Vol.2022-AVM-116 No.14
2022/2/25

視認率を考慮した人物特徴点の 3 次元座標推定精度の改善に対する検討

松村誠明 [1]　秋田健太 [2]　森本有紀 [2]　鶴野玲治 [2]　山本奏 [3]　青野裕司 [1]

概要：複数台の同期カメラで撮影された各映像に姿勢推定技術を適用し，取得した人物特徴点の 2 次元座標を用いて 3 次元再構成するとき，特に遠方に映る小さな人物や人物同士が交差するなど遮蔽が生じた際には姿勢推定技術の推定誤差が激しくなるため，3 次元再構成時の推定精度が低下する．本研究ではこれらの問題に対し，3 次元再構成時の最適化関数に視認率に応じた重みを付与することで推定精度の改善を図ったので報告する．

A Study for Improvement of 3D Pose Estimation Considering Visible Ratio of Human Feature Points

MASAAKI MATSUMURA[†1]　KENTA AKITA[†2]　YUKI MORIMOTO[†2]
REIJI TSURUNO[†2]　SUSUMU YAMAMOTO[†3]　YUSHI AONO[†1]

1. はじめに

近年機械学習の技術発展に伴い，画像処理技術を応用した実用サービスの検討が盛んに行われており，中でも人物の姿勢推定技術を用いたサービスは実用段階になりつつある．これら中にはスポーツにおける運動解析や採点等のほか，行動認識などが提供されているが，2 次元姿勢推定技術を用いた検討が中心である．2 次元姿勢推定技術で得られる情報は画面上における人物の特徴点(目・耳・鼻・関節等)の 2 次元座標のみであるが，人間は 3 次元的な動作を行うため，導入コストは増加するものの 3 次元姿勢推定を用いることで更なるサービス品質の向上が期待される．

2. 関連技術

2 次元姿勢推定技術は静止画像を入力とした検討が中心に進められており，特に画面内に複数名が同時に撮影された画像における被写体の 2 次元姿勢を一度に取得する方法として，広く用いられている技術が Cao らの提案する OpenPose[1]である．OpenPose は画面内における特徴点のヒートマップ(PCM: Part Confidence Map)と関節間の接続図(PAF: Part Affinity Fields)を VGG にて学習させることで高速かつ良好な推定精度を実現している．また，Papandreou らの提案する PersonLab[2]では PCM と PAF(Mid-range offsets)に加えて特徴点座標のハフ投票に用いる Short-range offsets，被写体を個体毎に分離するための mask と Long-range offsets を ResNet にて学習させることで高い推定精度を実現している．Cheng らは Wang らが提案した HRNet[3]を改良した HigherHRNet[4]を用いることで，従来手法では大小異なるサイズの被写体が画面内に混在する際に十分な認識率を維持できなかった問題に対し，多重解像度のネットワークを相互に接続することで解決する手法を提案している．

取得した 2 次元姿勢を 3 次元再構成にて 3 次元姿勢を推定する手法としては，複数台の同期カメラで撮影された各画像に 2 次元姿勢推定技術を適用し，取得した 2 次元姿勢に対してカメラ間で対応関係を構築した後，以下の式の最小化する Bundle Adjustment(以下 BA)によって特徴点の 3 次元座標を取得する試みが行われている．

$$\min \sum_{i,j} \left\| \boldsymbol{u}_{i,j} - F\left(\boldsymbol{p}_i | \boldsymbol{f}_j, \boldsymbol{r}_j, \boldsymbol{d}_j, \boldsymbol{R}_j, \boldsymbol{t}_j\right) \right\|_2^2 \qquad (1)$$

ここで i は特徴点番号，j はカメラ番号，$\boldsymbol{u}$ は 2 次元姿勢推定によって得られた特徴点の推定座標，$\boldsymbol{p}$ は特徴点の 3 次元座標，$\boldsymbol{f}$ はカメラ焦点距離，$\boldsymbol{r}$ はカメラ解像度，$\boldsymbol{d}$ はカメラ歪み，$\boldsymbol{R}$ はカメラ回転，$\boldsymbol{t}$ はカメラの 3 次元座標をそれぞれ表し，関数 F は特徴点の 3 次元座標 $\boldsymbol{p}$ をカメラのスクリーンに投影して 2 次元座標を得る関数を表す．BA では $\boldsymbol{u}$ のみを既知の値としてその他パラメータを未知数として導出させることもできるし，あらかじめカメラパラメータを別途導出しておき，$\boldsymbol{p}$ のみを未知数として導出させることもできる．しかしながら 2 次元姿勢推定技術は特に特徴点が遮蔽されたり明確に視認できないような状況や人物同士が近づいたり交差するようなシーンでは大きな誤差が重畳する．そのため，BA によって信頼のできる推定値を得るためには十分なカメラ台数を用意するなどして，多くのサンプリング点を確保する必要がある．

