25 cm 離れた位置に 2 つの魚眼カメラを設置した．その広角な画角によって装着者の身体全体を撮像することはできたものの，そのデバイスは装着者にとって非常に負担になるものであった (図 2.1) .

軽量・小型・単眼の魚眼カメラを用いた手法が Xu ら [22] と Tome ら [23, 24] によって提案された．Xu らは野球帽のつばに下向きに魚眼カメラを取り付けた．Tome らはヘッドマウントディスプレイに下向きに魚眼カメラを取り付けた．どちらの研究においても，魚眼カメラの光学特性による歪みを含んだ，固有のカメラ位置からの 3D ポーズアノテーション付き画像が必要になる．そのデータ不足を補うために大規模な合成データを生成し，学習データとして利用した.

Xu ら [22] は，直接回帰アプローチでカメラ位置から 3D ポーズを構成する各関節座標への 3D 単位ベクトルと距離を推定し，それらを組み合わせることで 3D ポーズを推定するモデル (Mo^{2}Cap^2) を提案した (図 2.2) ．3D 単位ベクトルは，推定した 2D ポーズの関節位置を入力として，魚眼カメラのパラメタとキャリブレーションツール ocamcalib [25] を使って取得した.

Tome ら [23, 24] は， 3D ポーズ拡張部にマルチブランチ・エンコーダデコーダモデル (xR-EgoPose) を使ったパイプラインアプローチを提案した．xR-EgoPose は 2D 関節位置のヒートマップから 3D ポーズの関節座標を推定する (図 2.3) ． この 3D ポーズ拡張部の推定モデルは，入力画像を必要とせず， 2D ポーズを構成する関節位置のヒートマップと一貫性のある 3D ポーズのデータセットで学習することができる．2D ポーズ推定部から出力するヒートマップは，2D ポーズの関節位置以外のオクルージョンなどの情報を含むため，ヒートマップと 3D ポーズを用いて学習する 3D ポーズ拡張部は複雑な姿勢に対する頑健性が高くなる.

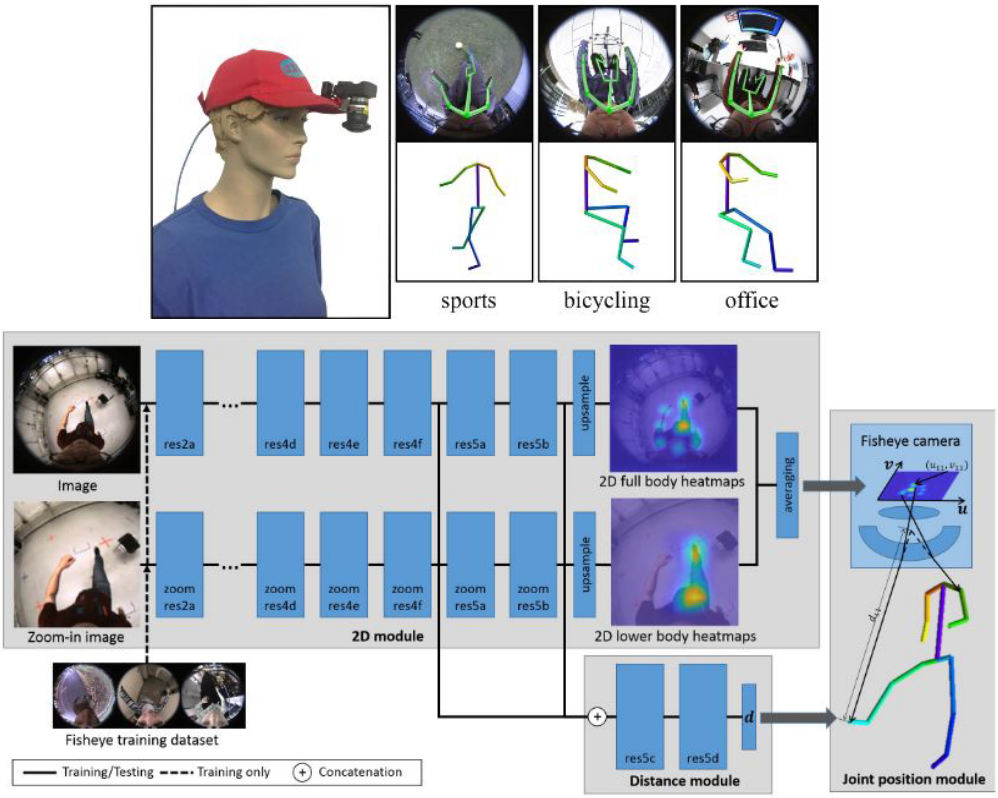

図 2.2　Xu らの提案した魚眼カメラを用いた 3D ポーズ推定.

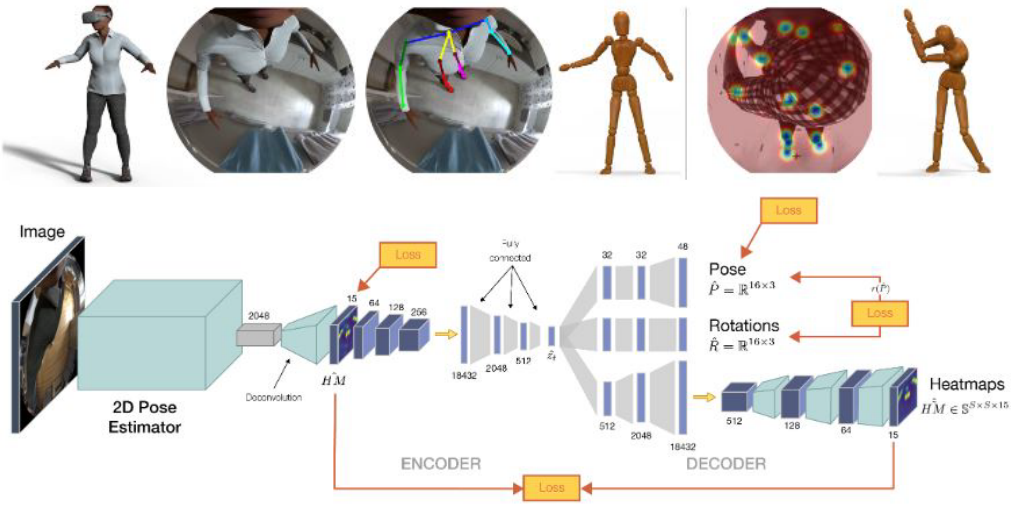

図 2.3　Tome らの提案した魚眼カメラの用いた 3D ポーズ推定.

第3章

全方位カメラを用いた自己 3D ポーズ推定

3.1 序言

本章では，身体に装着した全方位カメラで撮像した画像を用いた自己 3D ポーズ推定について述べる．3.2 節では，身体に装着する全方位カメラと撮像する画像について述べる．3.3 節では，学習と評価に用いるデータの収集方法とデータセットの構築について述べる．3.4 節では，自己 3D ポーズ推定に用いる推定モデルについて述べる．3.5 節では，収集したデータセットを用いた推定モデルの学習と評価について述べる．最後に，3.6 節で本章についてまとめる．

3.2 装着する全方位カメラと撮像される画像

身体に装着したカメラで撮像した画像から自己 3D ポーズ推定をおこなう場合，設置位置の近さから通常のカメラではその画角内に身体全体を撮像することが難しいためポーズ推定の精度は低くなる [26]. 先行研究 [22, 23, 24] では，広角な魚眼カメラを用いることで，近接したカメラ位置からでも身体の大部分を撮像し，精度の高い自己 3D ポーズを推定する方法を提案した．しかし，提案されている方法の魚眼カメラは単眼であるため，依然として画角に制限があり，手などの動きの大きい部位を撮像できない場合がある (図 2.2 ，図 2.3)．たとえば，手話で頭部を触る単語を表す場合，先行研究のカメラ設定では手を画像内に捉えられず， 3D ポーズ推定が失敗する可能性がある．

本章では，市販のネックマウントに全方位カメラ (Ricoh R Development Kit [1]) を取り付けることで，首の前方あたりに位置するレンズから撮像される画像を用いて，自己 3D ポーズ推定をおこなう．全方位カメラは背面同士に設置された 2 つの魚眼カメラで構成されており，その周囲全体を撮像することができる．全方位カメラを用いる

[1]Ricoh R Development Kit: `https://ricohr.ricoh/en/`

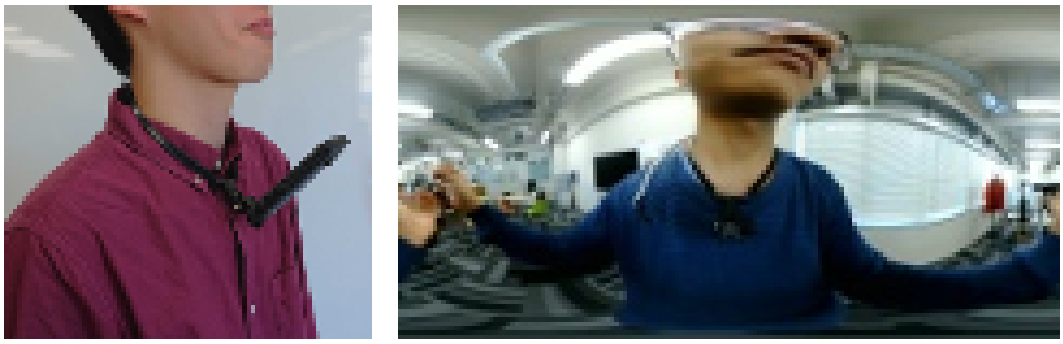

図 3.1　首の前方に装着した全方位カメラと撮像される正距円筒画像.

ことで，より幅広い動作について顔や手などを撮像することが可能になる．全方位カメラを装着した様子と撮像される画像の例を図 3.1 に示す．本章で用いる全方位カメラで撮像される正距円筒図法の画像には，通常の画像とは異なる以下の特徴がある．

- 歪み：画像の上部と下部が横に広がるように歪む.
- 分断：全周囲のどこかが境界となり画像の両端に分断される (たとえば装着者の腕が横端で切断されて，反対の横端から続きの腕が表示される場合などがある) .

3.3　学習と評価用のデータ収集

3D ポーズ推定モデルの学習と評価をおこなうために， 3D ポーズのアノテーション付き正距円筒画像のデータセットを構築する．以下の (1) と (2) のデータを同期して収集する．(1) 市販のネックマウントを使って，首の前方あたりに装着した全方位カメラで，正距円筒図法の画像を撮像する．(2) 全方位カメラを装着した人の対面に RGB-D カメラを設置し，深度データを用いて装着者の 3D ポーズを推定する．

収集するデータのフォーマットは以下である．

正距円筒画像： 全方位カメラを装着して，装着者が直立した状態でカメラの軸を地面に垂直になるように設定してから正距円筒画像を撮像する．これによって，体が傾いたとしても，身体の中心の関節 (head, neck, torso, waist) は常に画像の中心線上に射影される．

3D ポーズ： 3D ポーズを構成する上半身の 12 関節（head, neck, torso, waist, shoulders, elbows, wrists, hands）の 3D 空間の座標を取得する．

3D ポーズを取得する過程で，データの扱いを簡単にするために以下の 1 から 4 の正規化をおこなう．

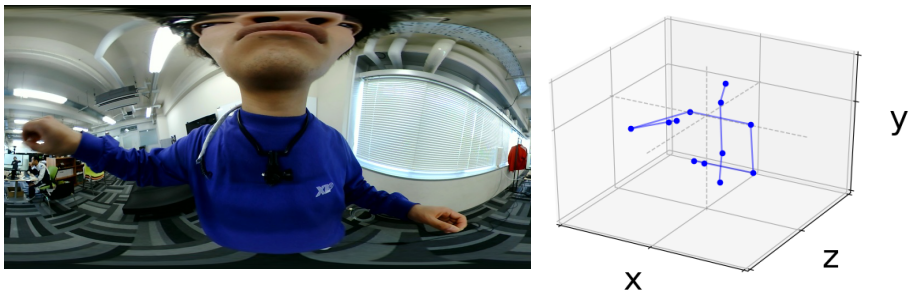

図 3.2　収録した正距円筒画像と 3D ポーズの例.

1. 全方位カメラの位置が 3D 空間の原点となるように，3D ポーズを構成する各関節座標を移動する．このとき，カメラの位置 (原点の位置) は実際の装着位置に従って決めており，首の関節位置から前方 (肩幅 ×0.4)，下方 (肩幅 ×0.1) の位置に固定する．
2. 両肩を結ぶ直線が x 軸と並行になるように，y 軸を中心に 3D ポーズを回転する．
3. 関節 neck と torso を結ぶ直線が y 軸と並行になるように，x 軸を中心に 3D ポーズを回転する．
4. 両肩の幅 (ユークリッド距離) が 1.0 となるように，3D ポーズを拡大，もしくは縮小する．

収集する正距円筒画像と，同期した 3D ポーズの例を図 3.2 に示す．このとき，2D 平面上の 2D ポーズの各関節位置は，全方位カメラの位置を原点とした 3D 関節の座標と正距円筒図法の投影式によって導かれる．

3D ポーズを取得するシステムは以下の機器で構成される．

- RGB-D カメラ： Intel RealSense Depth Camera D435 [2]
- 骨格推定ソフト： Nui Track ver 1.3.5 (on Windows x64) [3]

身体に全方位カメラを取り付けた装着者の対面に RGB-D カメラを設置する収録システムの構成を図 3.3 に示す．収録システムでは，全方位カメラで正距円筒画像を撮像すると同時に，RGB-D カメラと骨格推定ソフトを用いて 3D ポーズを取得する．取得した 3D ポーズは上述したフォーマットに従って正規化される．データの収録中に，装着者の動作によってオクルージョンが発生し，骨格推定ソフトを利用した 3D ポー

[2]https://www.intelrealsense.com/depth-camera-d435/
[3]https://nuitrack.com/

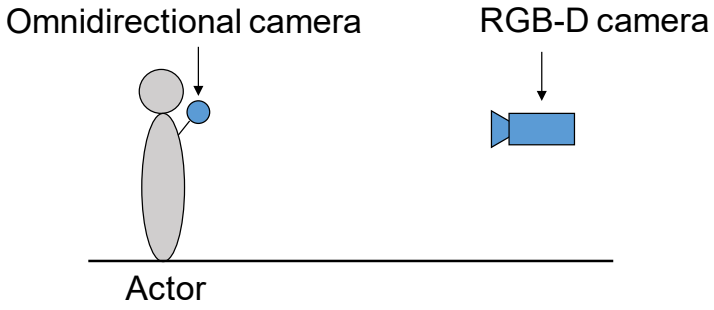

図 3.3　データ収録システムの全体の構成.

表 3.1　学習，検証，テスト用のデータセット.

Dataset	test	training	validation
A	18,050	94,713	18,943
B	17,045	95,550	19,111
C	19,754	93,293	18,659
D	17,695	95,009	19,002
E	19,320	93,655	18,731
F	20,627	92,565	18,514
G	19,215	93,742	18,749

ズの取得に失敗する場合がある．収録データの品質を保つため，3D ポーズの収録対象の全ての関節の推定が成功したときのみ，正距円筒画像と 3D ポーズを記録する．すなわち，本システムで収録するデータの時系列は不連続となっている．本章で扱う 3D ポーズ推定モデルにおいては，時系列の連続性は考慮せず，独立したフレームごとのデータとして学習，推定をおこなうため大きな問題にはならない．

本章では，日本手話に関するデータを収集する．手話・日本語大辞典 [27] にて定義される手の位置と動きを網羅する 16 例文を作成する．その例文は， 22 種類の手の位置すべて， 7 種類の手の動きすべて，そして 59 種類の手形のうち 38 種類を含んでいる．手形については，手の位置・動きを網羅する例文を作成するなかで，文意を保ちながら可能な限り異なる手形の単語を選んだ．例文は，75 種類の手話単語を含んでいる．

前述した 16 例文を使って 7 人の協力者から手話のデータを収録した．協力者は各例文を 3 回，もしくは 4 回おこなう．衣服，眼鏡の着用などの装飾品の指定はない．収録したデータについて，協力者ごとに分類し，さらに学習，検証，テスト用のデータが 5:1:1 になるように分類する．構築したデータセットを表 3.1 に示す．表中の数値は 3D ポーズのアノテーション付き画像の数を示す．Dataset 列のアルファベットは協力者の ID を示す.

図 3.4　OpenPose を用いた 2D ポーズ推定結果（左図）と提案手法の推定結果（右図）.

3.4　自己 3D ポーズ推定モデル

近年，畳み込みニューラルネットワークを用いた，さまざまな 3D ポーズ推定モデルが提案されている．しかし，既存の学習済み推定モデルは通常のカメラを用いて，離れた位置にいる人間を撮像した画像を対象としているため，身体に装着した全方位カメラで撮像された歪みと分断を含む正距円筒画像に用いることはできない．実際に，3D ポーズ推定よりも容易とされる 2D ポーズ推定であっても，高精度の 2D ポーズ推定をおこなう OpenPose [2] を用いた推定に失敗する (図 3.4)．図中の左図は OpenPose を用いた 2D ポーズ推定結果を示しており，左手首と鼻の推定に失敗していることがわかる．右図は，本章の手法によって推定した結果であり，画像上で分断されている左手首 (l-wrist) と左手 (l-hand) を含めてポーズ推定できている．なお，右図は 3D ポーズ推定結果を正距円筒図法の投影式を用いて 2D 平面上に置換して表している．

本章では，Mehta ら [9, 18] の提案したロケーションマップ法を用いた推定モデルを用いる．ロケーションマップ法は，歪みや分断を含まない通常の画像に対する 3D ポーズ推定モデルとして提案されたものである．しかし，2D 平面上の 2D ポーズの関節位置と 3D ポーズの関節座標との一貫性 (2D-3D 一貫性) を考慮した推定モデルであり，歪みや分断を含む正距円筒画像を入力とした自己 3D ポーズ推定にも用いることができる．

ロケーションマップ法は，画像平面上の 2D-3D 一貫性に着目し， 2D ポーズの関節位置から 3D ポーズの関節座標を推定する方法である．VNect [9] では，畳み込みニューラルネットワークを用いて，入力画像からヒートマップ $\boldsymbol{H}$ とロケーションマップ $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ の 4 つのマップを出力する．そして，出力した 4 つのマップから推定結果の 3D ポーズを得る．

ヒートマップは 2D 平面上の関節位置を表す二次元確率分布となっており，値の最も高い位置が 2D 関節位置として推定される．3D ポーズの関節座標は，ヒートマップによって決定された 2D 平面上の位置と同じ位置のロケーションマップの値によって得

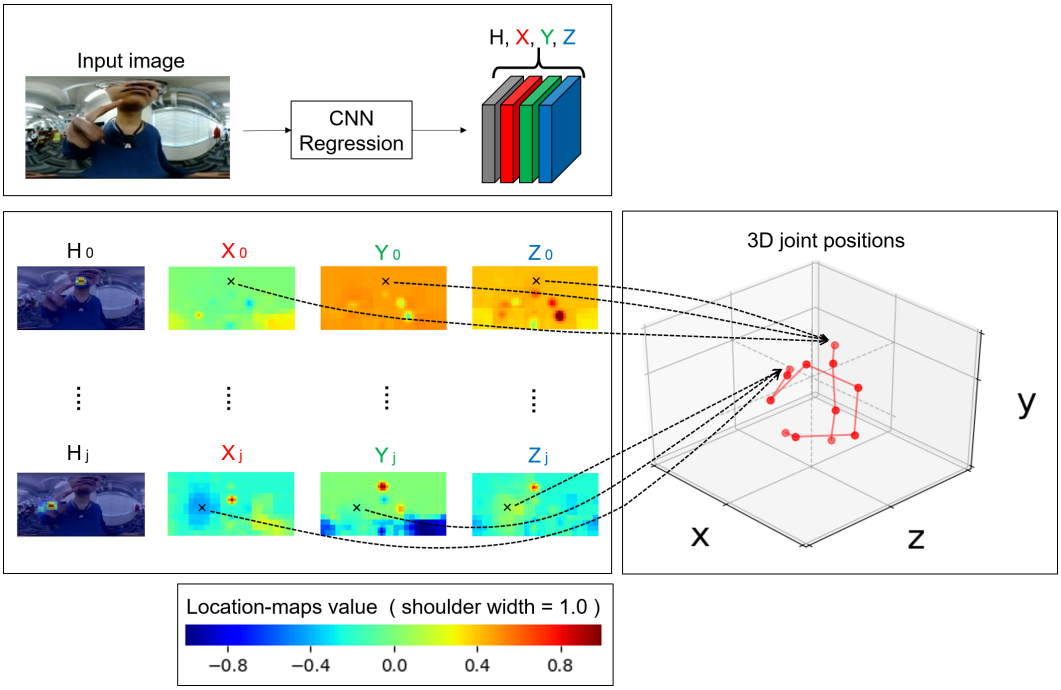

図 3.5　ヒートマップとロケーションマップを用いた 3D ポーズ推定モデルの全体像.

