様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 １３ 日現在

機関番号：８２６２６
研究種目：若手研究
研究期間：2020～2023
課題番号：２０Ｋ１９８３６

研究課題名（和文）自己教示学習を用いた人の体型と姿勢の3次元推定

研究課題名（英文）Self-supervised learning of 3D human body shape and pose

研究代表者
吉安　祐介（Yoshiyasu, Yusuke）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・主任研究員

研究者番号：１０７１２２３４

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究では、単眼画像から人の体型と姿勢の三次元推定を行う課題において、これまでボトルネックとなっていた画像に対して３D体型姿勢情報を付与する労力を軽減する自己教示学習について研究した。まず、研究分野で活用されている人間の画像と３Dデータセットを収集・準備し、トランスフォーマを用いた３D体型・ポーズ推定モデルの構築を行った。その上で、３Dラベルの使用を軽減する①Masked autoencoder (MAE)という自己教示学習手法で学習した画像特徴抽出器を用いる方法、②２D画像と３Dモデルの密な対応付け情報を用いる学習方法と③三次元形状生成モデルを用いる方法を開発した。

研究成果の学術的意義や社会的意義
開発したトランスフォーマを用いた３D体型・ポーズ推定モデルは、ベンチマークにおいて高い性能を示し、コンピュータビジョン分野のトップ国際会議CVPR2023にも採録されており、学術的な意義が高い。また、MAEという自己教示学習を用いる方法、２D-3D間の密な対応付け情報を用いる方法、三次元生成モデルを用いる方法は学習に必要となる３Dラベルを軽減するという点で有用である。

研究成果の概要（英文）：In this study, we investigated self-supervised learning for reducing the effort of annotating 3D body shape and pose information to images, which has been a bottleneck in the task of learning 3D human body shape and pose from images. First, we collected and prepared image and 3D datasets for 3D human mesh recovery commonly used in the research field and built a model for estimating 3D body shape and pose using a transformer. Then, we developed three methods to reduce the use of 3D labels: 1) a method using an image feature extractor learned by a self-supervised learning method called masked autoencoder (MAE), 2) a learning method using dense correspondence information between 2D images and 3D models, and 3) a method using a 3D shape generation model.

研究分野：三次元コンピュータビジョン

キーワード：深層学習　自己教師学習　体型　３Dポーズ　機械学習

１．研究開始当初の背景

深層学習の発展により、一枚の画像から人の体型と身体 3 次元姿勢を同時に推定することが可能になってきている。しかし、これまで提案されてきた人の 3D 認識技術は、画像分類、物体検出や 2D ポーズ認識と比較して、十分な精度や汎用性を発揮するに至っていない。その理由は、体型や姿勢の 3 次元学習に必要な大規模なデータセット、すなわち、自然な状況で撮影された画像にラベル付けして得られる「画像と 3D ラベルのペア」を大量に揃えることが極めて難しいことに起因する。

これまで提案された既存技術の多くは、事前に構築された体型の統計モデルに関するパラメータ（体型、関節角度、カメラパラメータ）を画像から回帰し、画像に対して身体形状のメッシュモデルを当てはめるという方法を用いる。2018 年に提案された Human Mesh Recovery という方法では、人手で 2D 画像上にラベル付けした関節点と、推定した 3D 身体モデルの関節点の投影位置を最小化することで、画像から体型と姿勢を学習する。しかし、体型統計モデルの範囲が青年に限られるため、こどもや老人などの幅広い年齢の体型を表現することができないことや、ニューラルネットを用いて画像入力から体型パラメータを推定することは非線形性が高く容易でないことから、必ずしも推定した体型モデルが画像に対して重なり合わない。シルエットなどを用いることで推定精度は高まるが、シルエットを手作業でラベリングするには労力を要する。一方、三次元計測技術を用いて画像と 3D データを同時に収集する方法もあるが、Kinect などの光投影式の形状計測装置は屋外における広範囲な計測が難しく、モーションキャプチャ（MoCap）や慣性センサ（IMU）などの接触式計測装置は体にセンサやマーカを貼り付けなければならないなど、これらの方法を自然な状況下での大規模なデータ収集に利用することは難しい。実際、この分野で頻繁に用いられる Human3.6M などのデータセットは、タイトな衣服を着用した数名のアクターが行った十数種類の動作を撮影した動画像と MoCap データから構成されており体型や服装などの多様性に欠ける。

