人間の３次元運動推定に用いる画像ポーズ推定器のデータセットと学習効果

○大桶夏津（東大）　池上洋介（東大）　山本江（東大）　中村仁彦（東大）

1.　はじめに

人間の動作を解析するうえで, 関節位置の時系列情報をはじめとした運動情報は重要な要素となる. 例として, Marianna [1] らは自転車を漕ぐ動作における足の回転速度と骨盤の揺動の関係を解析し, 損傷リスクの低い回転速度を提案した. また, ゴルフのスイング動作から腕の加速度や関節角を計算し, プロの動作と自分の動作を比較することができるサービスも存在する [2]. このように, 人間の運動解析はスポーツや日常生活などの様々な場面で役に立つ.

運動情報を計測するモーションキャプチャ技術には主に光学式モーションキャプチャとビデオモーションキャプチャがある. 光学式モーションキャプチャは被験者にマーカを貼りつけて計測を行う手法であり, 高い精度を実現できるが, マーカが運動の妨げになることもある. 一方で, ビデオモーションキャプチャはカメラ画像のみから運動の三次元再構成を行う手法であり, マーカの貼り付けの必要がないため, スポーツなどの激しい運動の解析に適している. ビデオモーションキャプチャの例として Ohashi ら [3] や Chen ら [4] の研究が挙げられる. これらは, まず複数の 2 次元画像中の人の関節の位置を検出し, それらの三角測量などにより三次元空間での人の関節位置を推定する. よって, これらの手法において人の運動を正しく三次元再構成するためには, 画像中の関節位置を精度良く推定することが重要となる.

OpenPose [5] や HRNet [6] などの画像ポーズ推定器の学習には COCO keypoints Dataset [7] (COCO Dataset) が用いられる. COCO Dataset は 20 万枚以上の画像に 25 万人以上の人物を含む非常に大きなデータセットで, 屋内, 屋外など様々な場面の画像を含んでいる. 一方で, 体操の跳馬や格闘技の上段蹴りなど, 主にスポーツで見られる特殊な姿勢のデータは少なく, COCO Dataset のみで学習した画像ポーズ推定器ではこのような姿勢を誤推定してしまう場合もある.

そこで, 本稿ではこのような特殊な姿勢への対応を目標とした画像ポーズ推定器の学習について述べる. まず, LSPE Dataset [8] や CrowdPose [9] といったスポーツ画像を多く含むデータセットと, 独自に撮影したテコンドーや体操の画像を含むデータセットを用いて転移学習を行う. また, 腕のひねりなどの詳細なバイオメカニクス的な解析を行うことを想定し, これらのデータセットに対して新たに関節点のアノテーションの追加を行う. このデータセットを用いて学習した画像ポーズ推定器について, 評価用データに対しての推定結果を議論する. また, 複数のカメラ画像に対して推定された関節位置から運動の三次元再構成を行い, COCO Dataset [5] のみで学習した画像ポーズ推定器を用いて同様に三次元再構成した結果と比較する.

図 1: 学習に使用した各データセットのサンプル画像. (a) COCO Dataset [7], (b) LSPE Dataset [8], (c) CrowdPose [9] からそれぞれ抜き出したもの.

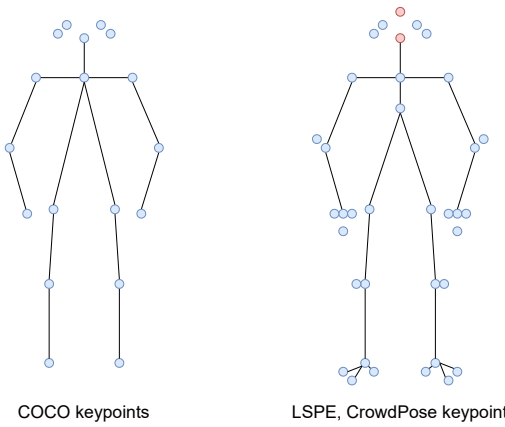


図 2: COCO Dataset [7] の 18 点のキーポイントと LSPE [8], CrowdPose [9] に新しく追加したものを含む 36 点のキーポイント. 独自データセットにつけた 34 点のキーポイントは 36 点のキーポイントから図中の赤い点で示される 2 点を除いたもの.