られる．3D ポーズの各関節 $j \in \boldsymbol{J}$ の 3D 空間での座標 x_j, y_j, z_j は次の式で表される．

$$
\begin{aligned}
row_j, col_j &= \mathrm{argmax}(\boldsymbol{H}_j), \\
x_j &= \boldsymbol{X}_j(row_j, col_j), \\
y_j &= \boldsymbol{Y}_j(row_j, col_j), \\
z_j &= \boldsymbol{Z}_j(row_j, col_j)
\end{aligned}
$$

$\boldsymbol{H}_j, \boldsymbol{X}_j, \boldsymbol{Y}_j, \boldsymbol{Z}_j$ は，関節 $j \in \boldsymbol{J}$ に対応するヒートマップとロケーションマップを表す．また， argmax はヒートマップの最大値となる 2D 平面上の位置 (行，列) を出力する関数である．ヒートマップとロケーションマップを用いた 3D ポーズ推定モデルの全体像を図 3.5 に示す．

学習時には，ヒートマップ $\boldsymbol{H}_j$ は画像平面上の関節位置を表す 2D ガウシャンマップとの L2 誤差で学習をおこなう．そして，ロケーションマップ $\boldsymbol{X}_j, \boldsymbol{Y}_j, \boldsymbol{Z}_j$ は以下の式で表される L2 誤差関数で学習をおこなう．

$$
\begin{aligned}
\mathrm{Loss}(h_j) &= ||\boldsymbol{H}_j - \boldsymbol{H}_j^{GT}||_2, \\
\mathrm{Loss}(x_j) &= ||\boldsymbol{H}_j^{GT} \otimes (\boldsymbol{X}_j - \boldsymbol{X}_j^{GT})||_2, \\
\mathrm{Loss}(y_j) &= ||\boldsymbol{H}_j^{GT} \otimes (\boldsymbol{Y}_j - \boldsymbol{Y}_j^{GT})||_2, \\
\mathrm{Loss}(z_j) &= ||\boldsymbol{H}_j^{GT} \otimes (\boldsymbol{Z}_j - \boldsymbol{Z}_j^{GT})||_2
\end{aligned}
$$

GT は教師データであることを示し，$\otimes$ はアダマール積を示す．$\boldsymbol{X}_j^{GT}$ は教師データ x_j の値の一様分布である．ロケーションマップの損失関数は，ヒートマップの教師データ $\boldsymbol{H}_j^{GT}$ によって， 2D 関節位置の周辺の誤差に重み付けして 3D 関節座標を学習していると解釈できる．

ロケーションマップ法のもっとも重要な特徴は， 2D ポーズの関節位置と 3D ポーズの関節座標 x_j, y_j, z_j からなる 2D-3D 一貫性のみによってマップを学習していることである．すなわち，ロケーションマップ法は，入力画像の光学的特徴や人間の骨格構造などを考慮せず学習をおこなうため，全方位カメラで撮像することによる歪みや分断の影響をうけずにモデルの学習をおこなうことができる．

Mehta らの提案した VNect [9] では，モデルのベースとして近年の画像認識研究で広く利用されている Residual Network (ResNet) [28] を用いている．VNect では，ResNet の res5a 以降の層に骨の長さなどの中間的な特徴を組み込んだアーキテクチャに変更している．本章では，ロケーションマップ法のベースとなるネットワークに Ke らの提案した High-Resolution Network (HRNet) [29] を用いる 3D ポーズ推定モデルを提案する．HRNet はモデルの全体を通して高解像度の特徴量を維持しながら，畳み込みによって低解像度となった特徴量とのマルチスケール混合をおこなうネットワークとなっている．HRNet をベースとしたロケーションマップ法を用いた提案モデルは， VNect よりもシンプルなアーキテクチャとなる．

3.5 推定モデルの学習と評価

HRNet [29] ベースの提案モデルと，Mehta らの提案したロケーションマップ法を用いた VNect [9] を比較評価する．また，提案モデルが歪みと分断を含む正距円筒画像において 3D ポーズを推定できることを確認する．Mehta らは，VNect のベースネットワークとして ResNet50 と ResNet100 を用いた推定モデルについて評価しており，計算複雑性と推定精度の観点から ResNet50 ベースの推定モデルが適当であると結論づけた．本章でも，VNect の計算複雑性とパラメータサイズを参考に提案モデルを実装する．

本節の評価では, ベースネットワークに HRNet-W24 と HRNet-W32 を用いて提案モデルを実装する．W24 と W32 はネットワーク内で畳み込まれる特徴量マップのチャネル数を示している．4 ステージの高解像度ネットワークを実装するため， HRNet-W24 ベースの提案モデルでは並行する 3 つのサブネットワークのチャネルは 48, 96, 192 となる．また， HRNet-W32 ベースの提案モデルでは，サブネットワークのチャネルは 64, 128, 256 となる．HRNet-W24 ベースの提案モデルの概要を図 3.6 に示す．図中の

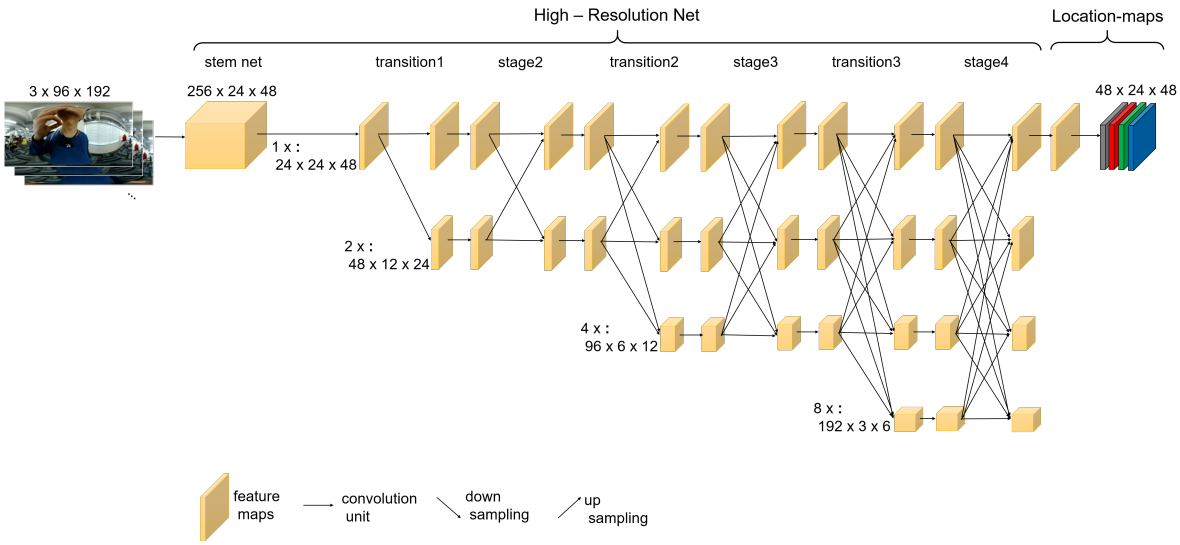

図 3.6　HRNet-W24 ベースの提案モデルのネットワークアーキテクチャ．

stem net では W の値に関係なく入力画像が $256 \times 24 \times 48$ に畳み込まれる．その後は，W に従ってサブネットワークに分岐していく．3D ポーズを構成する 12 の関節について，それぞれ 4 つのマップ $(\boldsymbol{H}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ が必要なため，出力の特徴量マップのチャネルは 48 となる．

提案モデルと VNect について，構築したデータセットを用いて leave-one-person-out 交差検証をおこなう．すなわち，6 人分の学習データと検証データ，そして残りの 1 人分のテストデータを用いて，モデルの学習と評価をおこなう．すべてのモデルについて，入力画像は 96×192 (縦 $\times$ 横) として，ヒートマップの分散は 1.0 とする．学習時のスケジュールはバッチサイズを 64，50 エポックとする．最適化アルゴリズムは Adam [30] を用いて，初期学習率は 0.001 に設定する．

評価指標として，関節毎の誤差平均を示す MPJPE (Mean Per Joint Position Error) と，ある閾値以下の誤差になる関節の割合を示す PCK (Percentage of Correct Keypoints) を用いる．誤差は 3D ポーズを構成する関節の推定結果と真値とのユークリッド距離とする．このとき，データ収集時に，肩幅が 1.0 になるように 3D ポーズを正規化していることに注意する．MPJPE は外れ値に大きく影響を受けるため，PCK の方が外れ値に対して頑健な評価指標といえる．また，MPJPE は平均値周辺の 3D ポーズを推定することで，良い結果を示す可能性があるため，3D ポーズ推定においては PCK の方が適切な評価指標といえる．

MPJPE の評価結果を表 3.2 に示す．表中には，いくつかの関節については個別に表示する．全体の平均は個別に表示されていない関節 (head, neck, torso, waist) も含んだ値である．検証データで撮像されている協力者は学習データに含まれているが，デー

表 3.2　検証データとテストデータの MPJPE の結果.

Model	Backbone	Out-scale	#Params	GFLOPs	Shoulders	Elbows	Wrists	Hands	All
検証データ									
VNect	ResNet50	12 x 24	14.5 M	1.70	0.019	0.053	0.069	0.079	0.044
VNect	ResNet100	12 x 24	33.5 M	2.97	0.019	0.052	0.067	0.076	0.043
Ours	HRNet-W24	24 x 48	16.7 M	1.70	**0.015**	**0.044**	**0.060**	**0.069**	**0.037**
Ours	HRNet-W32	24 x 48	29.3 M	2.70	**0.015**	0.045	0.062	**0.069**	0.038
テストデータ									
VNect	ResNet50	12 x 24	14.5 M	1.70	0.040	0.218	0.387	0.439	0.201
VNect	ResNet100	12 x 24	33.5 M	2.97	0.037	0.215	**0.379**	**0.434**	**0.200**
Ours	HRNet-W24	24 x 48	16.7 M	1.70	0.040	0.215	0.395	0.467	0.204
Ours	HRNet-W32	24 x 48	29.3 M	2.70	**0.036**	**0.202**	0.393	0.463	0.201

表 3.3　検証データとテストデータの PCK の結果.

Model	Backbone	PCK @ 0.1	PCK @ 0.2	PCK @ 0.3
検証データ				
VNect	ResNet50	91.98	98.53	99.56
VNect	ResNet100	92.42	98.64	**99.60**
Ours	HRNet-W24	**94.02**	**98.79**	**99.60**
Ours	HRNet-W32	93.91	98.75	99.59
テストデータ				
VNect	ResNet50	48.11	65.74	77.63
VNect	ResNet100	49.49	67.01	78.11
Ours	HRNet-W24	49.65	67.07	77.89
Ours	HRNet-W32	**50.45**	**68.09**	**79.12**

タ自体は学習データには含まれていない．一方で，テストデータは撮像されている協力者も学習データには含まれていない．すなわち，テストデータは検証データよりも推定が困難なデータである．提案モデルの出力する特徴量マップは VNect の 2 倍の大きさになっているが，これはベースモデルとして利用している HRNet が，高解像度マップを出力するという特徴を持っているためである．また， PCK の閾値を $0.1, 0.2, 0.3$ とした場合の評価結果を表 3.3 に示す．

提案モデル (HRNet-W24) は VNect (ResNet50) よりもパラメタ数の大きいモデルである．しかし，計算複雑性 (GFLOPs) とパラメタサイズ (#Params) を相対的に比較すると，計算複雑性の低いモデルであると言える．表 3.2 の MPJPE の結果から，提案モデルは検証データにおいて VNect よりもわずかに良い結果を示しているが，テストデータにおいては VNect の方が良い結果を示す．表 3.3 の PCK の結果を見ると，提案モデルは検証データとテストデータの双方において VNect よりも良い結果を示す．

すなわち，提案モデルは既知人物の入力画像に対しては真値により近い 3D ポーズを推定するものの，未知人物の入力画像に対してはより大きい外れ値を推定する場合があり，その外れ値によって提案モデルの MPJPE が低下している．以上のことから，HRNet ベースの提案モデルは 3D ポーズ推定の評価で，一般的に使われている PCK 指標において，計算複雑性と推定精度の観点でよりよい性能を示す．

図 3.7 に，テストデータの歪みと分断を含む画像を入力とする提案モデル (HRNet-W24) の推定結果の例を示す．グラフ中の青色の線は 3D ポーズの真値を表し，赤色の線は推定結果を表す．図に示されているとおり，ロケーションマップ法をもちいた推定モデルは歪みや分断を含む画像に対しても 3D ポーズを推定できる．

3.6 結言

本章では，身体に装着した全方位カメラで撮像した画像から，自己 3D ポーズ推定をおこなう方法について述べた．市販のネックマウントを使った全方位カメラの装着法を提案し，撮像される正距円筒図法の画像の特徴について述べた．推定モデルの学習と評価をおこなうデータを収集するために，全方位カメラの画像と，それに同期した 3D ポーズを取得するデータ収録システムを構築した．また，データ収録システムを使って，手話の動作を収録したデータセットを構築した．正距円筒画像を入力として自己 3D ポーズを推定するために，Mehta らの提案するロケーションマップ法 [9, 18] と画像認識モデル HRNet [29] を組み合わせたモデルを提案した．構築したデータセットを用いて推定モデルの学習と評価をおこない，HRNet をベースとした提案モデルがより良い精度で推定できることを示した．

本章を通して，データ収録システムによって正距円筒画像と 3D ポーズのデータセットを大量に収集し，ロケーションマップ法を用いた 3D ポーズ推定モデルを学習することで，歪みと分断を含む画像においても 3D ポーズを精度高く推定できることを示した．しかし，本章では RGB-D カメラを対面に設置した室内という限られた環境下で学習データを収録しており，実際の生活環境下で大量の学習データを集めることが全方位カメラを用いた自己 3D ポーズ推定における大きな課題である．

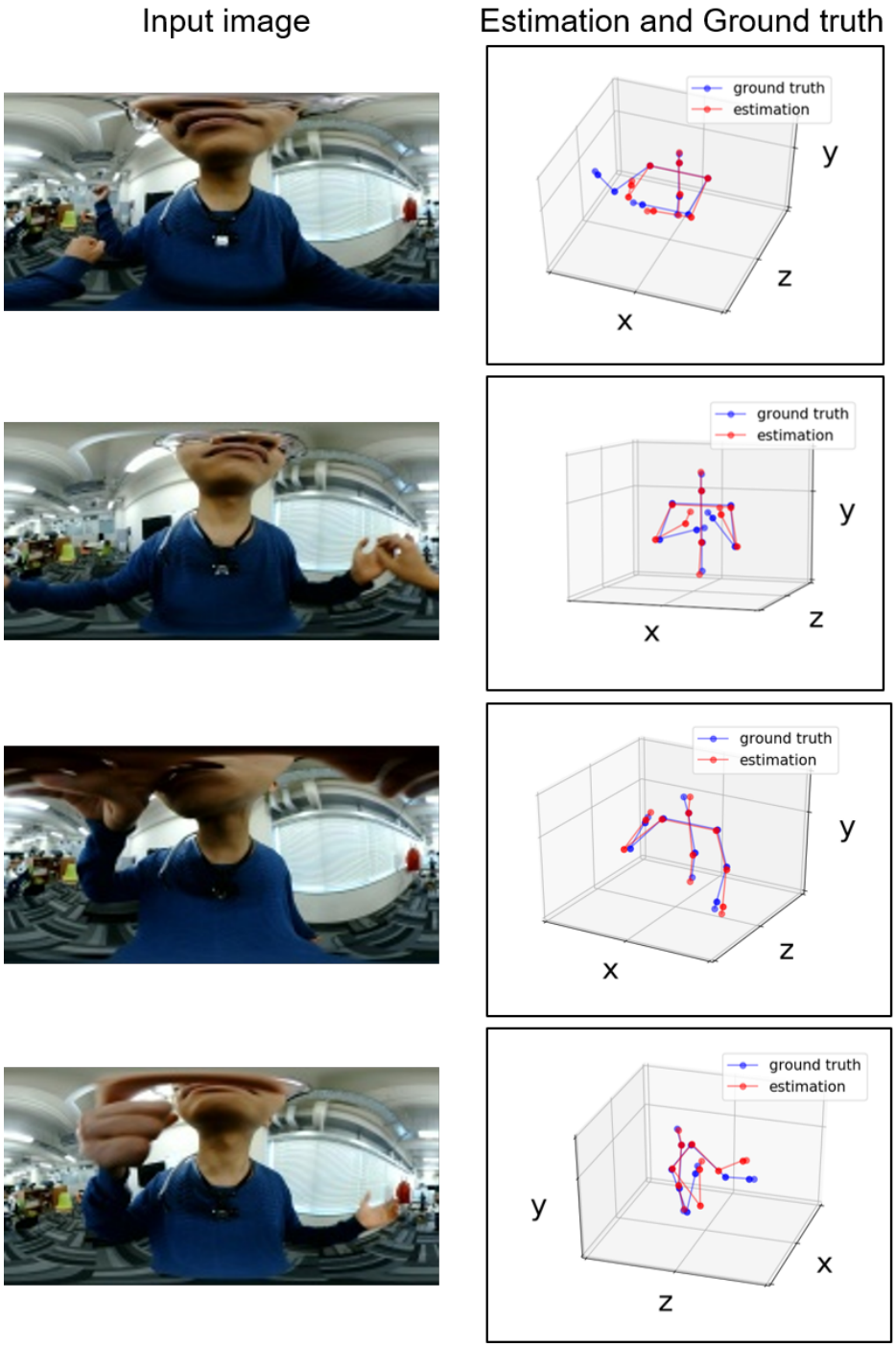

図 3.7　歪みと分断を含む画像を入力とした提案モデル (HRNet-W24) の推定結果の例.

第4章

3D 単位ベクトル化による 2D ポーズ推定と 3D ポーズ拡張の分離

4.1 序言

本章では，3D 単位ベクトル化による 2D ポーズ推定と 3D ポーズ拡張の分離について述べる．4.2 節では，本章で用いる全方位カメラの設定と撮像される画像について述べる．4.3 節では，合成データ生成ツールを用いた学習用データの収集と，評価用の実データの収集について述べる．4.4 節では， 3D 単位ベクトル化を組み込んだ自己 3D ポーズ推定モデルについて述べる．4.5 節では，収集したデータセットを用いた提案モデルの学習と評価について述べる．4.6 節では，提案モデルの特徴に着目したいくつかの追加調査について述べる．4.7 節では，評価と追加調査をもとに提案モデルについて考察する．最後に， 4.8 節で本章についてまとめる．

4.2 装着する全方位カメラと撮像される画像

3 章では，市販のネックマウントを用いて，全方位カメラ (Ricoh R Development Kit) を首の前方あたりに装着した．Ricoh R Development Kit は正距円筒画像で手や顔などの上半身を撮像することができた．一方で，カメラ内部でドームマスタ型から正距円筒図法に画像を変換するため，装着者の動作に対して画像の出力されるタイミングが遅れることや，カメラの重量が重いという特徴があった．

本章では，軽量 (27 g) で小型 (直径 37.6 mm) な全方位カメラ Insta360 Air [1] を，ボディマウントスティックを用いて胴体に装着する．装着している様子と撮像される画像を図 4.1 に示す．本章で用いる全方位カメラは，背面同士で設置される広角 (210°) の魚眼カメラを用いて， 2 つのドームマスタ型画像を撮像する．撮像される全方位カメラ画像は，カメラ内部での画像変換をおこなっていないため，装着者の動作とほぼ同時に出力される．

[1] https://www.insta360.com/product/insta360-air

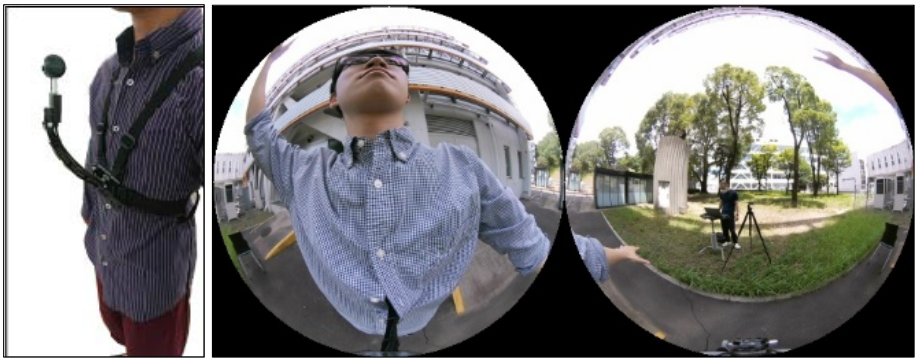

図 4.1　全方位カメラ (Insta360 Air) を装着した様子と撮像される画像.