２．研究の目的

本研究課題では、人の体型と姿勢の 3D 学習に要するラベリング作業を軽減する自己教示学習について研究する。この方法では、従来の教師あり学習のように推定結果と教師データの誤差を取るのではなく、異なる画像や異なるタスクについて自らが推定した結果を用いて、それらが整合するように作用する誤差や制約項を導入することで、ラベリングなしで深層学習ができる。また、学習時に 3D ラベルを用いないことから、屋外で撮影した画像を利用でき、より自然な状況の再現や対象の多様性も確保できる。本研究では、幅広い体型と 3D 姿勢を一枚の画像から推定する技術の学習を簡便にする自己教示学習を用いた方法を研究する。

３．研究の方法

本課題では、まず、研究分野で活用されている人間の画像と３D データセットを収集・準備し、単眼画像から体型と３D ポーズを生成する学習モデルを構築学習した後に、これまでボトルネックとなっていた画像と３D ラベルのペアデータセットの使用を軽減する学習方法や復元対象の多様性を向上する方法について検討考察するという研究方針とした。具体的には、（１）学習データセットとベースラインモデルの整備と（２）トランスフォーマを用いた体型ポーズ推定モデルの構築を行う。次に、データ効率を向上する（３）２D 情報を用いたアプローチと（４）自己教示学習と生成モデルを用いたアプローチを研究し、（５）研究総括する。

４．研究成果

（１）学習データセットとベースラインモデル整備

学習基盤構築のために、産総研 AI 橋渡しクラウド(AI Bridging Cloud Infrastructure、ABCI)上に、既存の身体姿勢・体型データセットをダウンロードし、三次元姿勢・体型の深層学習に使用できるように、画像、アノテーション、ベースラインモデルを整備・整理した。具体的には、MeshTransformer や MeshGraphormer を参考に、人間の画像に対して、3D 疑似ラベルを付与した Human3.6M、COCO、MuCO、UP3D、MPII などのデータセットを準備した。加えて、表情や手指の動きを含むエクスプレッシブなメッシュ形状を復元するために、SMPL-X 体型モデルと画像に対して人体三次元メッシュをラベリングしたデータセットである NeuralAnnot を導入した。

（２）トランスフォーマを用いた体型ポーズ推定モデル

単眼画像から人体三次元形状（体型とポーズ）を復元する学習モデル、Deformable mesh transFormer（DeFormer)を構築した（図１）。DeFormer は、Transformer デコーダ内にフィードバックループを形成し、入力画像に対してメッシュモデルを適合させる。デコーダは、1）疎なセルフアテンションと 2）変形メッシュクロスアテンションという身体メッシュ駆動型の効率的なアテンションモジュールからなり、標準的な Transformer のアテンションを用いた従来法では計算コストが高く活用が困難であった高解像度の画像特徴マップと高密度のメッシュモデルを効果的に利用することができる。その結果、Human3.6M と 3DPW というベンチマークにおいて

従来法を上回る性能を発揮した。また、手の三次元形状復元のベンチマーク Freihand や二次元キーポイントを推定する COCO ベンチマークにおいても従来法よりも高い性能を確認した。さらに、HRFormer という視覚トランスフォーマをバックボーンモデルとして導入することで、さらなる性能の向上を図った。

（３） 2D 情報を用いたアプローチ

３D ラベリングを簡略化する試みの１つとして、人体３次元姿勢・体型を推定するニューラルネットワークを 2D 画像にラベル付けされた密な対応付けに基づいて３次元身体姿勢・体型を学習する Deform&Learn という方法を開発した（図２）。これにより、モーションキャプチャなどの３次元情報を直接使わずに学習が可能となった。また、2D 画像上にアノテーションされた密な対応付けを教師データとして用いる身体３次元姿勢・体型学習手法（Deform&Learn）の拡張をおこなった。４視点から撮影した画像データから 2D-3D の密な対応付けを推定し、４視点すべての対応付け情報を同時に用いて非剛体レジストレーションを行い３次元体型・姿勢を復元した。復元した３次元データを疑似的な教師データとして用いることで、単眼画像から体型と３次元の姿勢を推定する深層学習モデルを学習できる。これにより、モーションキャプチャなどの高価な装置で計測した３D データを直接学習データとして用いなくとも、身体３次元推定が比較的精度よく実現できるようになり、従来法に比べて推定精度も向上した。