2.　スポーツ画像データセットを用いた画像ポーズ推定器の学習

2.1　データセット

画像ポーズ推定器の学習にあたり, 従来手法と同様の COCO Dataset [7] での学習の後, LSPE [8], CrowdPose [9] に新たにアノテーションを追加したものに加え, 独自に撮影した体操やテコンドーの画像を含むデータセットによる転移学習を行った. 各データセットの

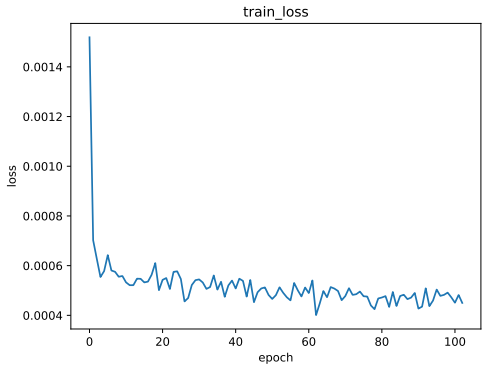


図 3: LSPE, CrowdPose, 独自データセットを用いた HRNet の学習における訓練データセットでの損失関数の推移. 横軸がエポック数, 縦軸が損失関数の値を表す.

画像の枚数, 人の数, アノテーションされたキーポイントの数, bounding box の有無を表 1 に示す. また, 各データセットの画像サンプルを図 1 に示す.

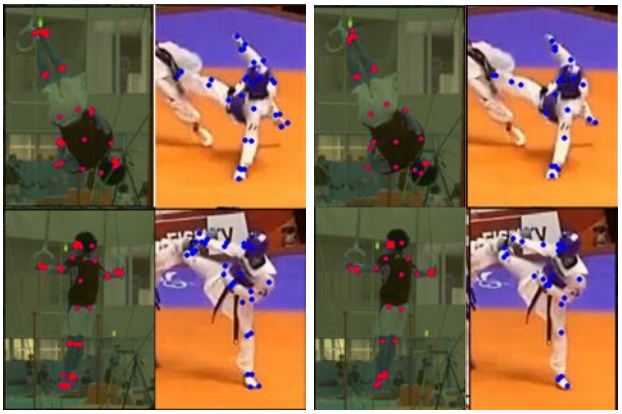
CrowdPose は 20,000 画像, 80,000 人にそれぞれ 14 点のアノテーションと bounding box が付いている. この中からスポーツ画像を中心に 5,000 画像を選択し, これらに 22 点のアノテーションを新たに追加してアノテーションの数を計 36 点とした. COCO Dataset [7] のアノテーションに含まれる 18 点のキーポイントと, 新しく追加したものを含む 36 点のキーポイントを図 2 に示す. これらの 5,000 画像の中から元の分と追加されたアノテーションが不正確なものを除き, 3,670 枚, 14,239 人を学習に使用した. 表 1 においてこれを CrowdPose_selected とする. LSPE は 10,000 画像, 10,000 人にそれぞれ 14 点のアノテーションが付いている. これらに同様に 22 点のアノテーションを追加し, アノテーションが不正確なものを除いた 9,701 枚, 9,701 人を学習に使用した. 表 1 においてこれを LSPE_selected とする. 独自のデータセットはテコンドーの試合のビデオ映像と独自に撮影した体操練習の映像からフレームをそれぞれ 724 枚と 777 枚切り出したもので, これらに 34 点のアノテーションと bounding box を付けた. この独自データセット計 1501 枚から 253 枚を評価用のデータセットとした.

2.2 学習条件

HRNet [6] の学習にあたり, まず COCO Dataset [7] のみで学習を行い, その後 LSPE [8], CrowdPose [9], 独自データセットを用いた転移学習を行った. 特徴抽出部の重みの初期値として Sun ら [6] が公開する HRNet の imagenet [10] での学習済みの重みを使用した. COCO Dataset, 転移学習時は共に Adam Optimizer[11] により学習率 0.001 で, COCO Dataset は 200, 転移学習は 100epoch の学習を行った.

2.3 画像ポーズ推定器の学習結果

まず, LSPE, CrowdPose, 独自データセットを用いた HRNet の学習の損失関数の推移を図 3 に示す. 図 3 において, 横軸がエポック数, 縦軸が損失関数の値を表す.



(a) Ground Truth (b) 推定値

図 4: 学習済みの HRNet [6] による評価用データを対象とした出力とその ground truth のサンプル. 画像中の青い点と赤い点が推定された, またはアノテーションされた関節位置を表す. 点の色は視認性の都合で画像により変えている. また, 推定値において尤度が 0.5 以下のものは表示されていない.