BA を用いて 3 次元姿勢を推定する技術として，Takahashi ら[5]は被写体の骨格長が不変であるという特徴を考慮し，

1　日本電信電話株式会社 人間情報研究所
2　九州大学
3　日本電信電話株式会社 スマートデータサイエンスセンタ

対象の被写体を 1 名に限定した環境にてカメラパラメータと特徴点の 3 次元座標を同時に推定する手法を提案している．また，Ohashi ら[6]は PCM の重ね合わせによって当該特徴点の 3 次元座標を推定する際に，既に 3 次元姿勢を生成した過去フレームの姿勢を参照し，次フレームにおける各被写体の予測姿勢を生成することで，当該特徴点が他者の同特徴点で遮蔽されるか否かによって影響の有無を切り替える手法を提案し，複数名が混在するシーンにおいてもロバストな 3 次元姿勢の取得を実現している．

3. 課題とアプローチ

前章にて記載した通り，人物特徴点の 2 次元座標を用いて 3 次元再構成で被写体の姿勢を推定するとき，2 次元姿勢推定技術に重畳する誤差に大きく影響を受ける．例えば図 1 のように 2 次元座標の推定誤差が小さいカメラと大きいカメラが混在している場合，従来の BA では推定誤差が大きい方に若干引き摺られるため結果的に特徴点の 3 次元座標の推定精度が低下することが分かる．

この問題に対し，Ohashi ら[6]は PCM の重ね合わせ時において，予測姿勢を用いた他者の同特徴点による遮蔽有無を考慮することで推定誤差の低下を軽減している．しかし 2 次元姿勢推定技術の誤差は他者の同特徴点による遮蔽だけでなく，他の部位に遮蔽されたり自己遮蔽によって精度が低下することも考えられる．2 次元姿勢推定技術はロバスト性が高く，多少の遮蔽に対しては確からしい 2 次元座標を推定する特徴を持つが，明確に視認できるカメラの推定精度と比較すると十分な推定精度とは言えないため，これらを同列に扱うことは避けた方が良い．

また，図 2 のように画面内に被写体が大きく映っている場合における特徴点の 2 次元姿勢推定誤差と，小さく映っている場合における特徴点の 2 次元姿勢推定誤差が同程度だった場合，3 次元化した際には前者の誤差よりも後者の誤差の方が大きな誤差を生みやすい．

そこで我々は BA の最適化計算途中のカメラパラメータや特徴点の 3 次元座標を用いて都度被写体のシルエットを生成し，シルエットに基づく特徴点毎の視認率 w を BA の計算式に組込むことで特徴点の 3 次元座標推定精度の向上を図る．

$$\min_{\{\boldsymbol{p}_i\},\{(\boldsymbol{R}_j,\ \boldsymbol{t}_j)\}} \sum_{i,j} w_{i,j} \left\| \boldsymbol{u}_{i,j} - F\left(\boldsymbol{p}_i \middle| \boldsymbol{f}_j, \boldsymbol{r}_j, \boldsymbol{d}_j, \boldsymbol{R}_j, \boldsymbol{t}_j\right) \right\|_2^2 \tag{2}$$

4. 提案手法

BA の最適化計算途中にて各被写体のシルエットを生成し，このシルエットを用いて各関節の視認率を計算することで 2 次元姿勢推定技術の推定誤差が大きいカメラの影響を軽減させるため，本章ではシルエットの生成方法と，それらを用いた重みの計算方法を述べる．