4.3　学習用の合成データ生成と評価用の実データ収集

3 章では，データを大量に収集することで，歪みや分断を含んだ画像を入力とした 3D ポーズ推定モデルを学習できることを示した．しかし，3D ポーズのアノテーション付き画像を大量に収集することは，据え置きの外部カメラを用いた場合でも難しい作業であり，身体に装着したカメラで撮像した画像についてはより一層困難である．さらには，自然な環境下で 2D / 3D ポーズのアノテーション付きの全方位カメラ画像を収集することは，たとえ高性能なモーションキャプチャ (MoCap) 環境を利用したとしても非常に時間のかかる作業である．

本章では 3DCG ソフトウェア [31] を使って，仮想的な全方位カメラで人間のボディモデルをレンダリングすることで，データ収集作業の困難さを解決する．多様な学習用データを収集するために，合成ヒューマンデータ生成ツール SURREAL [32] を利用する．人間のモデルとして SMPL ボディモデル [33] を，CMU MoCap データセット [34] のモーションデータを用いて動かし，SURREAL で提供されているテクスチャを用いて，さまざまな動作・テクスチャのボディモデルをレンダリングする．

より自然な画像とするために，3DCG ソフトウェア上でカメラと背景を現実の状況に近づけるように設定する．まずは，仮想カメラを実際の全方位カメラの装着位置と同じ位置に設定する．さらに，現実の環境下では装着者の動作の影響をうけてカメラ位置も摂動するため，仮想カメラの位置をレンダリングごとにわずかに移動させる．全方位カメラキャリブレーションツール ocamcalib [25] を用いて，全方位カメラ (Insta360 Air) の固有パラメータを取得し，仮想カメラにそのパラメータを設定することで，同じ光学的特徴をもつ全方位カメラとする．現実の全方位カメラを用いて背景画像 (屋外 50 種類，屋内 54 種類) を撮像し，レンダリングしたボディモデルの背景として合成する．

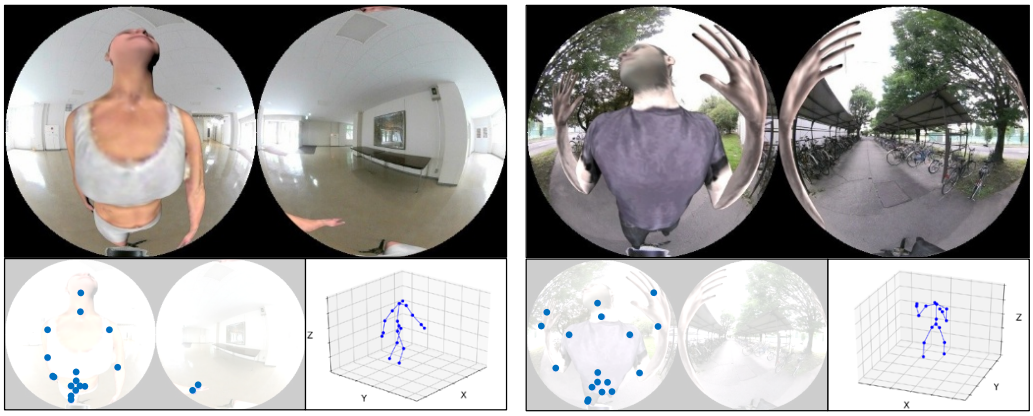

図 4.2　背景が屋内と屋外の 2D / 3D ポーズアノテーション付き合成画像の例.

生成した学習用データでは，ボディモデルを動かすときに適用した 3D モーションデータと，カメラキャリブレーションツール ocamcalib を使って， 2D / 3D ポーズの関節位置と座標を容易に取得できる．本章では，2D / 3D ポーズを構成する 18 関節 (head, neck, spine, pelvis, shoulders, elbows, wrists, hands, hips, knees, ankles) を収集する．3D ポーズの各関節座標は，全方位カメラの位置を原点とした 3D 空間で定義される．本章では，学習用データとして 151,280 の 2D / 3D ポーズアノテーション付きの合成画像を収集した．背景が屋内と屋外の学習用データの例を図 4.2 に示す．図中では，左下は 2D ポーズの関節位置を示し，右下は 3D ポーズの関節座標を示す.

実環境での評価用データも収集する．2 人の協力者にそれぞれ屋内と屋外で，5 種類の動作 (boxing, dancing, hands up, sitting, walking) をしてもらい，全方位カメラ画像と 2D / 3D ポーズを収集する．3.3 節の収録システムを用いて，全方位カメラ画像と同期した 3D ポーズを構成する関節座標を取得する．このとき， 3.3 と同様に 3D ポーズの正規化をおこなう．2D ポーズの関節位置は， 3D ポーズの関節座標を ocamcalib を用いて変換して取得する．各動作について，屋内と屋外で 500 フレームずつを収集し，全体で 5,000 の評価用の実データを収集する．評価用データを収集するときに用いた骨格推定ソフトは，合成データ生成に用いた CMU データセットの MoCap システムとは異なるため，学習用データと評価用データの 3D ポーズの骨格構造はわずかに異なっていることに注意する.

4.4 3D 単位ベクトル化を組み込んだ自己 3D ポーズ推定モデル

身体に装着した全方位カメラで撮像される画像から自己 3D ポーズを推定するパイプラインアプローチを用いた推定モデルを提案する．パイプラインアプローチは，入力画像から 2D ポーズを推定する 2D ポーズ推定部と，平面上の 2D ポーズから 3D ポーズを推定する 3D ポーズ拡張部を組み合わせることで 3D ポーズ推定モデルとする方法である．本章では， 2D ポーズ推定モデルと組み合わせる 3D ポーズ拡張部として，単純なフィードフォワードネットワークを用いる．

提案する 3D ポーズ拡張モデルは， Martinez ら [19] の提案する 2D ポーズの平面上の関節位置から 3D ポーズを推定するフィードフォワードモデルに，3D 単位ベクトル化 (4.4.1 節) と VD 損失関数 (4.4.2 節) を組み合わせることで，身体に装着した全方位カメラを用いた自己 3D ポーズ推定に適したモデルとなっている．

身体に装着した全方位カメラで撮像した画像を入力として，自己 3D ポーズを推定するパイプラインモデルの全体像を図 4.3 に示す．提案する 3D ポーズ拡張部は，推定された 2D ポーズの関節位置から変換された 3D 単位ベクトルを入力としている．全結合層のパラメータ w は，最初と最後の層以外の重みの数を示す．パラメータ b は残差ブロックの繰り返し回数を示す．もし，パラメータ $b = 0$ の場合，3D ポーズ拡張モデルは残差ブロックを持たない入力ブロックと最後の全結合層のみで構成されるネットワークモデルとなる．

提案する 3D ポーズ拡張部の推定モデルは，重みの数を表すパラメータ w と残差ブロックのくり返しを表すパラメータ b によって，モデルの大きさと複雑性を調整できる．提案モデルの最も重要な利点は， 3D ポーズ拡張部が 3D ポーズの関節座標と，原点から関節座標への 3D 単位ベクトルのみを用いて学習可能なことである．

4.4.1 3D 単位ベクトル化

推定された 2D ポーズの平面上の関節位置を，全方位カメラの位置を原点とした 3D 空間の関節座標に向けた 3D 単位ベクトルに変換する．3D 単位ベクトル化は，全方位カメラキャリブレーションツール ocamcalib [25] と，そこから得られる全方位カメラの固有パラメータで実装される．これによって， 2D 平面から 3D 空間へ入力の情報量が拡張される．

ocamcalib は平面上の関節位置 $[u, v]^\top$ を，カメラ位置を原点とした 3D 空間のベク

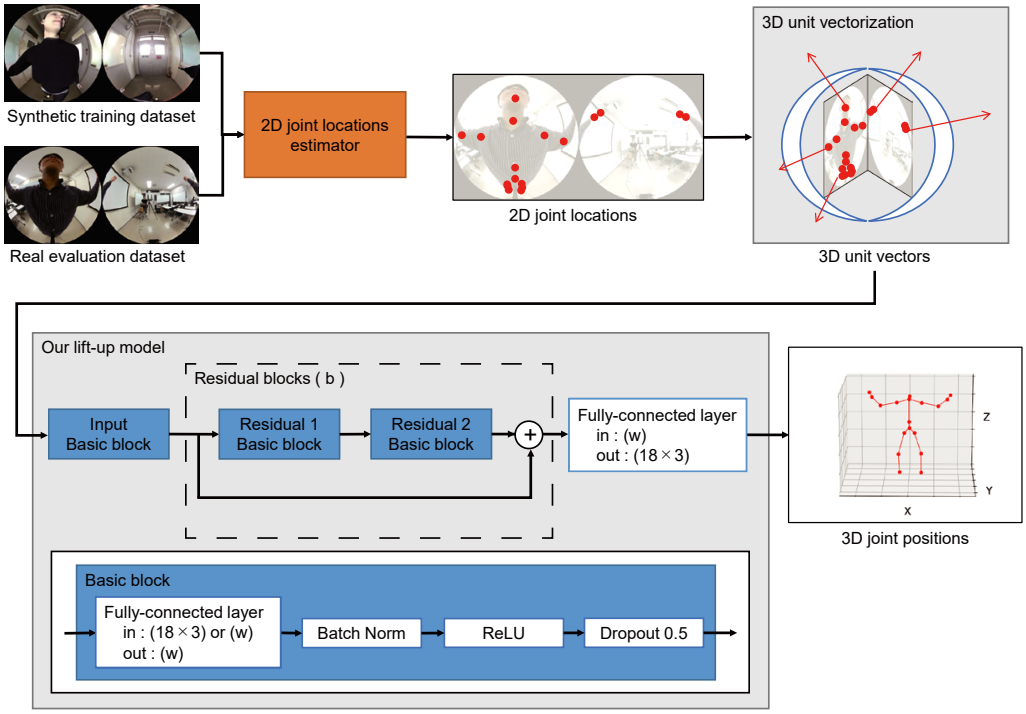

図 4.3　身体に装着した全方位カメラで撮像した画像から，自己 3D ポーズを推定するパイプラインプロセスの全体像.

トル $[x, y, z]^\top$ に変換する．全方位カメラの装着方向と座標系より 3 次元ベクトルは

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} u \\ f(\rho) \\ v \end{bmatrix} \tag{4.1}$$

となる（図 4.4 ）．ここで，$\rho = \sqrt{u^2 + v^2}$ であり，多項式関数 $f(\rho) = \alpha_0 + \alpha_1\rho + \alpha_2\rho^2 + \alpha_3\rho^3 + \alpha_4\rho^4 + \ldots$ はキャリブレーションツール ocamcalib によって得られる．最後に式 4.2 によって単位ベクトル $\boldsymbol{p}$ を得る．

$$\boldsymbol{p} = \frac{1}{\sqrt{x^2 + y^2 + z^2}}[x, y, z]^\top \tag{4.2}$$

4.4.2　VD (Vector and Distance) 損失関数

Martinez らの提案するモデルは L2 誤差を用いて学習をおこなった．全方位カメラの位置を原点とした 3D 空間において， 3D 関節座標は 3D 単位ベクトルとベクトルの大きさ（ 距離 ）に分けることができる．3D ポーズ拡張モデルの入力が 3D 単位ベクトルであること， 3D 関節座標が単位ベクトルと距離に分けられることから，以下

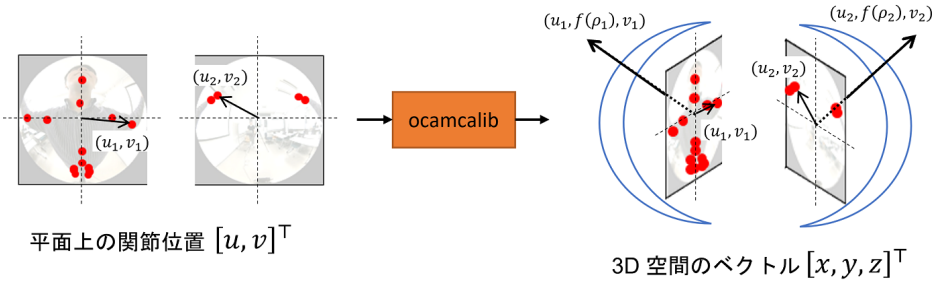

図 4.4　平面上の関節位置 $[u, v]^\top$ から 3D 空間のベクトル $[x, y, z]^\top$ への変換.

の式で表される VD 損失関数を導入する.

$$\mathrm{VDLoss}(\boldsymbol{P}_j) = \lambda_\theta \theta(\boldsymbol{P}_j^{GT}, \boldsymbol{P}_j) + \lambda_d \mathrm{D}(||\boldsymbol{P}_j^{GT}||, ||\boldsymbol{P}_j||)$$

ここで，$\boldsymbol{P}_j$ は任意の関節 $j \in \boldsymbol{J}$ の 3D 座標を表し，GT は真値を表す．VD 損失関数は，以下に示すコサイン類似度と距離の誤差によって構成される.

$$\theta(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \times ||\boldsymbol{y}||}, \mathrm{D}(x, y) = |x - y|$$

本章では，実験的調査によって係数を $\lambda_\theta = 1.0$ ， $\lambda_d = 0.1$ とする．VD 損失関数は，コサイン類似度の誤差を抑制しながら，入力の 3D 単位ベクトルを 3D 関節座標に近づける損失関数と解釈できる.

4.5　提案モデルの学習と評価

生成した学習データと実環境で収集した評価データを用いて，提案する 3D ポーズ拡張モデルの評価をおこなう．本章では，学習のために用意する計算機資源を考慮して，学習用データのうちランダムで選んだ 18,910 のデータを用いる．評価指標として全関節の誤差平均を示す MJPE (Mean Joint Position Error) と，ある閾値以下の誤差になる関節の割合を示す PCK (Percentage of Correct Keypoints) を用いる．誤差は 3D ポーズを構成する各関節の推定結果と真値とのユークリッド距離とする．評価ではデータの生成と収集時に，肩幅が 1.0 になるように正規化した 3D ポーズを，実際に計測した協力者の肩幅に復元した尺度 (mm) で示す.

本章では，先行研究で提案された単眼の魚眼カメラを用いた自己 3D ポーズ推定モデル Mo^{2}Cap^2 [22] ， xR-EgoPose [23, 24] と，3 章で検証した VNect [9] を比較対象として用いる．xR-EgoPose はヒートマップと 3D ポーズを使ったデュアルブランチ・エンコーダデコーダモデルを扱う．提案するパイプラインアプローチの 3D ポーズ拡

張部は， 2D ポーズ推定部を担う推定モデルを必要とする．評価において比較対象として用いる直接回帰アプローチの 3D ポーズ推定モデル VNect と Mo^2Cap^2 は，ネットワークモデルの内部で 2D モジュールと 3D モジュールに分けることができる．2D モジュールでは，ヒートマップ回帰を用いて 2D ポーズの関節位置を推定する．3D モジュールでは，それぞれが提案する方法によって 2D モジュールの推定結果から， 3D ポーズの関節座標を推定する．公平な評価とするため，パイプラインアプローチを提案する本章のモデルと xR-EgoPose は， 2D ポーズ推定部として Mo^2Cap^2 と Vnect の 2D モジュールの結果を使って学習と評価をおこなう．

はじめに，比較対象のうち直接回帰アプローチをとる Mo^2Cap^2 と VNect の学習をおこなう．Mo^2Cap^2 と VNect はベースとなるネットワークとして ResNet50 [28] を用いており， 256×512 の解像度の入力画像から， 32×64 のヒートマップを出力する．事前学習として，通常のカメラで撮像した実画像での 2D ポーズ推定タスクのデータセット MPII Human Pose Dataset [35] を用いて，推定モデルの 2D モジュールを学習する．通常カメラで撮像した実画像を用いた事前学習をおこなうことで，それぞれの推定モデルの低層での特徴量を学習する．つづけて，全方位カメラで撮像した画像について学習するために，生成した学習データを用いて，事前学習済みモデルの 2D / 3D モジュールをファインチューニングする．ファインチューニングはバッチサイズ 32 で 70 エポックおこなう．最適化アルゴリズムは Adam [30] を用いて，初期の学習率は 0.05 に設定する．実画像を用いて学習した低層の特徴量を残すために，ResNet50 ベースのネットワークの 13 層までのブロックの学習率は 0.001 に設定する．

提案する 3D ポーズ拡張モデルは, 生成した学習データで学習済みの比較対象 (VNect , Mo^2Cap^2) の 2D モジュールが出力した 2D ポーズ推定結果を用いる．学習スケジュールはバッチサイズ 32 で 70 エポックおこない，最適化アルゴリズムは Adam を初期学習率 0.001 に設定する．提案モデルのアーキテクチャは， 4.6.1 節の結果から重み 8 ($w = 8$) ，残差ブロックのくり返しなし ($b = 0$) とする．つまり，本節で用いる 3D ポーズ拡張モデルは残差ブロックを持たない， 2 層の全結合層をもつ推定モデルである．xR-EgoPose についても，同様の学習データとスケジュール，最適化アルゴリズムで学習をおこなう．

MJPE (mm) と PCK@30mm の評価結果を表 4.1 に示す．PCK@30mm は括弧内に示す．提案する 3D ポーズ拡張モデルは PCK@30mm では最も高い精度を示すものの，MJPE では xR-EgoPose より低い精度となる．この結果から，提案モデルは 2D ポーズ推定部の性能が良い場合には高い精度を示すものの， xR-EgoPose よりも頑健性が低く， 2D ポーズ推定部の外れ値により大きい影響を受けていると考えられる．また，

表 4.1　評価用データを用いた先行研究と提案モデルの比較結果.

VNect の 2D モジュールとの組み合わせ

model	boxing	dancing	hands up	sitting	walking	total
VNect	195 (11.56)	187 (11.42)	156 (12.16)	144 (12.66)	176 (11.61)	171 (11.88)
xR-EgoPose (p3d+hm)	**164** (16.60)	**161** (14.59)	**155** (16.86)	**123** (14.46)	**147** (15.43)	**150** (15.59)
Ours ($w = 8, b = 0$)	172 (**16.78**)	179 (**17.04**)	168 (**17.81**)	144 (**18.43**)	168 (**16.77**)	166 (**17.37**)

Mo^2Cap^2 の 2D モジュールとの組み合わせ

model	boxing	dancing	hands up	sitting	walking	total
Mo^2Cap^2	313 (8.22)	277 (9.35)	251 (10.38)	220 (3.57)	285 (7.23)	269 (7.75)
xR-EgoPose (p3d+hm)	**151** (**17.00**)	**165** (15.78)	**165** (17.02)	**130** (15.44)	**154** (16.31)	**153** (16.31)
Ours ($w = 8, b = 0$)	186 (16.68)	184 (**16.83**)	178 (**18.01**)	161 (**18.49**)	179 (**16.38**)	178 (**17.28**)

表 4.2　先行研究と提案モデルのパラメータサイズと実行時間.

model	params	estimated time on CPU (ms/frame)
VNect	(14.3 M) + 0.3 M	(454.230) + 11.506
Mo^2Cap^2	(21.3 M) + 69.8 M	(661.807) + 745.415
xR-EgoPose (p3d+hm)	(2D module) + 37.0 M	(2D estimation) + 25.830
Ours (w=8, b=0)	(2D module) + **0.0009 M**	(2D estimation) + **0.086**

提案モデルはパイプラインアプローチの 2D ポーズ推定部として， 2D モジュール部分を利用した Mo^2Cap^2 と VNect のどちらよりも良い精度を示している．

提案モデルのパラメータサイズと，CPU (Intel Xeon @ 2.20GHz) での実行時間を表 4.2 に示す．直接回帰アプローチである VNect と Mo^2Cap^2 については，その推定モデルの内部に入力画像から 3D ポーズを推定するまでの全体が含まれているため，3D モジュールを除いた 2D モジュールのみの値を括弧内に示す．提案モデルは，パラメータサイズと実行時間の観点で他の推定モデルよりも優れている．

VNect の 2D モジュールと組み合わせた 提案モデルの 3D ポーズ推定結果の例を図 4.5 に示す．図中の左列は入力となる全方位カメラ画像，中列は 3D ポーズの真値，右列は 3D ポーズの推定結果を示す．学習に用いた合成データと評価で用いた実データでは， 3D ポーズの関節座標を取得するために用いた MoCap システムが異なるため，3D ポーズの真値と推定結果の骨格構造はわずかに異なっている．

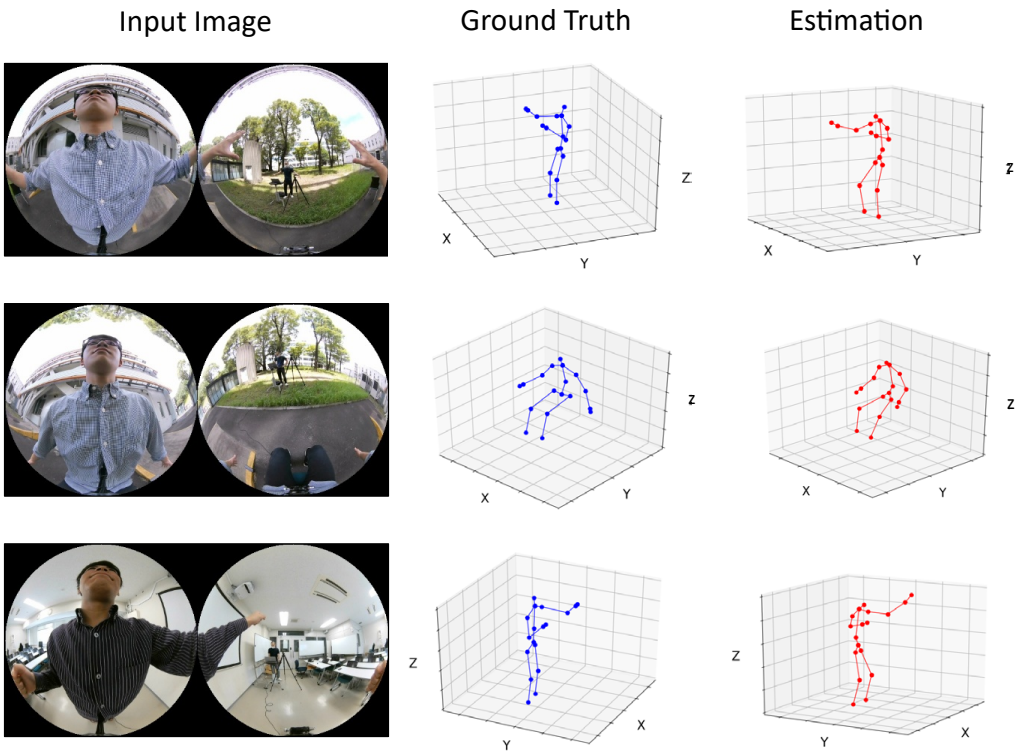

図 4.5　VNect の 2D モジュールと組み合わせた提案モデルの 3D ポーズ推定結果の例.