（４）自己教示学習と生成モデルを用いたアプローチ

MAE という自己教師学習技術で事前学習された Vision Transformer(ViT)モデルである ViTPose を DeFormer のバックボーンモデルとして用い、単眼画像からの人体三次元復元を試みた。画像から end-to-end に三次元復元することで比較的良好な復元結果を得ることができた。加えて、拡散モデルを用いた三次元人間形状生成技術を開発し、単眼画像からの人体３D 復元に応用した。この方法では、まず、画像データから３D ポーズを復元し、３D ポーズを条件とした拡散モデルにより三次元メッシュを生成する。したがって、画像-３D モデルのペアリングデータを用いる必要がない。従来のようにペアデータを用いてトランスフォーマモデルを学習した場合に比べて表面のノイズを軽減でき、ほぼ同等の精度（関節位置推定誤差での評価）が得られることがわかった。

（５）研究総括

本課題で開発したトランスフォーマに基づく人体三次元学習技術は、これまでの手法に比べて画像に対してメッシュを高精度に位置合わせできる。一方で、学習データのスケール、モデルの計算効率が課題となる。本課題で提案した、自己教示学習バックボーンモデルの利用、２D 情報を用いた３D 学習データに頼らない学習方法論、生成モデルを用いた３D メッシュに頼らない学習方法論は人間３D 学習におけるデータ効率向上の有用であると考える。今後は、リアルタイム動作を実現する計算速度の向上、効率的な学習フレームワークのスケールアップ手法に加えて、ディテールや色の表現、多人数復元、動物など多様な対象の復元等さらなる研究の推進と発展が期待される。

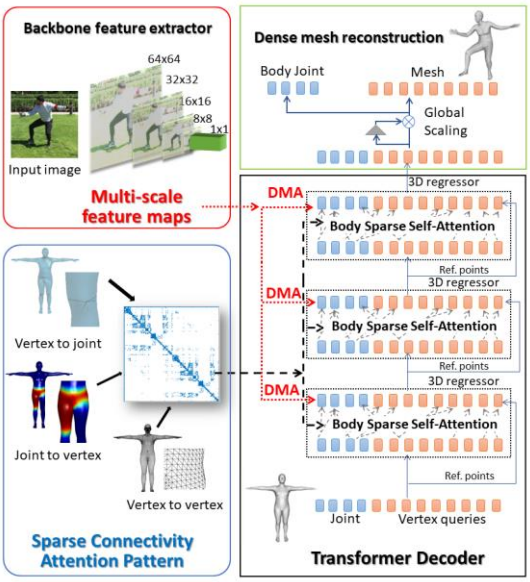

図１　トランスフォーマを用いた
人間３D 形状推定モデル

図２　密な対応付けを用いる方法

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Yoshiyasu Yusuke	４．巻 -
２．論文標題 Deformable Mesh Transformer for 3D Human Mesh Recovery	５．発行年 2023年
３．雑誌名 In proc. of Computer Vision and Pattern Recognition	６．最初と最後の頁 17006-17015
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/CVPR52729.2023.01631	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Yoshiyasu, Yusuke and Gamez, Lucas	４．巻 -
２．論文標題 Learning Body Shape and Pose from Dense Correspondences	５．発行年 2020年
３．雑誌名 Eurographics 2020 - Short Papers	６．最初と最後の頁 37-40
掲載論文のDOI（デジタルオブジェクト識別子） 10.2312/egs.20201012	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Yoshiyasu Yusuke、Samy Vincent、Imamura Yumeko、Ayusuwa Ko、Sagawa Ryuusuke、Yoshida Eiichi	４．巻 -
２．論文標題 Statistical Human Body Shape Model including Elderly People	５．発行年 2020年
３．雑誌名 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)	６．最初と最後の頁 4848-4853
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/EMBC44109.2020.9176459	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 Yusuke Yoshiyasu
２．発表標題 Deformable Mesh Transformer for 3D Human Mesh Recovery
３．学会等名 Computer Vision and Pattern Recognition (CVPR)（国際学会）
４．発表年 2023年

１．発表者名 Yusuke Yoshiyasu
２．発表標題 Learning Body Shape and Pose from Dense Correspondences
３．学会等名 Eurographics 2020 short paper（国際学会）
４．発表年 2020年

１．発表者名 Yusuke Yoshiyasu
２．発表標題 Statistical Human Body Shape Model including Elderly People
３．学会等名 EMBC 2020（国際学会）
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関