学習済みの推定器で評価用データセットを対象とした出力とその ground truth のサンプルを図 4 に示す.

また, 評価用データセットに対しての平均精度は 75.9 であった. 平均精度の計算にあたり, 正しく推定されたか否かの判断には以下の式 (1) が成り立つものを正しく推定されたものとした.

$$\sum_{i=1}^{N_{\text{keypoints}}} \exp\left(\frac{|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i|^2}{s}\right) \leq \epsilon \qquad (1)$$

式 (1) において, 左辺は object keypoint similarity (OKS) を表し, $N_{\text{keypoints}}$ は人一人当たりに定義されている関節点の総数, $\boldsymbol{x_i}$ は推定された関節 i の位置 [pixel], $\hat{\boldsymbol{x}}_i$ は ground truth における関節 i の位置, s は人物のサイズを表す. ここで, 人物のサイズ s をアノテーションで定義されたバウンディングボックスの面積の半分とした. また, ϵ は $0 \leq \epsilon \leq 1$ の値をとる閾値を表す.

2.4 考察

図 3 より, 学習データの損失関数が下がり, 学習が正常に進行していることが確認できる. 図 4 より評価用データにおいても関節点位置が概ね正しく推定できていることがわかる. また, 評価用データでの平均精度が 75.9 という点からある程度の汎化性能もあると考えられる. データセットやそのキーポイント数と評価指標が異なるため参考値ではあるが, OpenPose[5] の COCO Dataset のテストデータにおける平均精度は 65.3 となっている. 一方で, 肘や膝の外側に追加した特徴点が肘や膝の中心の特徴点と同じ位置に推定されることも多かった. これは肘や膝の外側の特徴点のアノテーションの位置にばらつきが大きく, 学習が進みにくかったものと考えられる.

表 1: 学習に使用したデータセットの画像枚数と人一人あたりにつけられたアノテーションのキーポイント数.

名前	画像数	人数	キーポイント数	bounding box の有無
CrowdPose[9]	20,000	80,000	14	有
CrowdPose_selected	3,670	14,239	36	有
LSPE[8]	10,000	10,000	14	無
LSPE_selected	9,701	9,701	36	有
Original Dataset	1,501	2,225	34	有

3. VMocap による動作の三次元再構成

3.1 ポーズ推定結果の三次元再構成による評価

以下の 2 つの推定器によりそれぞれ推定された画像上のポーズについて VMocap [3] による三次元再構成実験を行った.

- 推定器 1: スポーツ画像を含むデータセットで学習した HRNet [6]
- 推定器 2: COCO Dataset で学習した OpenPose [5]

OpenPose については Cao ら [5] が公開する学習済みモデルを使用した. VMocap では複数のカメラ画像を画像ポーズ推定器に入力し, 出力として得られた画像中の関節位置をもとに動作の三次元再構成を行うが, この画像ポーズ推定器として推定器 1 と 2 をそれぞれ使用し, 動作の三次元再構成を行った. 対象のデータとして, テコンドーの蹴り動作を撮影したものを用いた. 撮影の際に検証のため同時に光学式モーションキャプチャによる計測を行った.

3.2 三次元再構成結果

前節で定義した推定器 1 と 2 を用いて VMocap [3] により計算した人体の各関節の三次元位置の時間ごとの推定誤差を図 5 に示す. 図 5 において, 横軸が時間 [s], 縦軸が推定誤差 [mm] を表し, 青線がスポーツ画像を含むデータセットで学習した HRNet [6], 橙線が従来の OpenPose に対応する. 縦軸の誤差はいずれも光学式モーションキャプチャにより計測されたマーカ位置から計算した関節位置との距離の平均を表す. 関節位置の計算の例として, 肘については被験者の肘の外側と内側に貼り付けたマーカ位置の中点をとった. また, グラフ中の点線は被験者の蹴り動作の開始, 足が高く上がった時, 終了のおおよその時刻に対応し, それらの時刻で再構成された姿勢をそれぞれ示す.