4.6　提案モデルの追加調査

提案する 3D ポーズ拡張モデルの特徴について，いくつかの追加調査をおこなう．4.6.1 節では，提案モデルの重み (w) と残差ブロックのくり返し (b) パラメータについて評価する．4.6.2 節では，本章で提案した 3D 単位ベクトル化と VD 損失関数の有効性について評価する．4.6.3 節では，パイプラインアプローチの利点である真値を用いた 3D ポーズ拡張モデルの学習ついて評価する．最後に，4.6.4 節では，関連する研究でベンチマークとして利用されている $\mathrm{Mo}^2\mathrm{Cap}^2$ [22] のデータセットを用いた評価をおこなう．

4.6.1　提案モデルのパラメータ

提案する 3D ポーズ拡張モデルについて，重みパラメータを $w = 4, w = 8, w = 16$，残差ブロックのくり返しパラメータを $b = 0, b = 1, b = 2, b = 3$ の組み合わせで実装し，学習と評価をおこなう．MJPE (mm) とその標準偏差の結果を図 4.3 に示す．図中の左のグラフは，VNect の 2D モジュールとの組み合わせを，右のグラフは $\mathrm{Mo}^2\mathrm{Cap}^2$

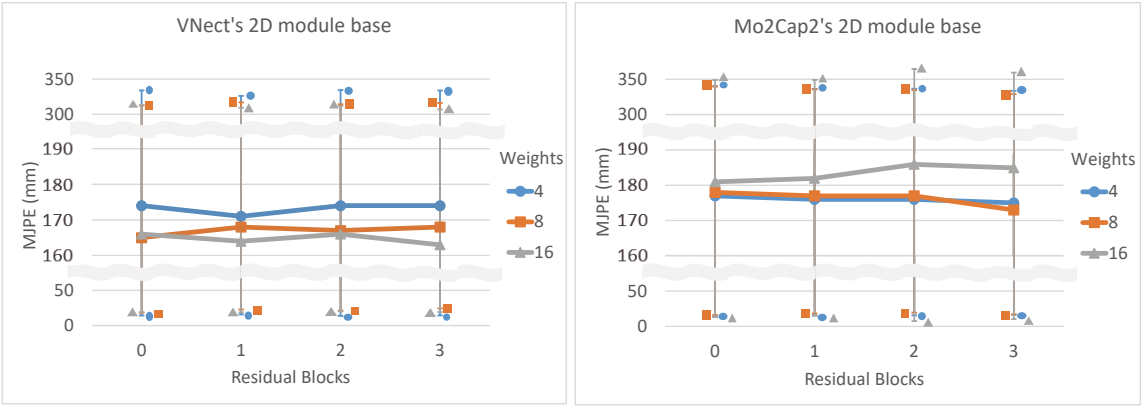

図 4.6　提案モデルの MJPE (mm) と標準偏差.

の 2D モジュールとの組み合わせを示す．標準偏差は視認しやすくするために，マーカーで位置を示した．このとき， VNect の 2D モジュールよりも， Mo^2Cap^2 の 2D モジュールは 2D ポーズ推定の結果が劣っていることに注意する．

推定モデルの重みが大きくなるほど 3D ポーズの推定精度はより高くなる．しかし，重みを大きくすると学習データに対する過学習がおこり， Mo^2Cap^2 の 2D モジュールによる精度の低い 2D ポーズ推定 (外れ値) に対する頑健性が低くなる．残差ブロックのくり返しについて見ると，適切な重みパラメータのなかでは，くり返し回数を増やすことで推定精度の向上が期待できる．4.5 節の提案モデルの評価においては，推定精度と汎化性，パラメータサイズの観点から重みパラメータ $w = 8$ ，くり返しパラメータ $b = 0$ とする．

4.6.2　3D 単位ベクトル化と VD 損失関数

3D 単位ベクトル化と VD 損失関数の有効性を調べるために， Martinez ら [19] の単純なフィードフォワードネットワークと，それに 3D 単位ベクトル化を追加した推定モデル， VD 損失関数で学習した推定モデルの MJPE (mm) と PCK@30mm の結果を表 4.3 に示す．PCK@30mm については括弧内に示す．

Martinez らの 3D ポーズ拡張モデルは，平面上の 2D ポーズの関節位置を入力として， L2 損失関数を用いて学習をおこなう．本章で提案する 3D ポーズ拡張モデルでは， 2D ポーズの関節位置から 3D 単位ベクトルに変換することで入力の情報量を増やしている．この 3D 単位ベクトル化は，VNect と Mo^2Cap^2 の両方の 2D モジュールと組み合わせた推定において， 2D ポーズのみを入力するよりも良い精度を示している．また， 3D 単位ベクトル化を適用した提案モデルにおいて， L2 損失関数と VD

表 4.3　提案モデルと Martinez らの MJPE (mm) と PCK@30mm の結果.

VNect の 2D モジュールとの組み合わせ

model	input	loss	total
Martinez et al.	2D location	L2	194 (16.61)
Ours	3D unit vector	L2	183 (**17.38**)
		VD	**166** (17.37)

Mo^2Cap^2 の 2D モジュールとの組み合わせ

model	input	loss	total
Martinez et al.	2D location	L2	228 (15.89)
Ours	3D unit vector	L2	199 (17.00)
		VD	**178** (**17.28**)

損失関数で学習した結果を比較すると， VD 損失関数で学習した推定モデルの方が良い精度を示している．以上のことから， 3D 単位ベクトル化と VD 損失関数はどちらも，本章で提案する全方位カメラを用いた自己 3D ポーズ推定モデルに有効性があると言える．

4.6.3　真値を用いた 3D ポーズ拡張モデルの学習

4.5 節の結果とともに，学習データの真値の 3D ポーズの関節座標とその 3D 単位ベクトルで学習した 3D ポーズ拡張モデルの MJPE (mm) と PCK@30mm の結果を表 4.4 に示す．PCK@30mm については括弧内に示す．また， 4.3 節で生成したものの，4.5 節での学習に用いなかったデータセット (exclusive synthetic dataset) の真値を用いて学習した結果も示す．なお，真値の 3D ポーズの関節座標とその 3D 単位ベクトルは 3D ポーズのアノテーションから非常に簡単に得られる．

2D ポーズ推定結果から 3D 単位ベクトルに変換したデータを学習に用いることで，2D ポーズ推定に失敗した 2D モジュール出力への頑健性を含んだ学習ができるため，一般的には真値を用いた学習よりも 2D ポーズ推定結果を用いるほうが良い推定モデルとなる．しかし，提案する 3D ポーズ拡張モデルでは， 3D ポーズアノテーションの真値のみを使った学習でも， 2D モジュールの推定結果を使った学習と同程度の精度を示している．また， 2D モジュールの学習に使われていない， データセット (ex. synthetic ground truth) の真値を用いた学習でも同等の精度を示している．

表 4.4　2D ポーズ推定結果を用いた学習と真値学習の評価結果.

VNect の 2D モジュールとの組み合わせ

training dataset (number of data)	total
2D estimation result (18,910)	**166** (17.37)
ground truth (18,910)	167 (17.27)
ex. synthetic ground truth (132,370)	168 (**17.42**)

Mo^2Cap^2 の 2D モジュールとの組み合わせ

training dataset (number of data)	total
2D estimation result (18,910)	178 (**17.28**)
ground truth (18,910)	177 (16.93)
ex. synthetic ground truth (132,370)	**176** (17.22)

4.6.4　ベンチマークデータセットを用いた評価

魚眼カメラを用いた自己 3D ポーズ推定に関する研究でベンチマークとして扱われている Mo^2Cap^2 データセット [22] を用いて，提案する 3D ポーズ拡張モデルと先行研究 [9, 18, 22, 23, 24] を評価する．Mo^2Cap^2 データセットでは，データ収集時に使用した魚眼カメラの固有パラメータが公開されていないため， 2D モジュールの推定結果から変換した 3D 単位ベクトルが利用できない．そのため，本節では 3D ポーズ拡張モデルをデータセットで公開されている 3D ポーズの関節座標とその 3D 単位ベクトルを用いた真値学習をおこなう．

本節では，残差ブロックのくり返しを 2 回 ($b = 2$) おこなう，重み 4096 ($w = 4096$) のネットワークアーキテクチャとなる 3D ポーズ拡張モデルを用いる．推定モデルのパラメータサイズは 67.5 M で， CPU 上での実行時間は 16.592 (ms / frame) である．学習は， $\lambda_\theta = 1.0, \lambda_d = 0.00001$ の係数とした VD 損失関数を用いて，バッチサイズ 1024 で 5000 エポックおこなった．

MJPE (mm) の結果を表 4.5 に示す．提案モデルの推定精度は， xR-EgoPose と Mo^2Cap^2 よりも劣るが，真値学習にも関わらず一部の動作では Mo^2Cap^2 よりも良い精度を示している．推定結果の 3D ポーズの例を図 4.7 に示す．図中では左列に入力画像，中列に 3D ポーズの真値，右列に 3D ポーズの推定結果を示す．

4.7　提案モデルについての考察

提案する 3D ポーズ拡張モデルは，組み合わせる 2D ポーズ推定モデルの性能に応じて，重みと残差ブロックのくり返しパラメータを調整できる．具体的には，推定精

表 4.5　Mo^2Cap^2 データセットを用いた先行研究と提案モデルの MJPE (mm) の結果.

Indoor

model	walking	sitting	crawling	crouching	boxing	dancing	stretching	waving	total
3DV' 17	48.76	101.22	118.96	94.93	57.34	60.96	111.36	64.50	76.28
VNect	65.28	129.59	133.08	120.39	78.43	82.46	153.17	83.91	97.85
Mo^2Cap^2	38.41	70.94	94.32	81.90	48.55	55.19	99.34	60.92	61.40
xR-EgoPose (p3d+hm)	**38.39**	**61.59**	**69.53**	**51.14**	**37.67**	**42.10**	**58.32**	**44.77**	**48.16**
Ours	40.28	84.71	117.79	98.15	49.78	54.19	99.81	59.69	68.11

Outdoor

model	walking	sitting	crawling	crouching	boxing	dancing	stretching	waving	total
3DV' 17	68.67	114.87	113.23	118.55	95.29	72.99	114.48	72.41	94.46
VNect	84.43	167.87	138.39	154.54	108.36	85.01	160.57	96.22	113.75
Mo^2Cap^2	63.10	**85.48**	96.63	92.88	96.01	68.35	123.56	61.42	80.64
xR-EgoPose (p3d+hm)	**43.60**	85.91	**83.06**	**69.23**	**69.32**	**45.40**	**76.68**	**51.38**	**60.19**
Ours	67.36	99.87	109.47	110.10	98.92	70.39	123.51	62.74	85.95

度の高い 2D ポーズ推定モデルと組み合わせる場合ほど，重みと残差ブロックのくり返しパラメータを大きくすることで， 3D ポーズ推定モデルの精度を上げることができる．

パラメータサイズが入力画像の解像度に比例して大きくならないことも提案モデルの利点である．画像から 2D / 3D ポーズを推定するモデルの精度を向上させる場合，もっとも単純な方法は入力画像の解像度を大きくして情報量を増やすことである．しかし，この単純な方法は Vnect ，Mo^2Cap^2 ，xR-EgoPose などのヒートマップを用いた推定モデルのパラメータサイズを大きくする．提案する 3D ポーズ拡張モデルでは，パラメータサイズは 3D ポーズを構成する関節数と 2D ポーズ推定部の性能で決まるため，入力画像の解像度増加にともなったパラメータサイズの増大を抑制できる．

提案する 3D ポーズ拡張モデルの最も優れた利点は 3D ポーズの関節座標とその 3D 単位ベクトルを用いた真値学習ができることである．この特徴によって，パイプラインアプローチにおける 2D ポーズ推定部と 3D ポーズ拡張部を別々に学習できるだけでなく，3D ポーズ拡張部の学習に必要なデータセットを独立して収集することが可能になる．

同様にパイプラインアプローチを用いている先行研究の xR-EgoPose も，3D ポーズの関節座標とそれを 2D 平面上に変換した関節位置のヒートマップを用いた真値学習が可能である．しかし，魚眼カメラ (もしくは全方位カメラ) の光学特性の歪みの度合いによって，同じ 3D ポーズでも平面上に変換される 2D ポーズは大きく変わる．この 2D-3D の一貫性によって， xR-EgoPose はカメラの固有パラメータごとにデータの収集と再学習を必要とする．

本章で提案した 3D ポーズ拡張モデルでは，パイプラインアプローチの中に 3D 単位ベクトル化を組み込むことで，データ収集と全方位カメラの固有パラメータの観点で，3D ポーズ推定における 2D-3D 一貫性を切り離すことができる．具体的には，全方位カメラの固有パラメータを適用した 3D 単位ベクトル化モジュールを用いることで，カメラ毎の光学系の違いを閉じ込めることができる．それによって，一般に公開されている利用可能な 3D MoCap データセットから容易に生成できる 3D ポーズの関節座標とその 3D 単位ベクトルを用いて， 3D ポーズ拡張モデルの独立したデータ収集と学習が可能である．すなわち，身体に装着した全方位カメラ (もしくは魚眼カメラ) を用いた自己 3D ポーズ推定モデルを構築する際に，非常に大きな課題となる 3D ポーズのアノテーション付き画像を収集する負担を軽減することができる．また，全方位カメラの光学系の違いを 3D 単位ベクトル化モジュールに閉じ込めることで，異なる光学系のカメラを用いた場合でも 3D ポーズ拡張部の再学習は不要となる．

4.8 結言

本章では， 3D 単位ベクトル化を用いた 2D ポーズ推定部と 3D ポーズ拡張部の分離について述べた．軽量で小型の全方位カメラの装着方法と，撮像される画像の特徴について述べた．大量の学習用データを用意するための人工的な合成データの生成と，評価用の実データの収集について述べた．全方位カメラ画像を入力とした自己 3D ポーズ推定をおこなうために， 3D 単位ベクトル化を組み込んだ 3D ポーズ拡張モデルと，学習時に用いる VD 損失関数を用いた 3D ポーズ推定モデルを提案した．提案する 3D ポーズ拡張モデルは CPU 上で実時間で動作する一方で，その推定精度も先行研究と比較して同程度であることを示した．提案する 3D ポーズ拡張モデルについていくつかの追加調査をおこない以下について示した．

- 重みと残差ブロックのくり返しパラメータによる推定精度と頑健性の変化.
- 3D 単位ベクトル化と VD 損失関数の有効性.
- 真値学習を用いた 3D ポーズ拡張モデルの評価.
- ベンチマークデータセットを用いた評価.

本章の研究を通して，提案する 3D 単位ベクトル化を組み込んだ 3D ポーズ拡張モデルは，身体に装着した全方位カメラ (もしくは魚眼カメラ) を用いた自己 3D ポーズ推定モデルを構築する際に，非常に大きな課題となる 3D ポーズのアノテーション付き画像を収集する負担を軽減できることを示した．

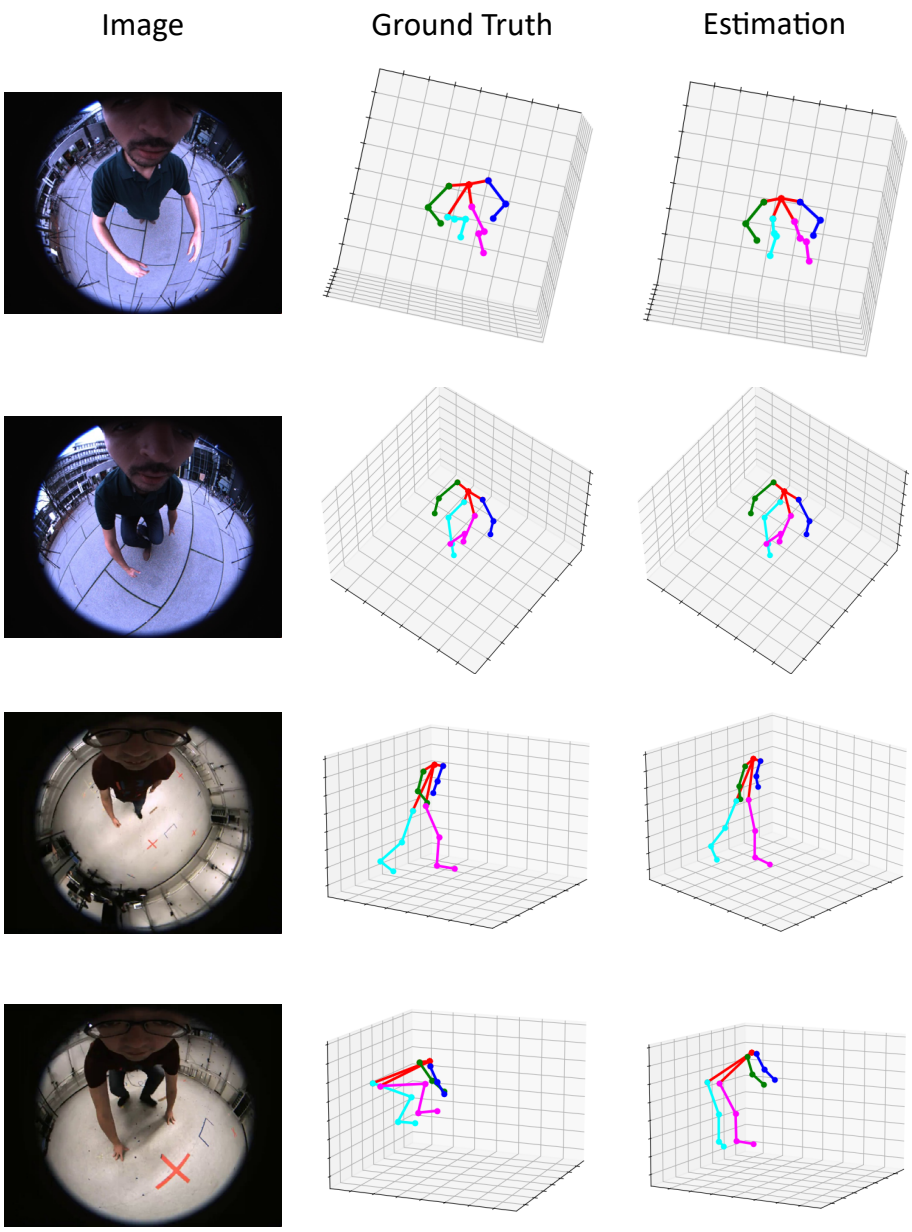

図 4.7　ベンチマークのデータセットを用いた提案モデルの推定結果の例.

第5章

装着型 MoCap システムのプロトタイプ開発

5.1 序言

本章では，話者中心の手話認識と翻訳への応用に向けた，全方位カメラを用いた装着型 MoCap システムのプロトタイプ開発について述べる．5.2 節では，提案システムのハードウェアの構成について述べる．5.3 節では，提案システムで身体動作を追跡する自己 3D ポーズ推定について述べる．5.4 節では，自己 3D ポーズ推定モデルの学習と評価をおこなう合成データの生成について述べる．5.5 節では，提案システムの推定精度，頑健性，実行速度について評価する．5.6 節では，実データを用いた提案システムの定性評価をおこなう．5.7 節では，提案システムの評価についての考察を述べる．最後に， 5.8 節で本章についてまとめる．

5.2 ハードウェア構成

手話は手指動作と非手指動作を用いながら，周辺の空間を使って表現する言語である．手話の自動認識に関する研究では，手話動作を撮像した画像をそのまま入力とする End-to-End モデル [3, 4, 5] のほか，画像から上半身の 3D ポーズ・手形・口形の特徴量を入力として手話認識をおこなうマルチチャネルモデル [6] も提案されている．すなわち，手話認識への応用に向けた装着型 MoCap システムでは，手話動作のもとで顔と手を含む上半身全体を撮像できる位置に全方位カメラを装着することが望ましい．

本章では，グースネックケーブルとヘッドバンドを用いて，あごの前方に小型の全方位カメラ (Insta360 Air) を装着する．また，持ち運び可能であり，かつオフライン環境でも利用できるように，ポータブルバッテリーで動作可能な小型の GPU 搭載シングルボードコンピュータ (Nvidia Jetson Nano [1]) で自己 3D ポーズ推定をおこなう．ハードウェアの構成と装着例を図 5.1 に示す．なお，プロトタイプで用いるカメラや

[1] https://developer.nvidia.com/embedded/jetson-nano-developer-kit

図 5.1　装着型 MoCap システムのハードウェア構成と装着例.

シングルボードコンピュータは比較的安価で入手できる汎用品であり，ハードウェアを構成する各デバイスはこの構成に限るものではない.

5.3　自己 3D ポーズ推定モデル

装着した全方位カメラで撮像される画像から，自己 3D ポーズを推定するために，2D ポーズ推定部と 3D ポーズ拡張部を組み合わせるパイプラインアプローチを用いる．3D ポーズ推定モデルを，画像から 2D ポーズを推定するモデルと， 2D ポーズから 3D ポーズへ拡張するモデルに分けることで，それぞれに適した推定モデルを選ぶことができる．また， 4 章で提案した 3D 単位ベクトル化を用いることで，データセットの収集やそれを使った学習を分離して実施できる.

2D ポーズ推定部として, 軽量な畳み込みニューラルネットワークである MobileNetV2 [36] を用いる．3D ポーズ拡張部には， 4 章で提案した 3D 単位ベクトル化を組み込んだ単純なフィードフォワードネットワーク [19] を用いる．フィードフォワードネットワークは，重み 256 ($w = 256$)，残差ブロックのくり返し 2 回 ($b = 2$) で構成する．3D ポーズ推定モデルの全体構成を図 5.2 に示す.