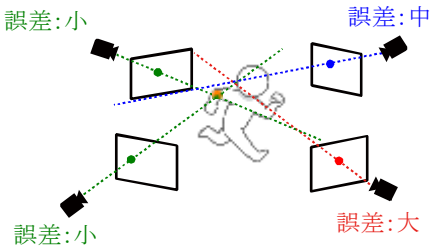

図 1　2 次元座標の推定誤差が
3 次元座標の推定精度に与える影響

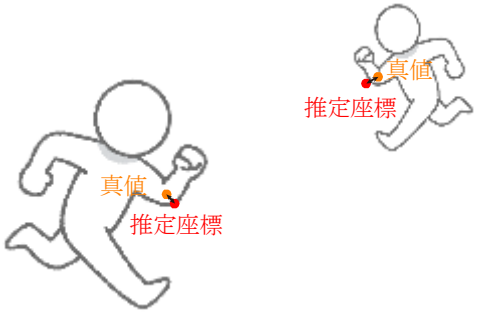

図 2　被写体のサイズと推定誤差の影響

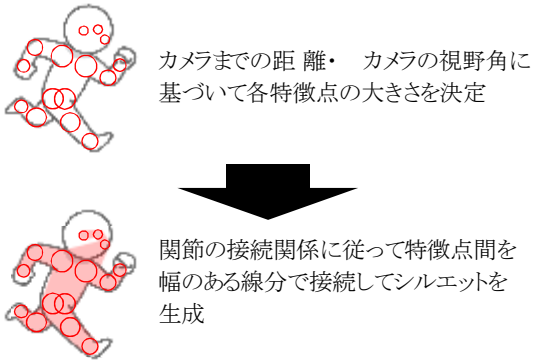

図 3　シルエットの生成

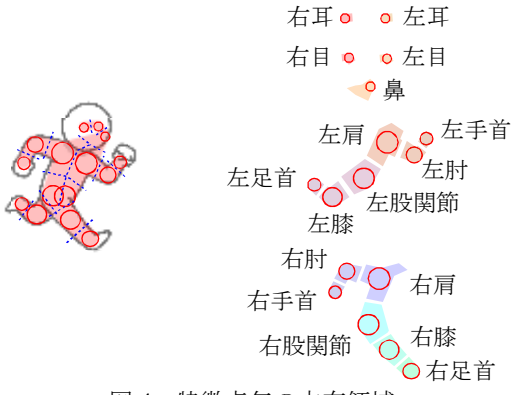

図 4　特徴点毎の占有領域

4.1 シルエットの生成

シルエットは BA の最適化計算途中における被写体の各特徴点の三次元座標を各カメラに投影し，カメラまでの距離・視野角に基づいてその大きさを決定し，関節の接続関係に従って特徴点間を幅のある線分で接続することでカメラから見た際の簡易的なシルエット生成を行う(図 3). なお，シルエットにおける各特徴点の占有領域は接続先の特徴点までの線分における中点で分断した形状とする(図 4).

Vol.2022-AVM-116 No.14
2022/2/25

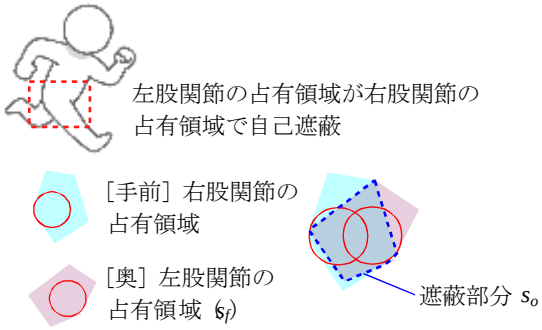

図 5　遮蔽部分の検出

4.2　画面内占有率に応じた視認率(占有領域視認率)

前章に記載の通りカメラの視野内に映る当該特徴点の大きさに応じて重みが変化するよう設計することが望ましい．そこで，当該カメラの総画素数を S，当該特徴点の占有領域を s_f として占有領域視認率を考慮した重み w_a を以下のように定義する．

$$w_a = \frac{s_f}{S} \tag{2}$$

4.3　自己/他者遮蔽を考慮した視認率(遮蔽時視認率)

当該特徴点が明確に視認できる際は大きな重みを割当て，他者や自身の体によって遮蔽され視認しにくい際は小さな重みを割り当てることが望ましい．そこで，全ての特徴点の占有領域を図 5 のように奥行が遠い順にスクリーンに投影し，当該特徴点の遮蔽部分 s_o を検出することで，遮蔽時視認率を考慮した重みを以下のように定義する(図 5 は自己遮蔽の例だが，他者遮蔽の場合も同様に処理する)．

$$w_o = \frac{s_f - s_o}{s_f} \tag{3}$$

5.　実験・結果

5.1　試験シーケンス

実写映像では関節座標の正解値を得ることが困難なため，本稿では慣性センサ式モーションキャプチャスーツである MVN[7]にて取得したスポーツや日常生活の動作を含むモーションデータ計 131 シーケンスの中からランダムに 4 つのモーションデータを選択し，撮影空間内に 4 人の人物が同時に収まるようランダム配置したモーションデータに対して Poser Pro 11[8]にて人体モデルを当てはめてレンダリングしたシーケンス 100 個を用いた．なお，各シーケンスは計 150 フレーム (30 fps)のデータとして生成した．