同じ 2 つの推定器について, VMocap により計算された三次元空間での右足の関節位置の確信度を表すスコアを図 6 に示す. 図 6 の縦軸のスコア s_{all} は以下のように計算される.

$$s_{all} = \sum_{i=1}^{n_{camera}} \sum_{j=1}^{n_{joints}} s_{i,j} \tag{2}$$

ただし, n_{camera} は撮影したカメラの台数, n_{joints} は関節数を表し, $s_{i,j}$ は三次元空間上での関節 i の位置をカメラ画像 j に投影した位置における画像ポーズ推定器

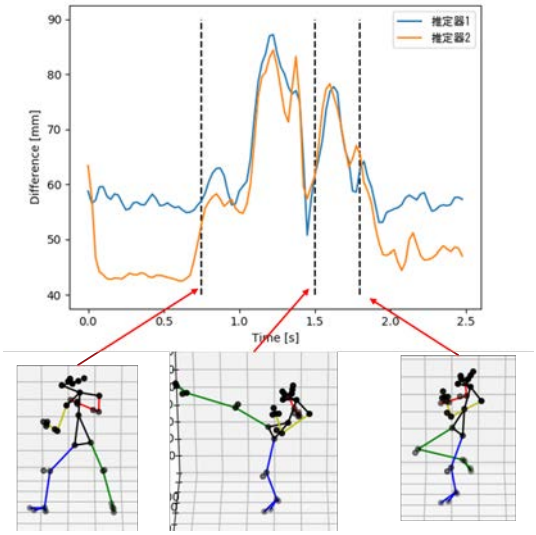


図 5: スポーツ画像を含むデータセットで学習した HRNet [6] (推定器 1) と従来のデータセットで学習した OpenPose [5] (推定器 2) を用いて再構成した三次元空間上の関節位置の時間ごとの誤差. 2 つの推定器の出力に共通する 20 関節について光学式モーションキャプチャで求めた関節位置からの誤差の平均をとった. グラフ中の点線は蹴り動作の開始時, 足が高く上がった時, 蹴り動作の終了のおおよその時刻に対応する. また, そのときに再構成された姿勢をそれぞれ示す.

が出力した関節の存在確率を表す. よって, スコア s_{all} が高いほど再構成された三次元空間上での関節位置の確信度が高いといえる. 今回の実験ではカメラ 4 台, 2 つの推定器で共通の右足の 6 点の関節についてスコアを計算したので $n_{camera} = 4$, $n_{joint} = 6$ となり理論上の最大値は 24 となる.

図 6 において, 横軸が時間, 縦軸がスコアを表し, 青線がスポーツ画像を含むデータセットで学習した HRNet [6], 橙線が従来の OpenPose に対応する. 図 6 中の点線も図 5 と同じ時刻にそれぞれ対応し, そのときの再構成された姿勢を同様に示す.

3.3 考察

図 5 より, スポーツ画像を含むデータセットで学習した HRNet [6] を用いて関節位置の三次元位置を再構成した場合に, 従来の推定器のほうが全体的に誤差は小さくなった. ただし, 足が高く上がっている時刻に関しては今回学習した推定器が従来とほぼ同じがやや上回る

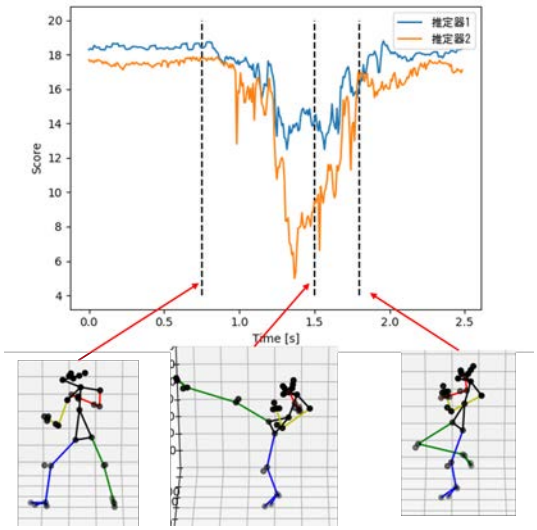


図 6: スポーツ画像を含むデータセットで学習した HRNet [6] (推定器 1) と従来のデータセットで学習した OpenPose [5] (推定器 2) を用いて再構成した三次元空間上の関節位置の時間ごとの確信度．確信度は式 (2) で計算され，この時の理論上の最大値は 24．グラフ中の点線は蹴り動作の開始時，足が高く上がった時，蹴り動作の終了のおおよその時刻に対応する．また，そのときに再構成された姿勢をそれぞれ示す．