5.4　学習と評価用の合成データ生成

自然な環境下で，身体に装着したカメラの画像と 3D ポーズのアノテーション付きデータを大量に収集することは，高精度の MoCap システムを利用したとしても時間のかかる作業である．特に，全方位カメラは撮像する画像に歪みを含むため， 2D 関節位置と 3D 関節座標の一貫性を保つことが難しく，たとえ人手でおこなったとして

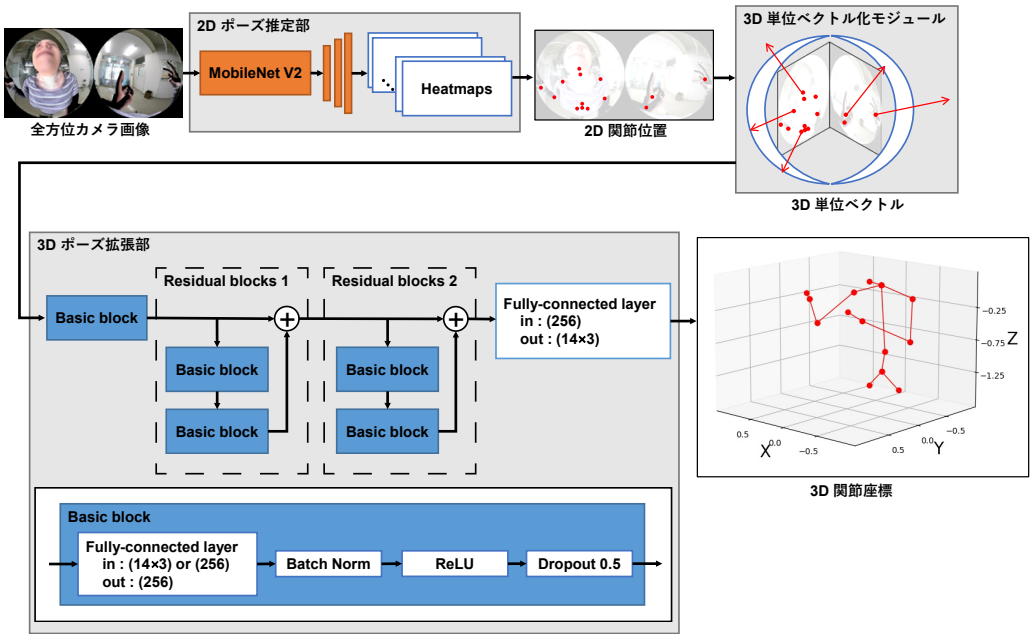

図 5.2　全方位カメラ画像を入力とした 3D ポーズ推定モデルの全体構成.

も正確なアノテーションを付けることは非常に難しい．そのため， 2D / 3D ポーズのアノテーション付き全方位カメラ画像を人工的に生成して，学習用と評価用のデータセットを構築する．基本となる手順は 4.3 節の方法を用いるが，異なる点を以下に述べる．

SMPL ボディモデル [33] を動かす 3D モーションデータとして， Nagashima ら [37, 38] の大規模手話データセット [2] を用いる．仮想的な全方位カメラの初期位置を，本章のハードウェアに合わせた位置に設定したうえで，レンダリング毎に正規分布に従った位置と回転の変動を仮想カメラに与える．このとき，回転はオイラー角で X 軸，Y 軸， Z 軸の順でおこなう．装着者の 2D / 3D ポーズとして， 14 関節 (head, neck, spine, pelvis, shoulders, elbows, wrists, hands, hips) を扱う．生成する合成データの例を図 5.3 に示す．

合成データを生成するにあたり，学習と評価でそれぞれ異なる 3D モーションデータと背景を用いる．また，全方位カメラの摂動に対する自己 3D ポーズ推定モデルの頑健性を評価するために，カメラの位置と回転の大きさを変えた評価データを生成する．生成した合成データセットを表 5.1 に示す．

[2] https://www.nii.ac.jp/dsc/idr/rdata/KoSign/

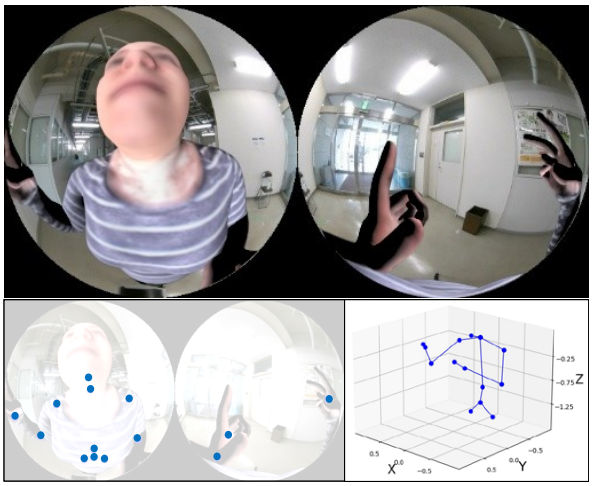

図 5.3　生成した合成画像 (上) と 2D 関節位置 (左下) と 3D 関節座標 (右下).

5.5　提案システムの学習と評価

自己 3D ポーズ推定モデルを構成する 2D ポーズ推定部と 3D ポーズ拡張部を，それぞれ個別に学習をおこなう．2D ポーズ推定部の推定モデルは， 2D ポーズのアノテーション付き合成画像を学習データとして用いる．3D ポーズ拡張部の推定モデルは，生成したデータのうち 3D ポーズのアノテーションのみを学習データとして用いる．

2D ポーズ推定部の学習では 4.5 節と同様に，はじめに 2D ポーズのアノテーション付き実画像データセット MPII Human Pose [35] を用いた事前学習をおこなう．生成した学習データを用いたファインチューニングでは， MobileNetV2 を構成する 16 ブロックのうち 1 から 15 ブロックの重みを固定して学習をおこなう．初期学習率を 0.001 とした Adam [30] を最適化アルゴリズムとして用いて，バッチサイズ 32 で 140 エポックの学習をおこなう．

3D ポーズ拡張部の学習では，生成した 3D ポーズの関節座標とその 3D 単位ベクトルを用いて学習をおこなう．最適化アルゴリズム，バッチサイズ，エポック数は 2D ポーズ推定部と同じである．4 章で提案された VD 損失関数は，最適な係数 $\lambda_\theta, \lambda_d$ を得るために調査が必要なため，本章では L2 損失関数を用いて学習をおこなう．

評価用に生成した合成データを用いて， 3D ポーズ推定の精度と実行時間，及びカメラ位置と回転の変動に対する頑健性を評価する．評価尺度は 3D ポーズを構成する関節座標の真値と推定値の誤差平均 (MPJPE) を用いる．

カメラの位置と回転をともに変動させた評価データに対する推定結果を表 5.2 に示す．学習データと同じカメラ変動 (位置： 8.75 mm ，回転： 5°) に対して，全関節の平均で 57 mm の誤差がある．カメラ変動の大きさにともなって推定の誤差が大きく

表 5.1　生成した学習用と評価用の合成データセット.

用途	モーション	背景画像	カメラ変動		データ数
			位置 (σ^2)	回転 (σ^2)	
学習	128 単語	室内：40 室外：40	8.75 mm	5°	22,872
評価	64 単語	室内：14 室外：10	0.00 mm	0°	2,528
			8.75 mm	5°	
			17.50 mm	10°	
			26.25 mm	15°	
			35.00 mm	20°	
			8.75 mm	0°	
			17.50 mm	0°	
			26.25 mm	0°	
			35.00 mm	0°	
			0.00 mm	5°	
			0.00 mm	10°	
			0.00 mm	15°	
			0.00 mm	20°	

表 5.2　カメラの位置と回転をともに変動させた評価データの推定結果.

カメラ変動		誤差平均 (mm)									
位置 (σ^2)	回転 (σ^2)	head	neck	spine	pelvis	shoulders	elbows	wrists	hands	hips	total
0.00 mm	0°	10	15	34	42	20	44	83	102	50	50
8.75 mm	5°	14	18	39	48	24	51	94	115	58	57
17.50 mm	10°	24	27	52	64	36	66	120	146	76	75
26.25 mm	15°	38	43	83	102	56	98	171	205	120	112
35.00 mm	20°	57	67	133	162	88	147	239	284	189	165

なっている．関節毎では，wrists や hands などの動きが大きく，オクルージョンの発生しやすい関節の推定について誤差が大きくなっている．また，体の中心線にある関節でもカメラ位置に近い spine よりも，pelvis や hips といったカメラから遠い関節の推定誤差が大きくなっている．これは，カメラ位置を原点とした 3D 空間のデータを用いて学習，評価をしていることが原因と考えられる．

自己 3D ポーズ推定モデルのパラメータサイズと，画像 1 フレームあたりの実行時間を表 5.3 に示す．持ち運び可能な GPU 搭載シングルボードコンピュータ Nvidia Jetson Nano で 230.031 ms (4.3 fps) で動作する．処理能力の高い GPU ありの Google Colaboratory では 42.413 ms (23.6 fps) で動作する．3D ポーズ拡張部はパラメータサイズも非常に小さく，1.0 ms 未満で動作している．モデルのパラメータサイズと実行

表 5.3　自己 3D ポーズ推定モデルのパラメータサイズと，画像 1 フレームあたりの実行時間.

パラメータサイズ		
2D ポーズ推定部	3D ポーズ拡張部	合計
9.605 M	0.290 M	9.895 M

デバイス	2D ポーズ推定部	3D ポーズ拡張部	合計
Nvidia Jetson Nano (Nvidia Tegra X1)	229.352 ms	0.679 ms	230.031 ms
Google Colab GPU (Nvidia Tesla T4)	42.259 ms	0.154 ms	42.413 ms

時間のほとんどは 2D ポーズ推定部に使われている.

カメラ位置と回転の両方，もしくはそのいずれかを変動させた評価データに対する全関節の誤差平均のグラフを図 5.4 に示す．位置と回転のどちらも変動が大きくなるにともなって，推定誤差も大きくなるが，回転の方が変動量に対する推定誤差への影響がより大きくなっている．また，位置と回転の変動を組み合わせた場合の推定誤差はそれぞれの誤差を足し合わせた程度の誤差になる.

5.6　実データを用いた定性評価

日常生活の環境において全方位カメラ画像と 2D / 3D ポーズのデータを収集することは非常に困難なので，全方位カメラ画像のみを収集して定性評価をおこなう．3D ポーズ推定の成功例を図 5.5 に，失敗例を図 5.6 に示す．また，入力となる全方位カメラ画像上に 2D ポーズ推定の結果を赤丸で示す．図中において，左列に外部カメラからの画像，中列に入力となる全方位カメラ画像と 2D ポーズ推定結果，右列に 3D ポーズ推定の結果を示す.

成功例では 2D ポーズ推定の精度が高く，失敗例ではいくつかの関節で推定が失敗している．すなわち， 2D ポーズ推定の精度が，その後の 3D ポーズ拡張の結果に大きく影響している．また， hands, wrists, elbows などの自由度の高い関節が 2D ポーズの推定に失敗しやすい傾向がみられる．同じような周辺の環境，人物の有無であっても，推定に成功する場合と失敗する場合があり，失敗する要因については明らかではないが，hands や wrists が入力画像内で確認しづらい位置にあるときに，全方位カメラ画像の右側のレンズに撮像される物体をその関節と誤認識する傾向がみられる.

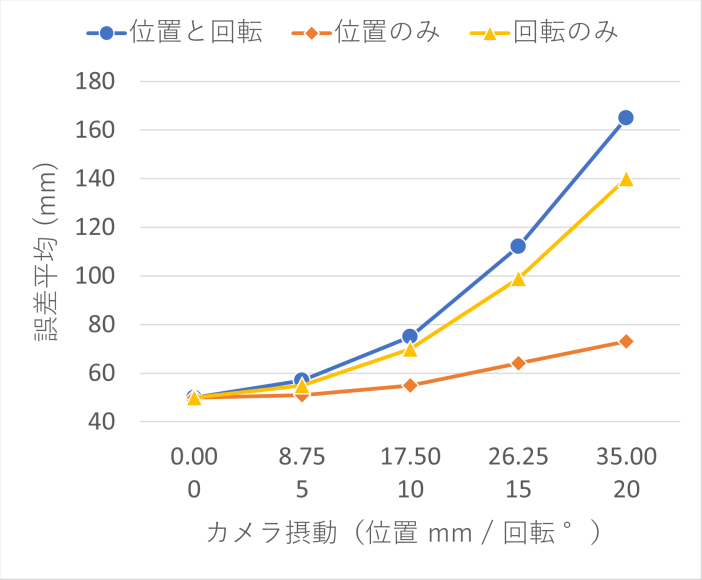

図 5.4　カメラの位置と回転の変動による推定誤差の推移.

5.7　プロトタイプシステムについての考察

提案する装着型 MoCap システムのプロトタイプに対する評価と今後の課題について述べる．合成データを用いた評価によって，全方位カメラの位置と回転の変動が推定精度に影響を与えることを示した．位置と回転のどちらの変動も大きくなるに従って，推定誤差への影響がより大きくなるが，回転の方が変動に対する誤差の影響が大きくなる．より精度の高い推定をおこなうためには，カメラ位置と回転を小さくする対策や，変動による影響を小さくする対策が必要である．たとえば，位置と回転の変動を小さくするために装着デバイスのデザインや材質を検討することや，回転の変動の影響を小さくするために，デバイスにチルトセンサを組み込んで取得した傾きを使って入力画像を前処理する方法などが考えられる．また，オクルージョンを発生しづらくするために，カメラ数の増加を含めた装着位置の検討なども必要である．

プロトタイプの自己 3D ポーズ推定モデルでは, 推定の実行時間は GPU 搭載シングルボードコンピュータで 230.031 ms / frame (4.3 fps) , GPU ありの Google Colaboratory で 42.413 ms / frame (23.6 fps) となることを示した．Cherniavsky ら [39] によると，人間が手話動画から意味を理解する場合に，15 fps を下回ると徐々に理解が難しくなり， 5 fps を下回るとより一層難しくなると言われている．本論文では，画像認識を用いた手話の自動認識・翻訳への応用を目的としており，直接の参考にはできないものの，プロトタイプの実行時間 (4.3 fps) は手話の自動認識には不十分と考えられる．3D ポーズ拡張部は高速に動作しているため，リアルタイムで動作させるためには 2D

ポーズ推定部について，より軽量で高速な 2D ポーズ推定モデルを用いる必要がある．

実画像を用いた定性評価によって， 5.5 節で用いた実画像データセットを用いた事前学習と，合成データを用いたファインチューニングによって 3D ポーズ推定ができることを示した．一方で， 2D ポーズ推定の精度低下の影響によって，自己 3D ポーズ推定が失敗する場合もあった．より高精度な推定モデルの学習をおこなうためには，合成データだけではなく実画像データの収集を含めたデータの拡充が求められる．

5.8 結言

本章では，全方位カメラを用いた装着型 MoCap システムのプロトタイプ開発について述べた．手話表現を考慮したカメラの装着位置と， GPU 搭載のシングルボードコンピュータを用いたハードウェア構成について述べた．装着した全方位カメラで撮像される画像から自己 3D ポーズを推定するモデルと，そのモデルを学習・評価するための合成データの生成について述べた．生成した合成データを用いて推定モデルを学習し，関節ごとの推定精度の他に，カメラの摂動（位置と回転）に対する頑健性と実行時間について評価した．また，日常生活の環境下で撮像した全方位カメラ画像を用いて， 3D ポーズ推定の定性評価もおこなった．

本章を通して，全方位カメラを用いた装着型 MoCap システムの開発におけるプロセスを示した．また，プロトタイプシステムを評価することで，提案システムにおける今後の課題と対応策について考察した．

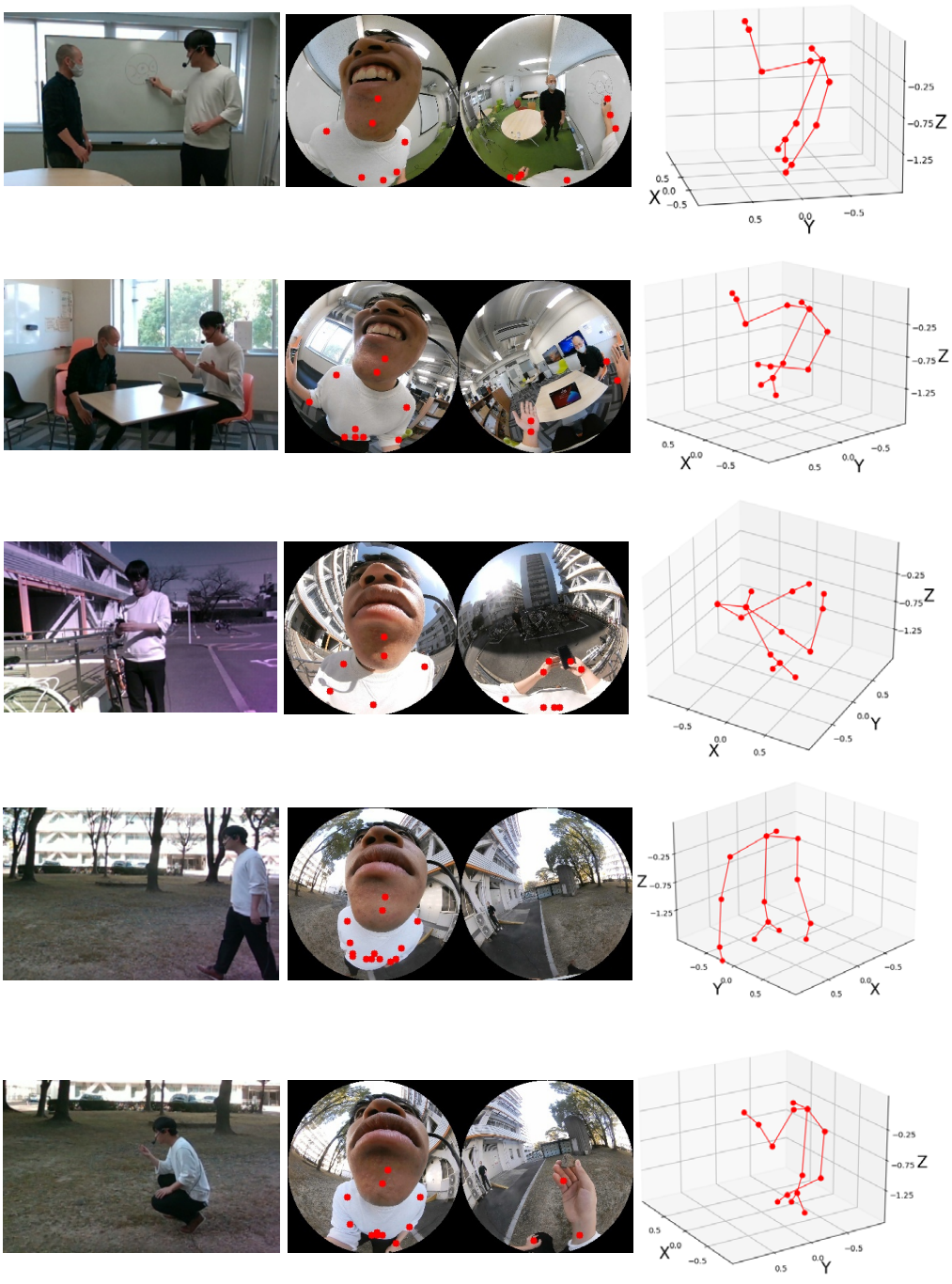

図 5.5　3D ポーズ推定の成功例.

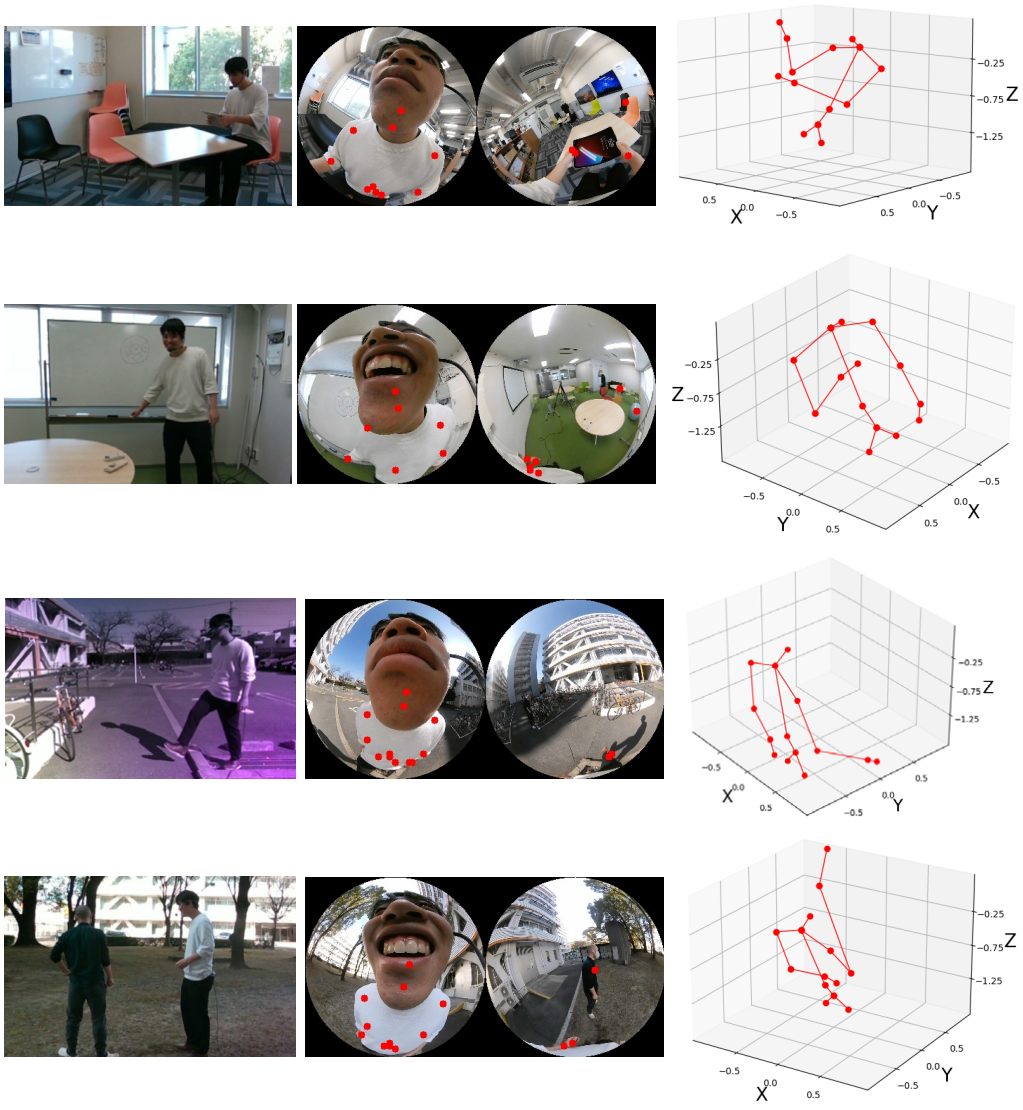

図 5.6　3D ポーズ推定の失敗例.

