# 透視投影を考慮した単眼カメラからの全身・手関節３次元推定再考

山本 和彦 *

**概要.** 機械学習技術の発展により, 単眼 RGB カメラ画像からのみで人の全身, または手関節の高精度の３次元位置推定が可能となってきている. 一般的にこれらは全身/手のバウンディングボックス検出とその中での詳細な関節位置推定の２段階でおこなわれる. つまり, 画像のどこに対象が写っていようとも画像中心にあるものとして推定される. しかし, 撮像面に透視投影されることを考えると, カメラの光軸から離れるほど, つまり画像の端にいくほど, 奥行きの差は画像中心方向への変位へと変換される. そうしたバウンディングボックスを切り抜くと, 関節は実際より横倒しになって潰れた見た目の画像となっており, これを正面に対象が写っているものとして推定しても誤った結果になってしまう. これを解決するため, 我々は機械学習が解釈するバウンディングボックス内の関節が実際には透視投影されたより広い範囲の画像から切り取られたものだと再解釈することによって, 既存モデルであっても追加学習することなく精度を向上させられることを発見した. 本稿では, 既存モデルの再評価を通して従来の姿勢推定での深度評価, 学習方法における問題点を浮き彫りにする.

## 1 はじめに

単眼 RGB カメラ画像のみから人の全身骨格, または手の関節３次元位置を高精度に推定することは, ビジョンベースのユーザインタフェースにとって非常に重要である. こうした技術は, 現在機械学習技術の発展により, 関節 [1] だけでなく, 表面の密な推定 [2] も含めて性能は飛躍的に向上している. これらは, 全身の場合人, 手関節の場合片手ごとの領域認識をおこない, バウンディングボックスを切り抜いたのち, その内部において腰 (全身の場合), または手首 (手関節の場合) を基準とした相対的な３次元座標を推定する, という２段階の処理をおこなうのが一般的である. つまり, 従来モデルは画面のどこに対象が写っていようともそれが画面中心にあるものとして推定をおこなう. しかしながら, 透視投影ではレンズ光軸から離れれるほど, 奥行き方向の差は画像中心へ近づく方向へ変換されて投影されるため, 画面端にある対象は実際より横倒しになって潰れた見た目の画像となっている (図 1). これを画像中心にあるものとして推定しても当然正しい出力は得られず, 過小評価された深度が推定されてしまうという問題がある.

この問題を解決するため本稿では, 認識対象がレンズ光軸からどのくらい離れているかを考慮して機械学習モデルの出力の再評価をおこなう. これによって, 既存モデルであっても追加学習をすることなく再評価をするだけで, 報告されているよりも格段に高い性能を観測できることを実証する. また,

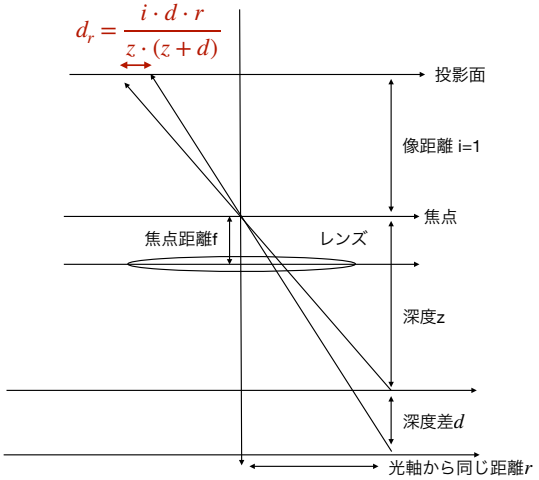


**図 1. レンズ光軸から離れるにしたがって深度差 $d$ は画像中心方向への変位 $d_r$ に変換されて投影される.**

このことから, 従来法における学習データセットの作成の仕方にも問題がある可能性があることを示す.

## 2 提案手法

ここでは手首を基準とした手関節相対位置推定を例として説明する. 全身推定の場合も基準が手首から腰となるだけで同様である. 我々は手の関節モデルとして MANO [3] に倣い片手 21 点を定義する. 図 2 のように光軸から離れた深度 $z$ の位置に手首位置 $P_w$, また, ある関節位置 $P_s$ があるとすると, 投影された関節位置 $p_s = [u_s, v_s]$ は画像上では, 投影された手首位置 $p_w$ を原点とした, レンズに垂直な仮想光軸 (図 2 赤点線) を中心として仮想焦点を挟んだ $P'_s$ にあるように見える. この結果画像上の手は横倒しになって潰れた見た目になってしまう. 機械学習モデルはこの画像から関節の相対深度 $d$ を推定