こともあることがわかった．これについて，元のデータセットに多く含まれる立って静止しているような姿勢に対しては，転移学習により推定精度が下がってしまった可能性が考えられる．一方で，従来の OpenPose [5] では撮影したカメラ 4 台のうち 2 台において被験者の右足が上がった姿勢の時に正しく推定できていなかったのに対し，今回学習した HRNet [6] では 4 台のカメラいずれも右足が上がった姿勢を正しく推定できていた．これは図 6 中のおよそ 1.0 s から 1.7 s の区間で従来の推定器のスコアが大きく下がっているのに対し，今回学習した推定器はスコアの減少が比較的小さいことに表れている．

4. おわりに

本稿では，スポーツ画像を含むデータセットを用いて画像ポーズ推定器の学習を行った．

1. スポーツ画像を含む独自のデータセットで HRNet [6] の学習を行い，評価用のデータセットでの平均精度が 75.9 となった．データセットが異なるため参考値ではあるが COCO Dataset [7] での OpenPose [5] の平均精度は 65.3 となっているため，今回の HRNet の学習が正常に進行し，評価用データセットに対してもある程度妥当な出力ができているといえる．
2. 学習した HRNet と画像ポーズ推定の従来手法それぞれについて複数台のカメラ画像に対して関節位置を推定し，運動の三次元再構成を行った．三次元位置の誤差は全体的に従来手法の方が小さかったが，被験者が右足を高く上げた姿勢での誤差は従来手法とほぼ同じか最大で約 8.4 mm 小さいことがわかった．
3. VMocap は一部のカメラで画像ポーズ推定が失敗してもロバストに三次元再構成できるよう設計されているため，今回の実験では従来手法でも精度良く三次元再構成できた．ただし，被験者が右足を高く上げた姿勢では従来手法が 4 台のカメラのうち 2 台で正しく推定できなかったのに対し，今回学習した HRNet では 4 台とも正しく推定できていた．このことから，従来手法では正しく三次元再構成ができない他の動作で今回学習した HRNet が有効になる可能性があると考えられる．

謝　辞　本研究は，令和 2-4 年度科学研究費補助金基盤研究 (A)「パーソナル・デジタルツインの獲得・記述・認証」20H00599，および東京大学と株式会社 NTT ドコモとの令和 2 年度共同研究「ビデオモーションキャプチャと 5G 通信による動作収集システムの確立」の支援を受けた．順天堂大学 原田睦巳教授，全日本テコンドー協会，株式会社 Xenoma から学習や検証用の画像データについて協力を受けた．

参　考　文　献

[1] L. Marianna, L. Emahnuel Troisi, A. Valeria. ”Motion capture system: A useful tool to study cyclist ’ s posture.” Journal of Physical Education and Sport, 2020, 20.4: 2364-2367.
[2] ”MySwing Professional — Noitom Motion Capture Systems”, https://www.noitom.com/myswing
[3] T. Ohashi, Y. Ikegami, K. Yamamoto, W. Takano, Y. Nakamura. ”Video motion capture from the part confidence maps of multi-camera images by spatiotemporal filtering using the human skeletal model.” 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018.
[4] H. Chen, P. Guo, Pengfei L. Hee Lee, G. Chirikjian. ”Multi-person 3d pose estimation in crowded scenes based on multi-view geometry.” European Conference on Computer Vision. Springer, Cham, 2020.
[5] Z. Cao, G. Hidalgo, T. Simon, S. Wei, Yaser Sheikh. ”OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields.” IEEE transactions on pattern analysis and machine intelligence 43.1 (2019): 172-186.
[6] K. Sun, B. Xiao, D. Liu, J. Wang. ”Deep high-resolution representation learning for human pose estimation.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[7] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. Lawrence Zitnick. ”Microsoft coco: Common objects in context.” European conference on computer vision. Springer, Cham, 2014.
[8] J. Sam, and M. Everingham. ”Learning effective human pose estimation from inaccurate annotation.” CVPR 2011. IEEE, 2011.
[9] J. Li, C. Wang, H. Zhu, Y. Mao, H. Fang, C. Lu. ”Crowdpose: Efficient crowded scenes pose estimation and a new benchmark.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[10] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei. ”Imagenet: A large-scale hierarchical image database.” 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.
[11] K. Diederik and J. Ba. ”Adam: A method for stochastic optimization.” arXiv preprint arXiv:1412.6980 (2014).