第6章

カメラ摂動に対する頑健性の調査

6.1 序言

本章では，装着した全方位カメラの摂動に対する自己 3D ポーズ推定モデルの頑健性について述べる．6.2 節では，学習と評価に用いるカメラ摂動を含む合成データの生成について述べる．6.3 節では，調査対象とする自己 3D ポーズ推定モデルについて述べる．6.4 節では，自己 3D ポーズ推定モデルの学習と，カメラ摂動に対する頑健性の評価について述べる．最後に，6.5 節で本章についてまとめる．

6.2 学習と評価用の合成データ生成

5.5 節で示したように，身体に装着した全方位カメラの摂動は，自己 3D ポーズ推定の精度に大きく影響を与える．本節では， 5 章のカメラ位置を基準として，カメラの位置と回転を変動させた学習と評価用の合成データを生成する．構築したデータセットを用いて，いくつかの自己 3D ポーズ推定モデルを学習，評価することでカメラ摂動の頑健性について示す．

合成データを生成する基本的な手順は 5.4 節の方法を用いるが，SMPL ボディモデル [33] を動かす 3D モーションデータとして CMU MoCap データセット [34] を用いる．このとき，肩幅が 350 mm になるように収集する 3D ポーズの骨格を正規化する．また，合成データを生成するにあたり，学習と評価でそれぞれ異なる 3D モーションデータと背景を用いる．

学習用データではレンダリングごとに, 3 次元空間において正規分布 $N(\sigma^2 = 17.50\,\mathrm{mm})$ に従った位置の移動と，$N(\sigma^2 = 10^\circ)$ に従った回転をすることでカメラを摂動させる．評価用データでは，位置と回転の変動を 5 段階（位置： $\sigma^2 = 0.00$ mm から 35.00 mm まで，回転： $\sigma^2 = 0^\circ$ から 20° まで）に分けて生成する．生成する合成データの例を図 6.1 に示す．この例では，合成データは同一の CMU MoCap データから生成されているが，異なるカメラ摂動が適用されている．上図は位置 $\sigma^2 = 0.00$ mm ，回転 $\sigma^2 = 0^\circ$ であり，下図は位置 $\sigma^2 = 35.00$ mm ，回転 $\sigma^2 = 20^\circ$ である．構築したデータセット

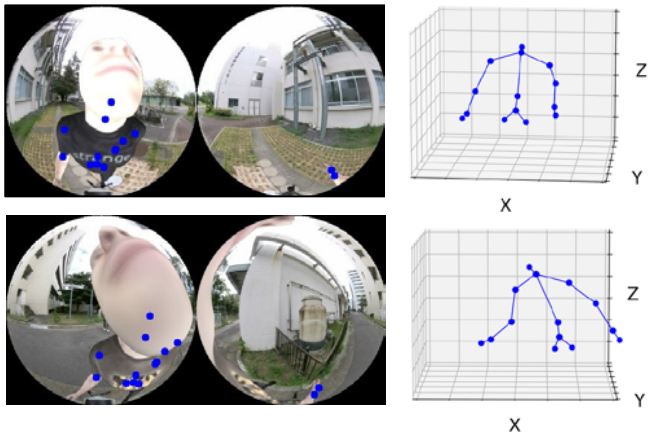

図 6.1　2D / 3D ポーズのアノテーション付き合成画像の例.

表 6.1　カメラ摂動を含む学習と評価用の合成データセット.

用途	背景画像	カメラ変動		データ数
		位置 (σ^2)	回転 (σ^2)	
学習	室内：40 室外：40	17.50 mm	10°	8,186
評価	室内：14 室外：10	0.00 mm 8.75 mm 17.50 mm 26.25 mm 35.00 mm	0° 5° 10° 15° 20°	2,088

を表 6.1 に示す.

6.3　調査対象の自己 3D ポーズ推定モデル

パイプラインアプローチを用いる自己 3D ポーズ推定モデルについて，カメラ摂動に対する頑健性を調査する．パイプラインアプローチは，既存の高精度な 2D ポーズ推定モデルを利用可能であること，画像を用いずに 3D ポーズのみを使って 3D ポーズ拡張部を学習できるという利点によって，学習データの収集とモデル学習の負担を軽減する手法である．本章では，3D 単位ベクトル化を組み込んだ推定モデル（提案モデル），Martinez らのフィードフォワードモデル [19]，Tome らの xR-EgoPose [23, 24] について評価する．

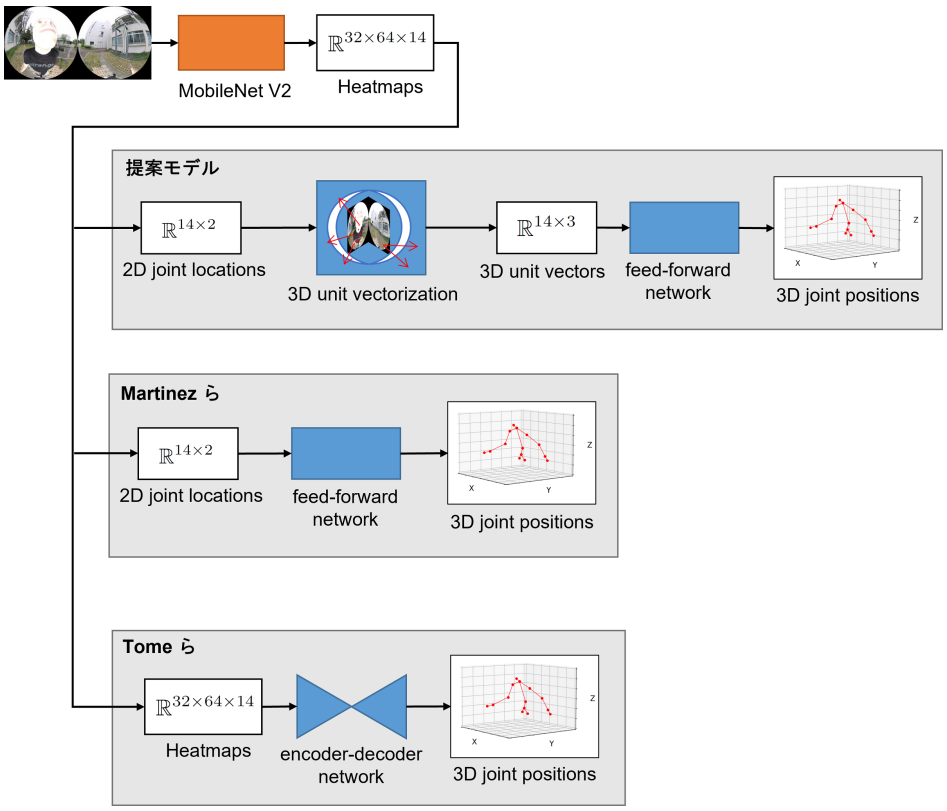

図 6.2　パイプラインアプローチを用いた自己 3D ポーズ推定モデルの全体構成.

パイプラインアプローチの 2D ポーズ推定部として MobileNetV2 [36] を用いる．2D ポーズ推定部は 128×256 ピクセルの合成画像を入力として，32×64 ピクセルのヒートマップを出力する．3D 単位ベクトル化を組み込んだ推定モデルの 3D ポーズ拡張部の重みと残差ブロックの繰り返しは 5.3 節と同じである．Martinez らの 3D ポーズ拡張部は，入力が 2D ポーズになっているが，重みと残差ブロックの繰り返しは提案モデルと同じである．xR-EgoPose は 32×64 ピクセルのヒートマップから，3D ポーズを推定するエンコーダ・デコーダモデルである．2D ポーズ推定部を含めたそれぞれの自己 3D ポーズ推定の構成を図 6.2 に示す．

6.4　推定モデルの学習と評価

自己 3D ポーズ推定の評価尺度として，3D ポーズを構成する関節座標の真値と推定値の誤差平均 (MJPE: Mean Joint Position Error) を用いる．また，2D ポーズ推定部の評価尺度として，2D ポーズを構成する平面上（ピクセル単位）の関節位置の真値と推定値の誤差平均 (MJLE: Mean Joint Location Error) を用いる．

まず 2D ポーズ推定部の MobileNetV2 の学習をおこなう．学習用データとして，

表 6.2　MJPE (mm) の評価結果 (位置 $\sigma^2 = 17.50$ mm，回転 $\sigma^2 = 10°$)．

model	head	neck	spine	pelvis	hips	shoulders	elbows	wrists	hands	all
提案モデル (VD loss)	81.44	61.56	74.60	86.72	102.39	66.42	109.93	**163.47**	**197.66**	113.15
提案モデル (L2 loss)	**26.55**	**30.82**	**70.83**	**82.84**	**96.58**	**43.36**	**99.19**	163.91	199.44	**101.14**
Martinez ら	35.27	41.79	89.48	98.59	119.03	57.41	113.19	168.96	204.46	113.66
xR-EgoPose	48.35	59.39	87.22	104.62	123.68	67.07	134.75	219.70	261.08	136.58

2D ポーズから生成したヒートマップと合成画像を用いる．初期学習率を 0.001 とした Adam を最適化アルゴリズムとして用いて，バッチサイズを 32，エポックを 140 とする．次に 3D 単位ベクトル化を組み込んだ推定モデル（提案モデル）の学習をおこなう．学習用データとして， 3D ポーズとそれを構成する関節への 3D 単位ベクトルを用いる．4.4 節で提案されている VD 損失関数は適切な係数を調査することが難しいため，VD 損失関数を用いた学習と， L2 損失関数を用いて学習した 2 つの推定モデルについて調査する．最適化アルゴリズム，バッチサイズ，エポックは 2D ポーズ推定部と同じである．xR-EgoPose は， 2D ポーズから生成したヒートマップと 3D ポーズを用いて学習をおこなう．また， Martinez らの推定モデルは， 2D ポーズと 3D ポーズを用いて学習をおこなう．それぞれの最適化アルゴリズム，バッチサイズ，エポックは提案モデルと同じである．

6.4.1　カメラ摂動に対する頑健性の評価

学習データと同じ摂動 (位置 $\sigma^2 = 17.50$ mm，回転 $\sigma^2 = 10°$) の評価データにおける結果を表 6.2 に示す．学習データの真値を使った 3D ポーズ拡張部においては， L2 損失関数で学習した提案モデルが最良の推定精度を示す．

段階的にカメラ摂動を大きくする評価データにおける 3D ポーズ推定の誤差平均 (MJPE) の推移を図 6.3 に示す．図中には 2D ポーズ推定部の評価を示すピクセル単位の 2D ポーズ推定の誤差平均 (MJLE) も示す．摂動が大きくなるに従って， 2D ポーズ推定部の推定結果が悪くなっている．特に，学習データの摂動より大きいカメラ変動の評価データ (位置 $\sigma^2 \geq 26.25$ mm，回転 $\sigma^2 \geq 15°$) において，2D ポーズ推定部の精度の悪化が大きくなる．3D ポーズ拡張部による 3D ポーズ推定の精度は， 2D ポーズ推定部の精度の悪化に従って悪くなる．提案モデル (L2 損失関数) はすべての摂動の評価データにおいて最良の推定結果を示す．すなわち， 3D 単位ベクトル化を組み込んだ推定モデルはカメラ摂動に対する頑健性が高いと言える．

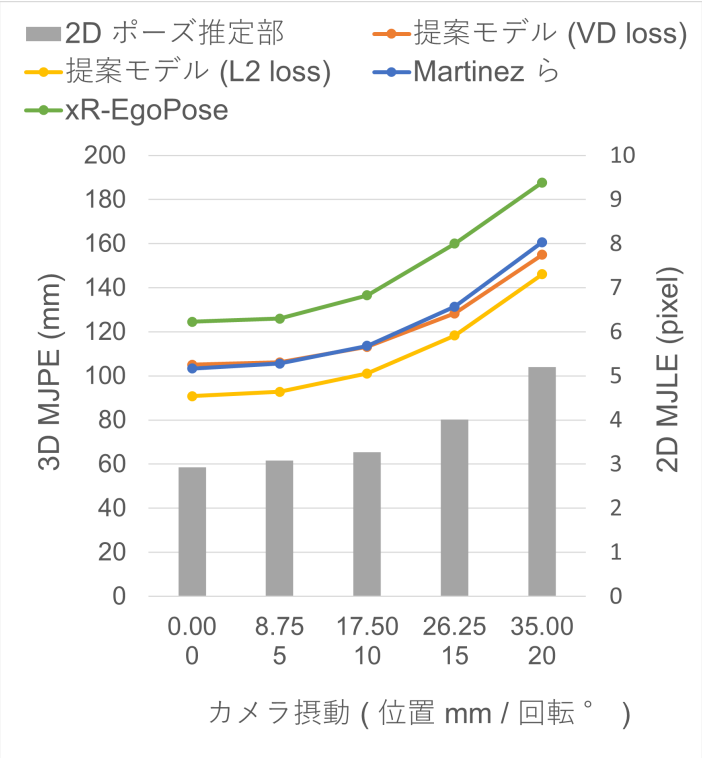

図 6.3　カメラの位置と回転の変動による推定誤差の推移.

6.4.2　2D ポーズの真値を入力とする 3D ポーズ拡張部の評価

2D ポーズ推定部の出力の代わりに， 2D ポーズの真値を入力とする 3D ポーズ拡張部の評価結果を図 6.4 に示す．2D ポーズ推定部の出力に真値を用いることで，パイプラインアプローチの 3D ポーズ拡張部のみに注目して推定精度と頑健性を評価する．ここで，図中に現れる 2D ポーズ推定部のわずかな誤差 (MJLE) は， 2D ポーズをピクセル単位に変換する際の丸め誤差によって発生している．提案モデル (L2 損失関数) と xR-EgoPose は学習データの摂動以下 (位置 $\sigma^2 \leq 17.50$ mm ，回転 $\sigma^2 \leq 10°$) の評価データで同程度の推定誤差となっている．しかし，提案モデル (L2 損失関数) は学習データより大きい摂動の評価データにおいても推定精度の悪化を抑えられている．すなわち， 3D ポーズ拡張部のみの評価においても， 3D 単位ベクトル化を組み込んだ推定モデルはカメラ摂動に対する頑健性が高いと言える．

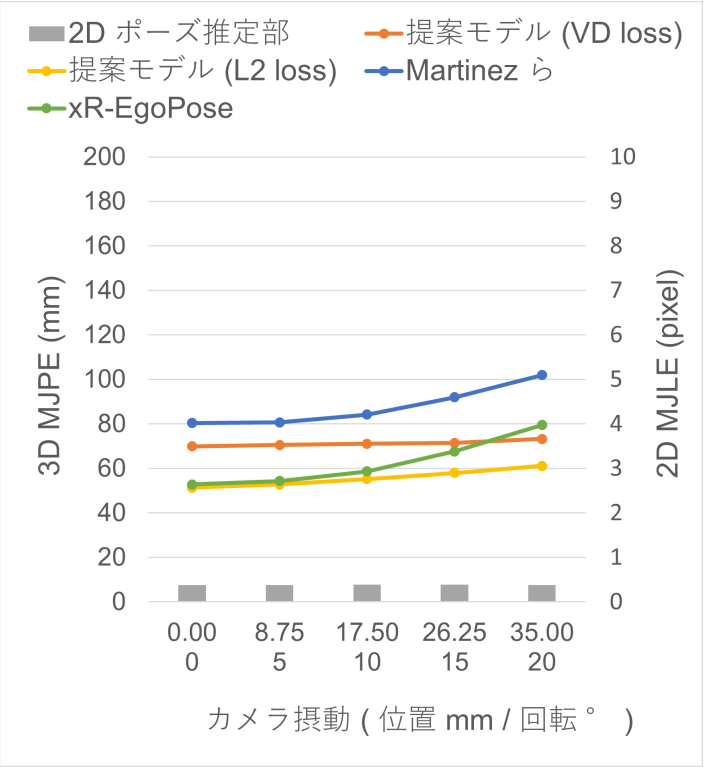

図 6.4　2D ポーズの真値を入力とする 3D ポーズ拡張部の評価結果.

6.4.3　2D ポーズ推定部の出力を用いて学習した 3D ポーズ拡張部の評価

xR-EgoPose は，2D ポーズ推定部によって得られる学習データのヒートマップを用いて学習することで，オクルージョンの発生しやすい複雑な姿勢への頑健性を持つ自己 3D ポーズ推定モデルとなる [23, 24]．学習データの合成画像を入力とする 2D ポーズ推定部の出力（ヒートマップ，もしくは 2D ポーズ）と，3D ポーズの真値を用いてそれぞれの推定モデルを学習する．段階的に摂動の大きくなる評価データにおける評価結果を図 6.5 に示す．すべての推定モデルにおいて真値で学習した場合（図 6.3 ）よりも良い推定精度となる．これは，2D ポーズ推定部が推定に失敗する場合に対して頑健な学習をしているためと考えられる．特に，xR-EgoPose は提案モデル (L2 損失関数) と同程度の推定精度を示す．しかし，依然として 3D 単位ベクトル化を組み込んだ推定モデル (L2 損失関数) はカメラ摂動に対する高い頑健性と推定精度を示す．

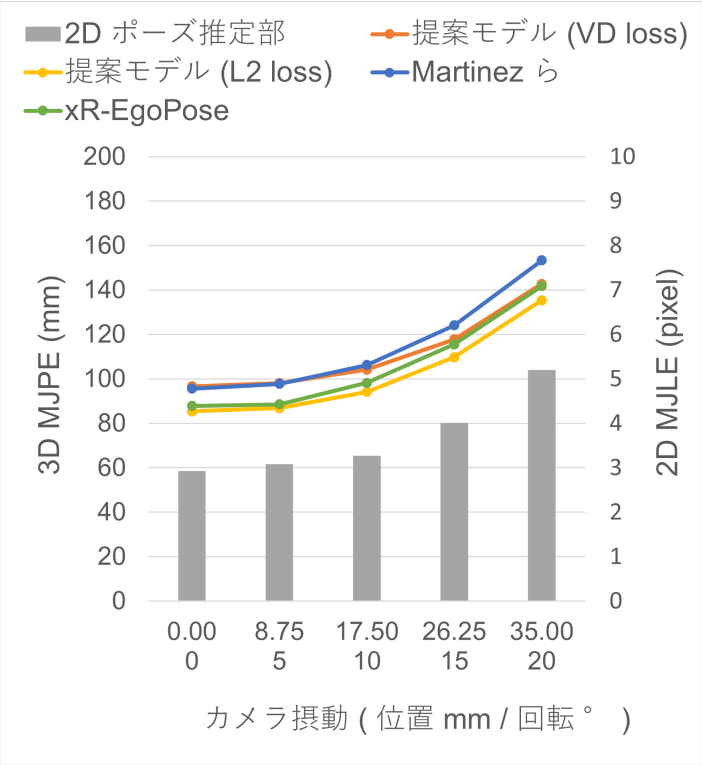

図 6.5　2D ポーズ推定部の出力を用いて学習した 3D ポーズ拡張部の評価結果.