5.2　撮影空間設定

撮影空間は図 6 のように中心を向いて周回する CG の空間上に計 12 台の FullHD カメラ(30°間隔)を設置し，偶数番のカメラは水平視野角 36.83°を，奇数番の水平視野角を 50.93°とした(各カメラの設定は表 1 参照)．これらのカメラにてレンダリングしたシーケンスの例を図 7 に示す．

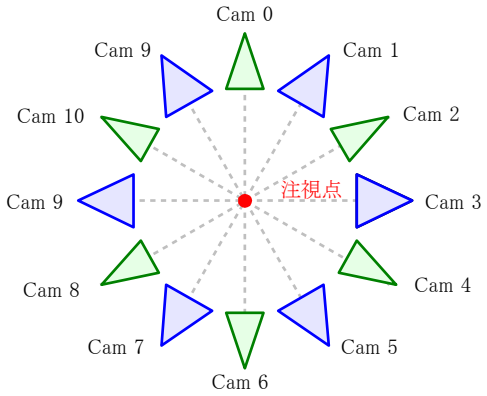

図 6　カメラ配置

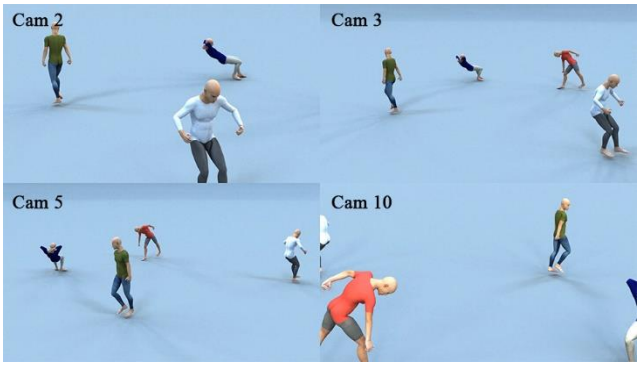

図 7　レンダリングイメージの例

表 1　カメラ設定

Cam ID	水平視野角 [°]	視点(x, y, z) [m]	注視点(x, y, z) [m]
0	36.83	10.000, 3.000, 0.000	0.000, 0.500, 0.000
1	50.93	8.660, 3.000, 5.000	
2	36.83	5.000, 3.000, 8.660	
3	50.93	0.000, 3.000, 10.000	
4	36.83	-5.000, 3.000, 8.660	
5	50.93	-8.660, 3.000, 5.000	
6	36.83	-10.000, 3.000, 0.000	
7	50.93	-8.660, 3.000, -5.000	
8	36.83	-5.000, 3.000, -8.660	
9	50.93	0.000, 3.000, -10.000	
10	36.83	5.000, 3.000, -8.660	
11	50.93	8.660, 3.000, -5.000	

5.3　実験条件

レンダリングしたシーケンスの各フレームに対して現状最も広く利用されている 2 次元姿勢推定技術である OpenPose[1]を適用して得た姿勢データに対し，類似度が高い特徴点を持つ姿勢同士を検出することによりカメラ間で被写体の紐付けを行ったデータに対して，カメラパラメータ固定した特徴点の 3 次元座標のみを計算する BA を実行する．BA の最適化には Ceres Solver[9]を使用し，出力した結果と，正解となるモーションデータにおける各特徴点の 3 次元座標との距離にて評価を行う．なお，正解となるモーションデータにおける関節定義と 2 次元姿勢推定技術に

て出力される特徴点の定義とでは一致しない部分があるため，本評価においては双方のデータに含まれる左右両方の肩・肘・手首・股関節・膝・足首の計 12 関節の平均誤差で評価を行った．

5.4 実験結果

生成した 100 個のシーケンスに対して従来手法(純粋に Ceres Solver にて BA を行った結果)と提案手法の結果を図 8 に，従来手法との差を図 9 に記す．

誤差全体に対しての改善量は微小なため，図 8 では大きな差が確認できないが，図 9 にて差分を見ると傾向が把握できる．占有領域視認率 w_a のみを考慮した際にはシーケンス依存性が高く，効果的なシーケンスにおいては最大 12.99mm 程度の改善が確認されたが，悪化するシーケンスも多くシーケンス平均では 0.74mm (1.74%)の改善に留まった．遮蔽時視認率 w_o を考慮した際には悪化するシーケンスは少なく平均的に良好な結果が得られ，全てのシーケンス平均では 1.62mm (3.79%)の改善が確認された．2 つの重みを組合せた w_aw_o で最適化した結果は平均 2.13mm (4.99%)の改善という最も高い値が確認できた．