* ヤマハ株式会社

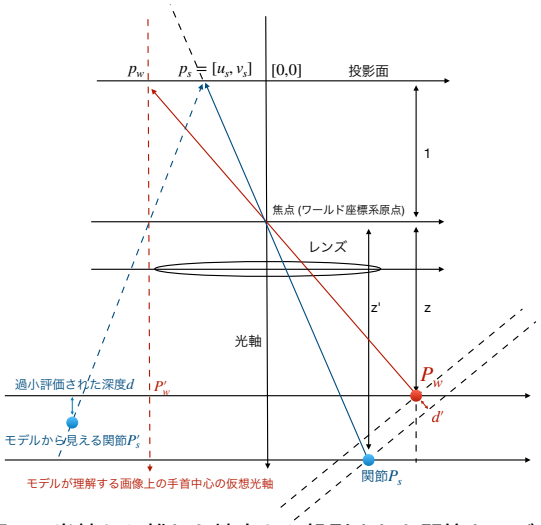


**図 2. 光軸から離れた地点から投影された関節をモデルは基準関節を通る仮想光軸に直交する平面上で評価してしまう.**

して出力する. 従来はこれをそのままレンズ光軸に沿った深度として利用していたが, これは過小評価されたものであるため, 実際には大きな誤差を生じさせる. そこで我々はレンズ焦点と, 基準となる手首座標を結んだ軸に沿った距離 (図 2 赤矢印) としてこの深度を再評価する. この再定義された関節の深度は,

$$z' = \frac{(|P_w| + d') \cdot |P_w|}{p_w \cdot [u_s, v_s, 1]^T} \tag{1}$$

と表すことができる. $d'$ は補正された深度で,

$$d' = (1 + ProjNet(|p_s|, d)) \cdot d \tag{2}$$

である. ここで $ProjNet() : \mathbb{R}^{21+20} \rightarrow \mathbb{R}^{20}$ は片手 21 点の正規化された関節画像座標の中心からの距離とオリジナルの機械学習モデルが出力した各関節の相対深度を入力として, 過小評価された相対深度を補正するスケールを出力するニューラルネットワークである. 本稿では全結合層と ReLU 活性関数のみ 3 層, 各隠れ層 128 次元のシンプル構造を採用した.

## 3 評価

まず, $ProjNet()$ を InterHand2.6M [4] の公開モデルの出力と学習データセットで事前学習したうえで, 検証データセットで提案手法適用有り無しの比較をおこなった. 結果, すべての入力画像において劇的に精度が改善し, MPJPE (関節間誤差の平均値) は適用無し 13.34mm に対して適用有りで 8.01mm と劇的に減少した. 図 3 に結果画像の一例を示す. オリジナルの出力と比較して自然な手の形状が彫られていることがわかる. また, 手の分布する深度の範囲も過小評価されていたものが修正されていることがわかる. これによって, 既存モデルにおいては

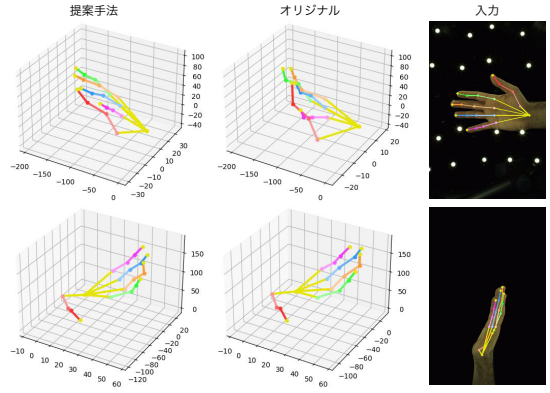


**図 3. 提案手法では過小評価されていた各関節の深度が修正されて自然な手の形になっている.**

深度評価の仕方に問題があり, 大きな誤差を生み出す要因となっていたことがわかる.

次に, 本稿での仮定を検証するため, 上述の InterHand2.6M で事前学習した $ProjNet()$ をそのまま MediaPipe [1] の出力に対して適用した. InterHand2.6M の検証データセットで実験した結果, 適用無しの MPJPE で 35.24mm, 適用有りで 28.74mm とこちらも誤差が大幅に減少した. これによって既存モデルが光軸から離れた対象の深度に対して同様のエラーモードを有しており, 同一の修正モデルを適用可能ということが確認できた.

## 4 まとめ

透視投影のメカニズムを考慮して画面に投影される対象の画像を再評価することで, 従来の人間の全身/手の 3 次元姿勢認識の評価方法の問題を 単純な方法で修正できることを示した. 本稿での知見から, 学習時においても正解の深度データをそのまま利用するのではなく透視投影を考慮したデータ拡張をおこなって学習させると, 従来モデルをより高性能にすることができると期待できる.

## 参考文献

[1] MediaPipe: A Framework for Perceiving and Processing Reality, C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. G.Yong, J. Lee, W. Chang, W. Hua, M. Georg, M. Grundmann, *CVPR*, 2019.

[2] DensePose: Dense Human Pose Estimation In The Wild, R. Güler, N. Neverova, I. Kokkinos, *CVPR*, 2018.

[3] Embodied Hands: Modeling and Capturing Hands and Bodies Together, R, Javier, T. Dimitrios, B. Michael J., *SIGGRAPH Asia*, 2017.

[4] InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image, M. Gyeongsik, Y. Shoou-I, W. He, Sh. Takaaki, L. K. Mu, *ECCV*, 2020.