6.5　結言

本章では，装着した全方位カメラの摂動に対する自己 3D ポーズ推定モデルの頑健性について述べた．カメラ摂動を再現するために，位置と回転の変動を段階的に大きくした評価用の合成データの生成について述べた．頑健性を調査する対象として，パイプラインアプローチを用いるいくつかの自己 3D ポーズ推定モデルについて述べた．生成した合成データセットを用いて，以下の学習と評価をおこなった．

- 真値を用いて学習した自己 3D ポーズ推定の評価
- 2D ポーズ推定部の出力に真値を用いて 3D ポーズ拡張部のみに注目する評価
- 2D ポーズ推定部の出力を用いて学習した 3D ポーズ拡張部の評価

本章を通して，本研究で提案する 3D 単位ベクトル化を組み込んだ自己 3D ポーズ推定モデルはカメラ摂動に対する頑健性が高いことを示した．

第7章

手話認識モデルへの適用

7.1 序言

本章では，全方位カメラを用いた装着型 MoCap システムによる，既存の手話認識モデルへの適用について述べる．7.2 節では，本章で用いる手話認識モデルについて述べる．7.3 節では，対面カメラと装着した全方位カメラを用いた手話認識モデルの学習と評価用のデータ収集について述べる．7.4 節では，手話認識モデルの入力形式に適用するための，全方位カメラで撮像した画像に対する 2D ポーズ推定について述べる．7.5 節では，対面カメラと全方位カメラから収集したデータを用いた手話認識モデルの学習と評価について述べる．最後に，7.6 節で本章についてまとめる．

7.2 適用する手話認識モデル

Li ら [1] の手話単語の認識モデル (Pose-TGCN) を用いる．Li らは，インターネットにあるアメリカ手話 (ASL) データベースや動画サイトから，手話単語の動画を収集することで大規模なデータセットを構築した．さらに，アーキテクチャの異なる複数の手話認識モデルを，構築したデータセットを用いて学習し，比較評価した．Pose-TGCN は， Li らの実験において， 2D ポーズを入力とするアーキテクチャの中で最良の精度を示した手話認識モデルである．収集した動画のフレーム画像の例を図 7.1 に示す．

図 7.1　Li ら [1] の収集した手話単語のフレーム画像の例.

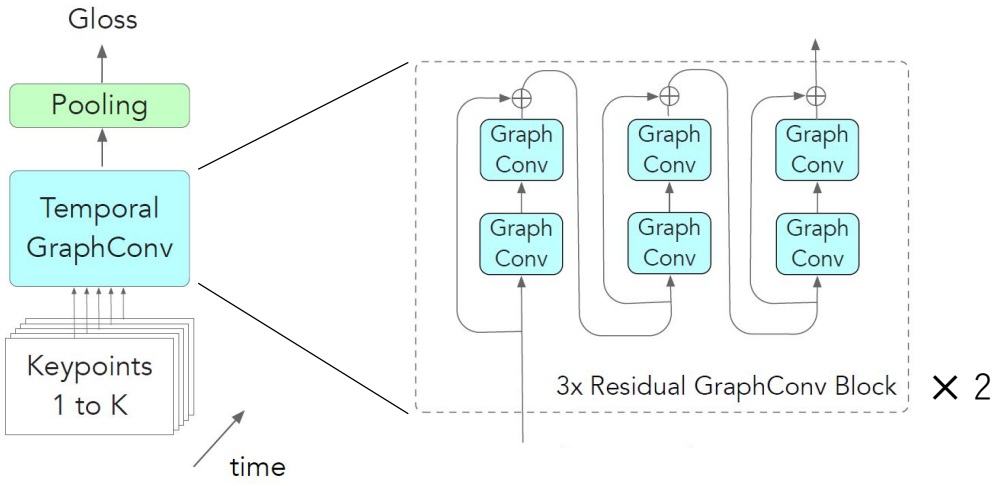

図 7.2　Li ら [1] の手話認識モデル (Pose-TGCN) のアーキテクチャ．

Pose-TGCN は動画内の人物から得られる 2D ポーズ (画像平面上の関節位置) の時系列データを，関節位置から得られるグラフとその系列で畳み込むネットワークモデル (Temporal Graph Convolution Networks) である．1 から T の連続フレームを考えた場合，入力は $\boldsymbol{X}_{1:T} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, ..., \boldsymbol{x}_T]$ より得られる．ここで，$\boldsymbol{x}_i \in \mathbb{R}^K$ は K 次元の 2D 関節座標を表す．Li らの用いた Pose-TGCN は 2 層のグラフ畳み込みネットワークブロックからなる．ここで，グラフ畳み込みネットワークブロックは 3 回の残差ネットワークを持つグラフ畳み込み層で構成される．本章で用いる手話認識モデルのアーキテクチャを図 7.2 に示す．

7.3　学習と評価用データの収集

全方位カメラを用いた装着型 MoCap システムを用いた手話認識モデル (Pose-TGCN) を学習・評価するために，全方位カメラで収録した画像から推定した単語ごとの時系列 2D ポーズを収集する必要がある．本節では，まず学習と評価用の手話単語の動画を収録する方法について述べる．

身体に装着した全方位カメラを用いて手話単語の動画を収録するために，5 章のヘッドバンド付き全方位カメラ (Insta360 Air) を用いる．また，同期した対面カメラからの動画を収録するために logicool STREAMCAM [1] を用いる．全方位カメラを装着した手話者の前方に対面カメラを設置し，両方のカメラを用いて手話単語の動画をフレー

[1] https://www.logicool.co.jp/ja-jp/products/webcams/streamcam.960-001301.html

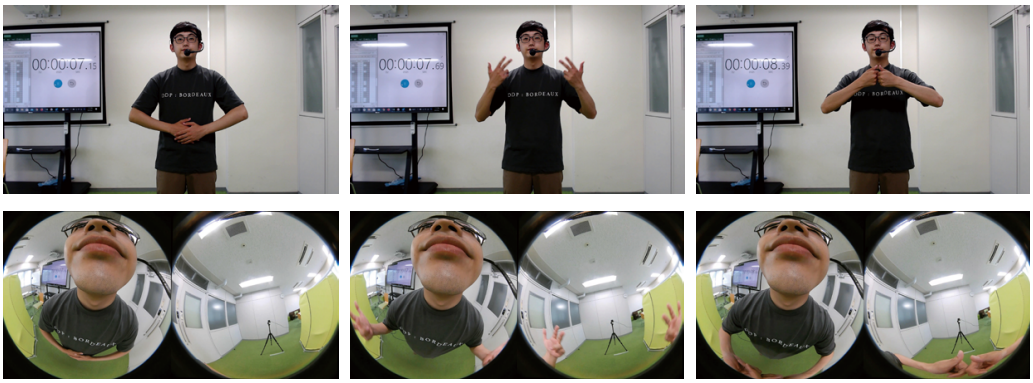

図 7.3　収録した手話単語動画の例（上段：対面カメラ，下段：全方位カメラ）．

ムレート 30 fps で収録する．両方のカメラから撮像できる位置にストップウォッチを設置し，表示されるカウンタを参照して，動画の収録後に 0.01 秒の精度で同期処理をおこなう．収録データの例を図 7.3 に示す．図中の上段は対面カメラ，下段は全方位カメラで撮像した動画のフレーム画像である．また，各列の対面カメラと全方位カメラの画像は同期している．

本章では，手話単語として Li [1] らの WLASL100 の 100 単語を対象とする（表 7.1）．7 人の協力者に，手話単語を 3 回ずつ実施してもらい，その様子を全方位カメラと対面カメラで収録する．収録した動画を同期処理した後に，さらに 1 回ずつの手話動作に分割する．これによって，協力者毎に 300 の手話動画を収録することになる．なお，協力者は全員が手話未習得者であり，事前に Li らのデータベースの動画を用いて，手話単語を練習してから収録を実施する．

7.4　対面カメラ画像と全方位カメラ画像のポーズ推定

収録した動画から，手話認識モデルの入力となるフレーム毎の 2D ポーズを収集する方法について述べる．対面カメラで撮像した画像は，高性能 2D ポーズ推定ツール OpenPose [2] を用いて，フレーム毎に上半身と手指の 55 点の関節位置を取得する．OpenPose で取得した対面カメラ画像の 2D 関節位置を平面上に図示した画像を図 7.4 に示す．図中において赤い点は推定した関節位置を示し，白線は関節位置をもとにした骨格を示す．

全方位カメラで撮像した画像から，手話認識モデルの入力となる 2D ポーズを取得するプロセスを図 7.5 に示す．まず，2D ポーズ推定の精度を向上するために，2D ポーズ推定部では OpenPose を用いて 8 つの関節（鼻，首，両肩，両肘，両手首）を推定す

表 7.1　手話認識の対象とする 100 単語.

accident	africa	all	apple	basketball	bed	before
bird	birthday	black	blue	book	bowling	brown
but	can	candy	chair	change	cheat	city
clothes	color	computer	cook	cool	corn	cousin
cow	dance	dark	deaf	decide	doctor	dog
drink	eat	enjoy	family	fine	finish	fish
forget	full	give	go	graduate	hat	hearing
help	hot	how	jacket	kiss	language	last
later	letter	like	man	many	medicine	meet
mother	need	no	now	orange	paint	paper
pink	pizza	play	pull	purple	right	same
school	secretary	shirt	short	son	study	table
tall	tell	thanksgiving	thin	thursday	time	walk
want	what	white	who	woman	work	wrong
year	yes					

図 7.4　OpenPose [2] を用いて取得した対面カメラ画像の 2D ポーズ.

る．OpenPose を用いた 2D ポーズ推定部の詳細は 7.4.1 節で後述する（図 7.5.(a) ）．全方位カメラ画像に対して OpenPose で推定できる関節は 8 つのみであるため， 3D ポーズ拡張部も同様に 8 関節として再学習をおこなった．再学習をおこなった 3D ポーズ拡張部， 3D 単位ベクトル化モジュールを組み合わせて， 2D ポーズ推定部によって出力された 2D 関節位置から 3D 関節座標を推定する．手指については MediaPipe [40] を用いて 3D 関節座標を推定する．2D ポーズ推定部で得られた手首の位置を中心として切り抜いた画像から， MediaPipe を用いて手指の 3D 関節座標を推定し， 3D ポーズ拡張部から出力される 3D 関節座標と結合する（図 7.5.(b) ）．得られた 3D 関節座標の XZ 平面を参照することで 2D 関節位置を取得する．なお，上述のプロセスで取得されない 5 関節（腰，両目，両耳）については，同期フレームの対面カメラ画像から取得した 2D 関節位置を用いて補完する（図 7.5.(c) ）．

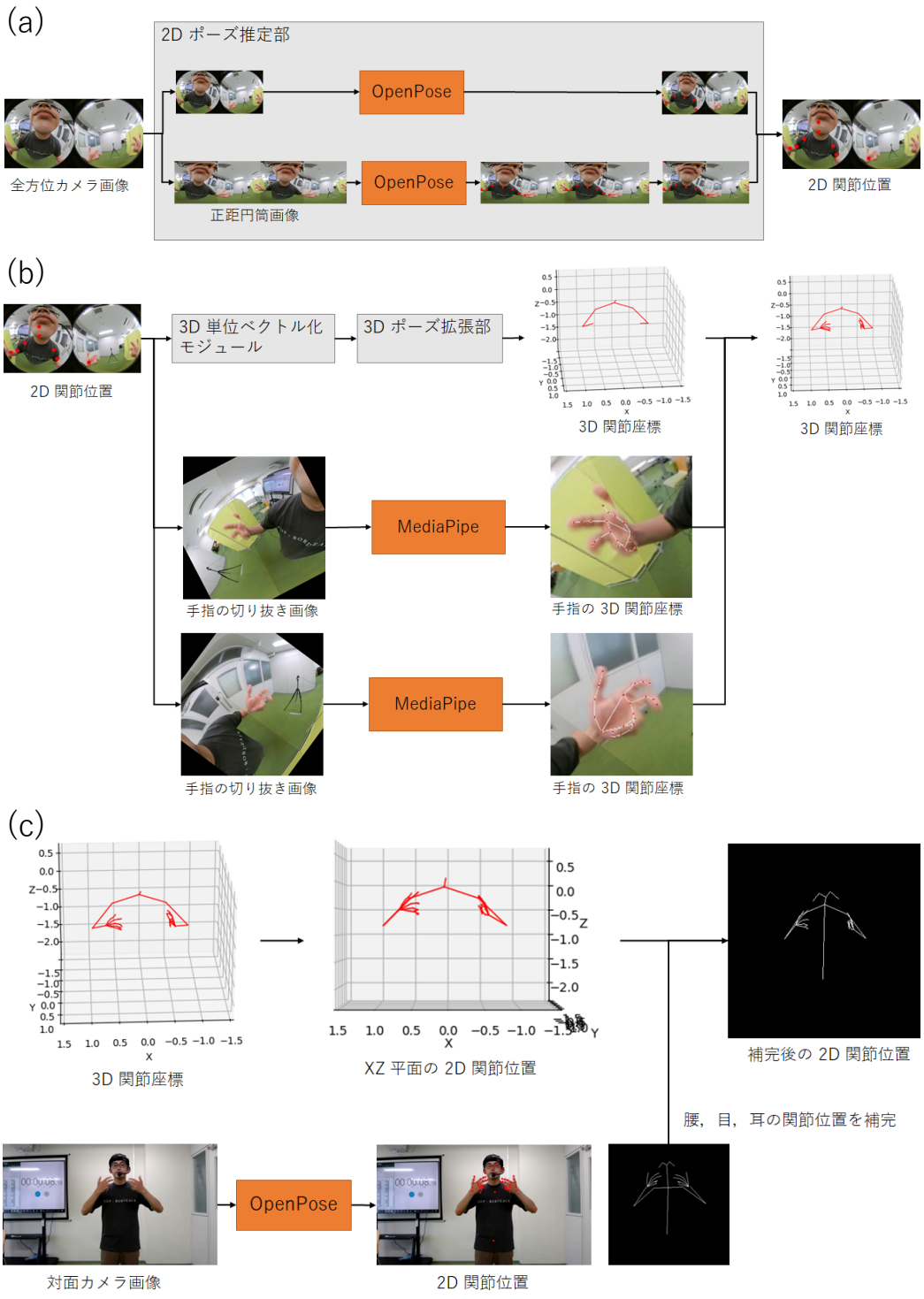

図 7.5　全方位カメラ画像から 2D ポーズを取得するプロセス.

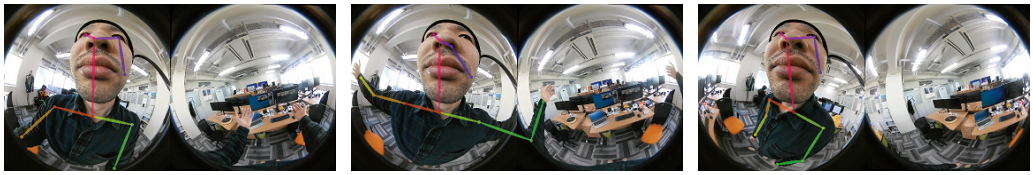

図 7.6　OpenPose を用いた 2D ポーズ推定の例.

7.4.1　OpenPose を用いた 2D ポーズ推定

5.6 節の定性評価において，合成データを用いた学習によって全方位カメラ画像から 3D ポーズを推定できることを示したが，図 5.6 に示すようにいくつかの場面で推定精度の低下が見られた．本節では，より高精度の推定をおこなうために OpenPose [2] を 2D ポーズ推定部に用いる方法を述べる．

全方位カメラ画像は 2 つの魚眼レンズで撮像された画像で構成されており，そのままの画像を入力すると OpenPose での推定に失敗する場合がある (図 7.6) ．図中の左列では，画像右側の魚眼レンズによって撮像された両手の推定に失敗している．中列の画像では，左肘の位置が左右の魚眼レンズで近いため，一人の人物として推定できている．右列の画像では，左側の魚眼レンズに上半身が収まっており，画像が歪んでいても 2D ポーズを推定できている．推定に失敗する場合に対応するために，全方位カメラで撮像した画像に前処理をしてから OpenPose による 2D ポーズ推定をする．

OpenPose を用いた 2D ポーズ推定をおこなうために，全方位カメラ画像に対して「正距円筒図法への変換」と「画像の結合」の前処理をおこなう．

正距円筒図法への変換　：全方位カメラの 2 つの魚眼レンズで撮像された画像について，それぞれの画像が連続するように正距円筒図法へと変換する．変換後の画像と OpenPose を用いた 2D ポーズ推定結果の例を図 7.7.(a) に示す．

画像の結合　：正距円筒図法への変換によって， OpenPose を用いた 2D ポーズ推定を部分的に可能にした．しかし，全方位カメラはその全周囲を正距円筒図法で撮像するにあたり，必ず画像平面上に境界が生じる．正距円筒画像を横方向に結合することで，画像中に境界によって分断されない人物が撮像され， OpenPose を用いた 2D ポーズを推定できる．推定した 2D ポーズの関節位置を組み合わせることで，結合前の正距円筒画像における 2D ポーズ推定の結果を得ることができる．画像の結合を用いた 2D ポーズ推定のながれを図 7.7.(b) に示す．図中の赤丸は，推定した関節位置を示す．

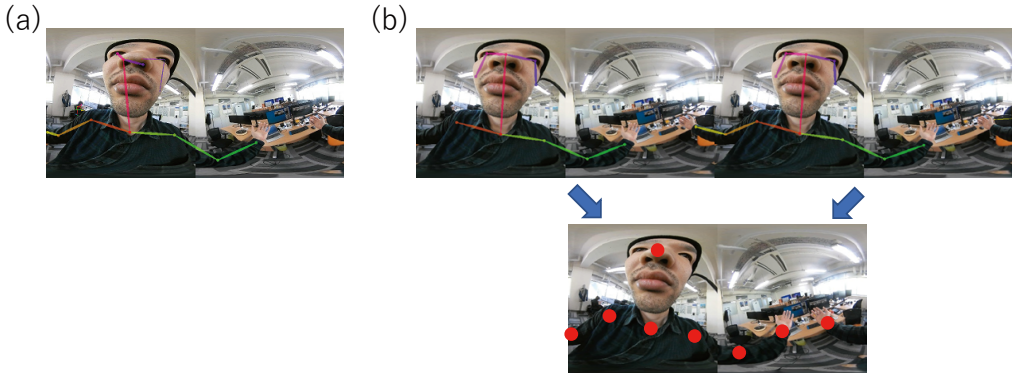

図 7.7　(a) 全方位カメラ画像から変換した正距円筒画像に OpenPose を用いた 2D ポーズ推定の例．(b) 結合した正距円筒画像に OpenPose を用いた 2D ポーズ推定の例．

正距円筒図法への変換と画像の結合をおこなうことで，一部の関節ではあるが OpenPose を用いた高精度の 2D ポーズ推定ができることを示した．7.4 節の全方位カメラ画像を入力とした 2D ポーズ推定部では，元画像と並行して，「正距円筒図法への変換」と「画像の結合」をした画像に対して OpenPose を用いた 2D ポーズ推定をおこない，精度の高い方の推定結果を 2D 関節位置として出力する（図 7.5.(a)）．

7.5　対面カメラと全方位カメラのデータを用いた学習と評価

対面カメラと全方位カメラのそれぞれで収集した 2D ポーズの系列データを用いて leave-one-out 交差検証をおこなう．すなわち，収集した 7 人のデータセットのうち，6 人分を用いて手話認識モデルを学習し，残りの 1 人分を用いてテストする．すべての協力者の組み合わせで学習とテストを実施し，その平均を用いて評価する．評価指標には top-N 分類精度を用いる．top-N 分類精度は分類問題においてモデルの推定結果の上位 N 以内に正解ラベルが含まれる確率を示す指標である．

Li [1] らと同様に， 収集した 2D ポーズを関節位置を頂点とした全結合グラフ（行列 $\boldsymbol{A} \in \mathbb{R}^{K \times K}$）として扱い，連続する 50 フレームを手話認識モデルの入力とする．Li らの学習済みモデルの重みを初期値として，バッチサイズ 64 ， 100 エポック，最適化アルゴリズムを Adam でファインチューニングする．top-N= $\{1, 5, 10\}$ の結果を表 7.2 に示す．ここでは，全方位カメラのデータで学習したモデルを全方位カメラのテストデータで評価し，対面カメラのデータで学習したモデルを対面カメラのテストデータで評価する．

表 7.2　全方位カメラと対面カメラを用いた手話認識モデルの評価結果.

	top-1	top-5	top-10
全方位カメラの手話動画	50.32	77.99	88.48
対面カメラの手話動画	85.48	98.27	99.33

評価結果より，全方位カメラの手話動画から収集したデータを用いた top-1 分類精度は 50.32% ，対面カメラから収集したデータを用いた場合は 85.48% であり，その他の top-N= $\{5, 10\}$ においても，対面カメラで収集したデータの方が良い分類精度を示した．これは，OpenPose のみを用いて 2D ポーズ推定をおこなった対面カメラの画像に対して，全方位カメラ画像は複数の手法を組み合わせているため，収集する 2D ポーズの精度が低いことが要因と考えられる．特に，全方位カメラ画像からの画像切り抜き時の境界処理や，身体との重なりにって MediaPipe を用いた手指の 3D 関節座標推定に失敗することが多い．全方位カメラ画像からの 2D ポーズ収集に失敗する例を図 7.8 に示す．図 7.8.(a) と (b) では右手が全方位カメラの境界によって潰れて， MediaPipe による推定に失敗している．図 7.8.(c) では両手の重なりによって，MediaPipe による推定に失敗している．

7.6　結言

本章では，身体に装着した全方位カメラで撮像した手話動作による，既存の手話認識モデルへの適用について述べた．まず， Li [1] らによる手話単語認識モデル (Pose-TGCN) について述べた．次に， Li らが用いたデータセット WLASL100 の 100 単語について，全方位カメラと対面カメラで同期した手話動画の収録方法について述べた．続けて，収録した手話単語の動画から，手話認識モデルの入力形式に従った 2D ポーズを収集する方法について述べた．全方位カメラと対面カメラの手話動画から収集した 2D ポーズの系列データを用いて，手話認識モデルを学習し，評価をおこなった．