なお，本手法の導入に伴い，BA 最適化計算のイテレーション回数は従来手法が平均 1.9 回 だったのに対して平均 25 回程度まで増加した．また，従来手法の計算時間が平均 9msec だったのに対し，提案手法では平均 1700msec 程度にまで増加した．そのため，提案手法では全体の 93%程度がシルエット生成・重みの算出に使われていることが分かった．

6. 結論・今後の課題

本稿では，複数台の同期カメラで撮影された各映像に姿勢推定技術を適用し，取得した人物特徴点の 2 次元座標を用いて 3 次元再構成するときの BA 最適化計算において，各特徴点の占有領域視認率と遮蔽時視認率の 2 つを重みとして付与することで推定精度の改善を図った．この結果，従来手法比で最大で平均 4.99%の改善を確認した．

今後の課題として，従来手法比で誤差が増加するフレームに対する考察を深め，最適化計算内で誤差が増加するケースに該当しそうな場合は占有領域視認率と遮蔽時視認率の重みを調整するなどすることで，さらなる改善が期待できる．また，速度の面においては，本実装は試験的な取り組みで行ったため，最適化の各イテレーションにて CPU にてシルエット描画を実施していたため，膨大な計算時間が必要だった．しかし，シルエット描画は極めて単純描画方法で良いため，GPU 等に計算させることで，大幅な速度改善が期待できる．

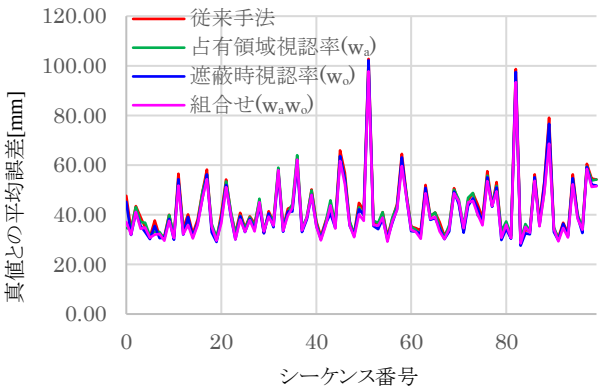

図 8　実験結果

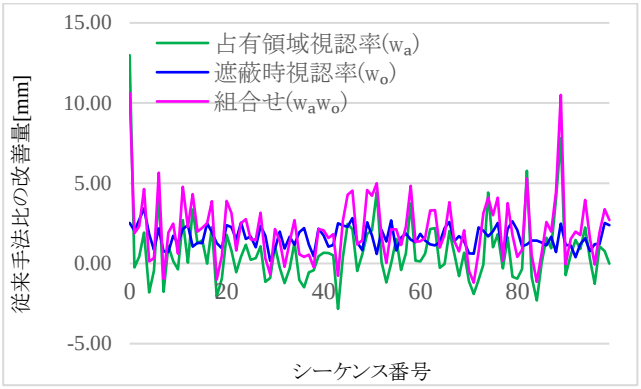

図 9　従来手法との比較

参考文献

[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[2] G. Papandreou, T. Zhu, L. C. Chen, S. Gidaris, J. Tompson, and K. Murphy, “PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model”, ECCV, 2018.

[3] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep high-resolution representation learning for visual recognition. CoRR, abs/1908.07919, 2019.

[4] Sun, K.; Xiao, B.; Liu, D.; Wang, J. D. Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5686–5696, 2019.

[5] K. Takahashi, D. Mikami, M Isogawa, and H. Kimata, “Human Pose as Calibration Pattern; 3D Human Pose Estimation with Multiple Unsynchronized and Uncalibrated Cameras”, CVPR, pp.1888-1895, 2018.

[6] T. Ohashi, Y. Ikegami, and Y. Nakamura. Synergetic reconstruction from 2d pose and 3d motion for wide-space multi-person video motion capture in the wild. Image and Vision Computing, 104:104028, 2020.

[7] “MVN Animate”. https://www.xsens.com/products/mvn-animate

[8] “Poser Pro”. https://www.e-frontier.com/smithmicro/poser/poser-pro.html

[9] “Ceres Solver”. http://ceres-solver.org/