本章を通して，対面カメラで収録して OpenPose で 2D ポーズを収集する手法に比べて，手話単語分類の精度は低くなっているものの，全方位カメラを用いた装着型 MoCap システムを利用した 2D / 3D ポーズ推定によって，既存の手話認識モデルを利用できることを示した．

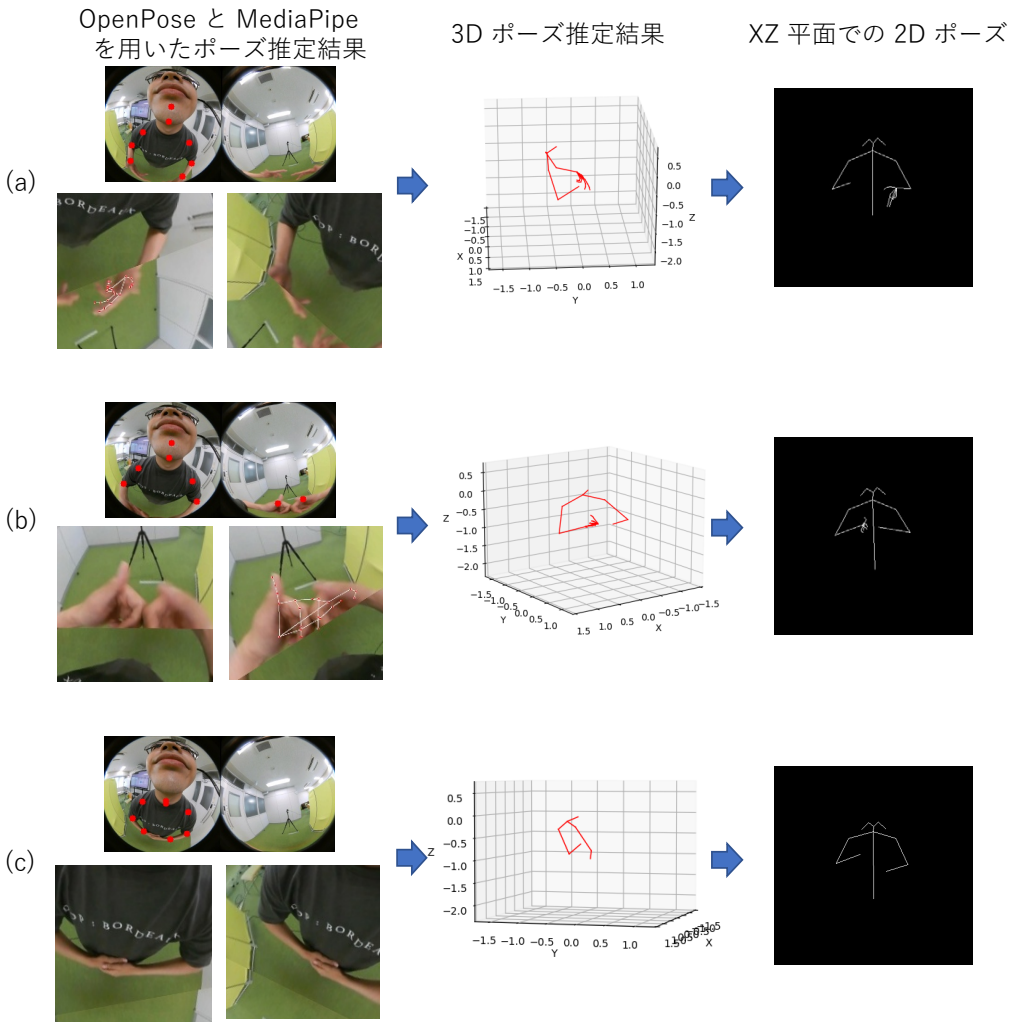

図 7.8　全方位カメラ画像からの 2D ポーズ収集の失敗例.

第8章

結論

8.1 本研究のまとめ

本論文は，全方位カメラを用いた装着型モーションキャプチャ (MoCap) システムに関する研究をまとめたものである．はじめに，画像認識などの情報技術を用いた手話の認識・翻訳に関する既存研究のいくつかの課題に対応できる「話者中心の手話認識と翻訳」について述べた．話者中心の手話認識と翻訳を実現するためには，手話者の身体動作と周辺環境の情報を同時に取得する必要があり，そのための仕組みとして全方位カメラを用いた装着型 MoCap システムを提案し，その要素技術の研究をおこなった．以下に研究のまとめである 3 章から 7 章の成果を要約し，最後に今後の課題を述べる．

3 章では，市販のネックマウントを使って身体に装着した全方位カメラ（Ricoh R Development Kit）で撮像される正距円筒図法の画像から 3D ポーズを推定する方法について述べた．データ収録システムを開発し，収集した画像と 3D ポーズのデータセットを用いて，畳み込みニューラルネットワークを用いた推定モデルを学習，評価した．大量の学習用データを収集することで，身体に装着した全方位カメラの画像から 3D ポーズを推定できることを確認し，自然な環境下で大量の学習データを集めることが大きな課題であることを示した．

4 章では， 3D ポーズ推定を 2D ポーズ推定部と 3D ポーズ拡張部に分離するパイプラインアプローチにおいて，3D 単位ベクトル化を組み込む推定モデルについて述べた．合成データ生成ツールを利用して大量の学習用データを人工的に作成し，全方位カメラ（Insta360 Air）を用いた 3D ポーズ推定モデルの学習と評価をおこなった．3D 単位ベクトル化を組み込んだ推定モデルは，既存の類似研究と比較して推定精度は最良ではないものの，既存研究にはない重要な利点があることを示した．その利点はデータセットの収集と学習の両方において 2D ポーズ推定部と 3D ポーズ拡張部を分離できることである．すなわち，2D ポーズ推定部に向けては 2D ポーズアノテーション付き画像を収集し， 3D ポーズ拡張部に向けては一般に公開されている 3D MoCap

データセットから，3D ポーズとその単位ベクトルを収集することでモデルを学習できる．それによって，身体に装着した全方位カメラの画像から 3D ポーズを推定するにあたっての大きな課題であった，自然な環境下での大量の学習データを収集する負担を軽減できることを示した．

5 章では，3 章・4 章で得られた知見を利用したプロトタイプシステムの開発について述べた．ヘッドバンドとグースネックケーブルを用いて小型の全方位カメラを装着し，GPU 搭載シングルボードコンピュータを用いたハードウェアを構成した．合成データ生成ツールを使って収集したデータセットを用いて学習した 3D ポーズ推定モデルをプロトタイプシステムのハードウェア上で動作させ，ポーズ推定の精度，カメラの位置と回転の摂動に対する頑健性，実行速度を評価した．製作したプロトタイプシステムは，3D ポーズ推定の精度，頑健性と実行速において改良が必要であった（詳細は 5.7 節を参照）．

6 章では，身体に装着した全方位カメラの摂動に対する自己 3D ポーズ推定モデルの頑健性について述べた．カメラ摂動を再現するために，カメラの位置と回転の変動を段階的に大きくした評価用の合成データセットを構築した．パイプラインアプローチを用いるいくつかの自己 3D ポーズ推定モデルに対して，合成データセットを用いて学習と評価をおこない，3D 単位ベクトル化を組み込んだ自己 3D ポーズ推定モデルがカメラ摂動に対する頑健性が高いことを示した．

7 章では，身体に装着した全方位カメラで撮像した動画による，Li [1] らの手話認識モデル（Pose-TGCN）への適用について述べた．Li らが用いたデータセット WLASL100 の手話単語について，身体に装着した全方位カメラと対面カメラで同期した動画を収録してデータセットを構築した．全方位カメラ画像について，4 章の 3D ポーズ推定モデルに OpenPose [2] と MediaPipe [40] を組み合わせて，手話認識モデルの入力形式に適用する 2D ポーズを収集した．また，対面カメラ画像については Li らと同様に OpenPose のみを用いて 2D ポーズを収集した．それぞれのカメラから収集したデータセットを用いて，手話認識モデルを学習し，評価をおこなった．対面カメラと OpenPose で 2D ポーズを収集する方法に比べて，全方位カメラを用いる方法は手話単語分類の精度は低くなったものの，全方位カメラを用いた装着型 MoCap システムを利用した 2D / 3D ポーズ推定によって，既存の手話認識モデルを利用できることを示した．

8.2 本研究の貢献

身体動作情報を取得するための全方位カメラを用いた装着型 MoCap システムに関する成果が本研究の主な貢献である．ジャイロセンサ等のついた衣服を身につける場合に比べ，小型の全方位カメラを装着する負担は小さくなるため長時間の利用を期待できる．また，カメラの位置を工夫することで，顔の表情などのセンサの取り付けが困難な部位についてもデータ収集が可能である．さらに，全方位カメラを用いることで，装着者の身体動作だけでなく周辺の環境を画像情報として同時に取得できる．身体動作と周辺の人・物体の情報を同時に取得し，関係性を含めた処理をおこなうことで，話者中心の手話認識・翻訳だけでなく，さまざまな用途への応用を期待できる．たとえば，日常生活において飲料を飲む身体動作をしている場合に，何を飲んでいるかの情報を画像から判断することも可能となる．

本研究の最も大きな貢献は， 4 章の 3D 単位ベクトル化による 2D ポーズ推定部と 3D ポーズ拡張部の分離である．ディープラーニングを中心とした近年の機械学習の発展により， 3D ポーズ推定においては大量のアノテーション付きデータセットの存在が最も重要である．しかし，身体に装着した全方位カメラ（もしくは魚眼カメラ）の画像と同期した 3D ポーズを収集することは，高性能なモーションキャプチャ環境を利用しても非常に時間のかかる作業である．そのため，自然な環境下で収録した大量のデータは存在せず，本研究を含めた既存研究では合成データ生成ツールを用いて大量のデータセットを生成し，学習をおこなった．3D 単位ベクトル化を組み込んだ 3D ポーズ推定モデルはデータの収集と学習の両方において 2D ポーズ推定部と 3D ポーズ拡張部を分離できる．すなわち，2D ポーズ推定部に向けては 2D ポーズのアノテーション付き画像を収集し， 3D ポーズ拡張部に向けては一般に公開されている 3D MoCap データから 3D ポーズとその 3D 単位ベクトルを収集することでそれぞれのモデルを学習できる．このデータ収集の分離によって，身体に装着した全方位カメラの画像から 3D ポーズを推定するにあたっての大きな課題であった，自然な環境下でのデータ収集の負担を軽減できる．

8.3 今後の課題

今後の課題として大きく 2 つをあげる．1 つ目は，全方位カメラ（もしくは魚眼カメラ）を用いた装着型 MoCap システムによる 3D ポーズ推定の精度と実行速度の向上である．5 章のプロトタイプ開発と評価によって示されたように，カメラの位置と回転の摂動やオクルージョンに対する頑健性と実行速度の向上が今後の課題である．

2 つ目は，提案システムを適用する手話の認識・翻訳モデルに関する研究である．7 章において， 2D ポーズを入力とする手話単語を対象とする認識モデルへの適用について述べた．より難しい既存の手話文の認識・翻訳モデルについての適用を検討すると共に，話者中心の手話認識・翻訳モデルについても検討が必要である．すなわち，周辺環境を画像情報として利用できることや，特定の利用者に最適化できるといった装着型 MoCap システムの利点を活かした話者中心の手話認識・翻訳モデルの研究が今後の課題である．

謝辞

本研究を遂行し学位論文としてまとめるにあたり，多くの方のご支援，ご協力を賜りました．皆様に感謝の言葉を申し上げます．

まず，名古屋工業大学 大学院工学研究科 情報工学専攻の酒向 慎司准教授に厚く御礼申し上げます．2018 年 3 月に酒向准教授が面識のなかった私とお会いくださり，博士後期課程での指導教員を引き受けて頂けたことから私の研究の道が拓かれました．酒向准教授には研究活動に限らず，さまざまな場面でご指導，ご助力頂きました．心より感謝致します．

同専攻の佐藤 淳教授，玉木 徹教授には，本論文の副査となることを快く承諾して頂き，審査会においては貴重なご意見，ご助言を頂きました．また，同専攻の坂上 文彦准教授にも，審査会にご出席頂き，貴重なご意見，ご助言を頂きました．心より感謝致します．

九州工業大学 大学院情報工学研究院 知能情報工学研究系の齊藤 剛史教授には，本論文の外部審査員となることを快く承諾して頂き，福祉分野への応用の立場から，貴重なご意見，ご助言を頂きました．心より感謝致します．

名古屋工業大学 大学院工学研究科 情報工学専攻の李 晃伸教授に厚く御礼申し上げます．李教授・酒向准教授が運営する研究室に所属したことで，快適に，集中して研究活動に取り組むことができました．秘書の花田さんには事務手続きなど，さまざまな場面でご助力頂きました．また，李・酒向研究室の学生の皆さんとの交流が，私の博士後期課程をより充実したものにしてくれました．心より感謝致します．

同専攻の徳田 恵一教授には，博士後期課程の入学にあたり，さまざまなご助力を頂きました．同専攻の片山 喜章教授には， 2011 年に修了した博士前期課程での指導教員として研究指導をして頂きました．また，博士後期課程に入学後も折にふれて気にかけて頂きました．東京工業大学 情報理工学院 情報工学系の篠田 浩一教授には，着手前の研究の方向性や進路についてご相談させて頂き，酒向准教授をご紹介頂きました．皆様にご助力頂いたおかげで，博士後期課程に入学し，研究に取り組むことができました．心より感謝致します．

最後に博士後期課程への進学について理解し，生活を支えてくれた家族に心より感謝致します．

2023 年 1 月　三浦 哲平

参 考 文 献

[1] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1448–1458, New York, NY, USA, March 2020. IEEE.

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, January 2021.

[3] Runpeng Cui, Hu Liu, and Changshui Zhang. A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training. *IEEE Transactions on Multimedia*, 21(7):1880–1891, 2019.

[4] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018.

[5] Necati C. Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10020–10030, New York, NY, USA, June 2020. IEEE.

[6] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Multi-channel Transformers for Multi-articulatory Sign Language Translation. In *European Conference on Computer Vision*, pages 301–319. Springer, 2020.

[7] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs. *International Journal of Computer Vision*, page 1311 – 1325, October 2018.

[8] Zheng Lihong, Liang Bin, and Jiang Ailian. Recent Advances of Deep Learning for Sign Language Recognition. In *International Conference on Digital Image Computing: Techniques and Applications*, pages 1–7, November 2017.

[9] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *ACM Transactions on Graphics*, 36:1–14, July 2017.

[10] Gary J Grimes. Digital Data Entry Glove Interface Device, November 1983. US Patent 4,414,537.

[11] Shinichi Tamura and Shingo Kawasaki. Recognition of Sign Language Motion Images. *Pattern Recognition*, 21(4):343–353, 1988.

[12] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In *International Conference on Language Resources and Evaluation*, pages 1911–1916, 2014.

[13] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured Prediction of 3D Human Pose with Deep Neural Networks. In *British Machine Vision Conference*, pages 130.1–130.11, September 2016.

[14] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1263–1272, July 2017.

[15] Sijin Li and Antoni B. Chan. 3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network. In *Asian Conference on Computer Vision*, pages 332–347, 2015.

[16] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep Kinematic Pose Regression. In *European Conference on Computer Vision*, pages 186–201, 2016.

[17] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D Human Pose Estimation in

the Wild Using Improved CNN Supervision. In *International Conference on 3D Vision*, pages 506–516, October 2017.

[18] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-Shot Multiperson 3D Pose Estimation from Monocular RGB. In *International Conference on 3D Vision*, pages 120–130, September 2018.

[19] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A Simple Yet Effective Baseline for 3d Human Pose Estimation. In *IEEE International Conference on Computer Vision*, pages 2659–2668, October 2017.

[20] Zhou Xiaowei, Zhu Menglong, Leonardos Spyridon, and Daniilidis Kostas. Sparse Representation for 3D Shape Estimation: A Convex Relaxation Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1648–1661, August 2017.

[21] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. EgoCap: Egocentric Marker-Less Motion Capture with Two Fisheye Cameras. *ACM Transactions on Graphics*, 35(6), November 2016.

[22] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo2Cap2: Real-time Mobile 3D Motion Capture with a Cap-mounted Fisheye Camera. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):2093–2101, May 2019.

[23] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xR-EgoPose: Egocentric 3D Human Pose From an HMD Camera. In *The IEEE International Conference on Computer Vision*, pages 7727–7737, October 2019.

[24] Denis Tome, Thiemo Alldieck, Patrick Peluse, Gerard Pons-Moll, Lourdes Agapito, Hernan Badino, and Fernando De la Torre. SelfPose: 3D Egocentric Pose Estimation from a Headset Mounted Camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, October 2020.

[25] Davide Scaramuzza, Agostino Martinelli, and Roland Siegwart. A Toolbox for

Easily Calibrating Omnidirectional Cameras. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5695–5701, October 2006.

[26] Hao Jiang and Kristen Grauman. Seeing Invisible Poses: Estimating 3D Body Pose from Egocentric Video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3501–3509, July 2017.

[27] 竹村茂. **手話・日本語大辞典**. 廣済堂出版, 1999.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, June 2016.

[29] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep High-Resolution Representation Learning for Human Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5686–5696, June 2019.

[30] Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, volume 3, May 2015.

[31] Blender. https://www.blender.org/.

[32] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from Synthetic Humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4627–4635, July 2017.

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6), October 2015.

[34] Carnegie Mellon University Motion Capture Database. http://mocap.cs.cmu.edu/.

[35] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, June 2014.

[36] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, New York, NY, USA, June 2018. IEEE.

[37] Yuji Nagashima. Constructing a Japanese Sign Language Database and Developing Its Viewer. In *International Conference on PErvasive Technologies Related to Assistive Environments*, page 321 – 322, New York, NY, USA, June 2019. ACM.

[38] Keiko Watanabe, Yuji Nagashima, Daisuke Hara, Yasuo Horiuchi, Shinji Sako, and Akira Ichikawa. Construction of a Japanese Sign Language Database with Various Data Types. In *International Conference on Human-Computer Interaction*, volume 1032, New York, NY, USA, July 2019. Springer.

[39] Neva Cherniavsky. *Activity Analysis of Sign Language Video for Mobile Telecommunication*. PhD thesis, University of Washington, 2009.

[40] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A Framework for Building Perception Pipelines, June 2019.

研究業績

学術論文 (査読あり)

- Teppei Miura and Shinji Sako. “Simple yet effective 3D ego-pose lift-up based on vector and distance for a mounted omnidirectional camera”, Applied Intelligence, May 2022. DOI:10.1007/s10489-022-03417-3
- Teppei Miura and Shinji Sako. “3D human pose estimation model using location-maps for distorted and disconnected images by a wearable omnidirectional camera”, IPSJ Transaction on Computer Vision and Applications, vol.12, no.4, August 2020. DOI:10.1186/s41074-020-00066-8

国際会議 (査読あり)

- (ショート) Teppei Miura, Shinji Sako, and Tsutomu Kimura. “3D Ego-Pose Lift-Up Robustness Study for Fisheye Camera Perturbations”, In The 18th International Conference on Computer Vision Theory and Applications (VISAPP2023). 掲載予定（採択通知あり）.
- (ポスター) Teppei Miura and Shinji Sako. “SynSLaG: Synthetic Sign Language Generator”, In The 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS ’21), no.90, pp.1-4, October 2021. DOI:10.1145/3441852.3476519

研究会

- 三浦哲平，酒向慎司．“周辺環境を含むライフログ収集をめざした装着型 MoCap システムの提案”，電子情報通信学会技術研究報告，HCG シンポジウム， 2021-12-15.
- 三浦哲平，酒向慎司．“3D モーションデータを用いた手話データ生成ツール”，電子情報通信学会技術研究報告 vol.121 no.203，福祉情報工学研究会 2021-10 pp.1-5，

2021-10-12.

- 三浦哲平，酒向慎司．“手話認識への応用を目的としたモバイル MoCap システム〜 OpenPose を利用した 3D ポーズ推定の精度向上〜”，電子情報通信学会技術研究報告 vol.121 no.52，福祉情報工学研究会 2021-11 pp.54-58，2021-06-01.
- 三浦哲平，酒向慎司．“全天球カメラを用いた 3D ポーズ推定〜手話認識への応用に向けて〜”，電子情報通信学会技術研究報告 vol.120 no.161，福祉情報工学研究会 2020-7 pp.9-14，2020-09-01.
- 三浦哲平，酒向慎司．“ウェアラブルな全方位カメラの画像を入力とした 3D ポーズ推定 〜手話の認識と翻訳に向けて 〜”，電子情報通信学会技術研究報告 ISSN 2432-6480 vol.119 no.235，パターン認識・メディア理解研究会 PRMU2019-32 pp.5-10，2019-10-11.

助成金

- 公益財団法人 中部電気利用基礎研究振興財団, 出版助成 (2020).
- JST 次世代研究者挑戦的研究プログラム JPMJSP2112 (2021/10 - 2022/3).

ツール，データセットの公開

- (ツール) “SynSLaG: Synthetic Sign Language Generator” `https://github.com/TeppeiMIURA/SynSLaG`
- (データセット) “Simple yet effective 3D ego-pose lift-up based on vector and distance for a mounted omnidirectional camera” `https://drive.google.com/drive/u/1/folders/1ps92B1bYN5QuAWQfrq2Db8K3mQw5gqny`
- (データセット) “3D human pose estimation model using location-maps for distorted and disconnected images by a wearable omnidirectional camera” `https://drive.google.com/drive/u/1/folders/1XcQuS1dhSYggvb05CbQH-Qgd5OcBGyOe`