機械学習とのハイブリッド逆投影法による 鍵盤楽器演奏時の高精度マルチカメラハンドトラッキング

山本 和彦 †

† ヤマハ株式会社

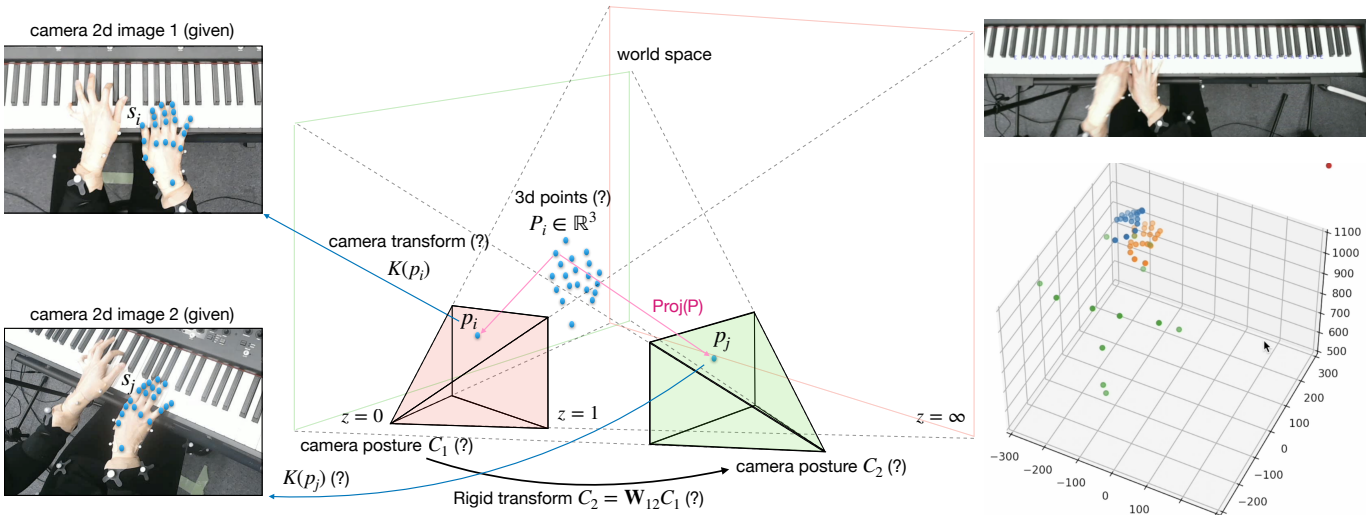

図 1　鍵盤楽器演奏時の複数視点のカメラ画像と音が与えられたとき, 両手の手関節の 3 次元位置とカメラ姿勢, 鍵盤平面を非接触で同時推定する (左). これは逆投影問題として定式化することによって効率的に解くことができる. 複数視点カメラ画像からのバンドル調整と機械学習とのハイブリッドな手法によって, 従来手法では困難な自己遮蔽の大きい状況でも正確に手の 3 次元形状を絶対スケールとともにキャプチャすることができる (右).

概要

モーションキャプチャにおいて, 外部物体との複雑なインタラクションを伴う手指動作の詳細なトラッキングをすることは最も難易度の高いタスクである. とくに, 楽器演奏時には, 手にグローブやセンサ類を装着する従来手法では自然な動きを阻害してしまう, 楽器や手そのものによる遮蔽が発生してしまう, という致命的な問題がある. そこで本論では, 複数の RGB カメラと音の情報のみを使い, 手に一切のセンサ類を装着することなく鍵盤演奏時の両手の詳細な 3 次元モーションキャプチャをおこなう手法を提案する. これを実現するため, バンドル調整で一般的な再投影誤差の最小化ではなく逆投影問題として定式化し, 機械学習とのハイブリッドな手法でこれを効率的に解く. 検証の結果, 演奏動作を妨げることのない高精度の非接触ハンドトラッキングが実現できたことを示す.

1　はじめに

楽器演奏時の手指動作の詳細なキャプチャをすることは高い需要があるにもかかわらず, 非常に難易度が高く, 挑戦的なタスクとなっている. 従来, 手にマーカーやグローブなどのセンサを装着する手法が一般的であるが, 手の自然な動きを阻害したり, 手の自己遮蔽によるセンサ認識ミスなどの致命的な問題が発生する. 手に一切のセンサ類を装着せず画像情報 (深度含む) からの動作認識 [1] も深層学習の発展に伴い性能が向上しているが, それぞれの手の関節と両手の絶対位置関係を同時に推定することが困難であったり, 手首などを基準とした相対位置しか計算できない, 一部が遮蔽された手関節の推定精度が著しく低下するといった問題がある.

そこで本論では鍵盤楽器演奏時の両手の手指の関節の詳細な動きを複数の RGB カメラと音情報を使って非接触でトラッキングすることのできる手法を提案する. 提案手法では複数視点のカメラ画像と音の情報から手関節の 3 次元位置とカメラ姿勢, 鍵盤平面を同時推定する. これを実現するために, カメラ画像から 2 次元の手関節位置推定をまずおこない, 世界座標系にある 3 次元手関節点からのレイがそれぞれのカメラへ到達し, 推定された 2 次元関節点へと投影されるプロセスを逆に辿ることでつじつまを合わせる, 逆投影問題として定式化をおこなう. このとき, 補助的に 2 次元手関節変位を入力として 3 次元化をおこなう事前学習済みの深層学習モデルを使う. これによって, 楽器や手自体によって遮蔽された関節においても安定して良好な 3 次元位置推定をおこなうことができる. また, 音の情報を利用することによって従来手法では決定することができなかった絶対スケールを定めることができ, かつ鍵盤表面の平面も求めることができるようになる.

提案手法の評価のため, 手と手を除いた全身に小型の光学式のマーカーを貼り付け, 提案手法と同時に鍵盤楽器演奏の収録をおこなって比較した. その結果, 完全非接触で演奏を妨げることのない高精度のハンドトラッキングが実現でき

たことを示す.

2 関連研究

複数視点のカメラ画像が与えられたときにカメラ姿勢とオブジェクト位置を同時推定する手法は, バンドル調整 [2] として知られており, SLAM (Simultaneous Localization and Mapping) や 3 次元再構成をおこなう COLMAP [3] などのための基礎技術となっている. 一般的なバンドル調整では, まず複数視点の画像が与えられ, それぞれの画像からの特徴点抽出をおこなう. 抽出された特徴点のうち, 異なる画像間で対応関係にある可能性が高いものを選出し. それらが同じ 3 次元位置から投影されたものとみなして辻褄が合うように再投影誤差の最小化をおこなう. しかし, カメラ配置が非常に疎であり視点が大きく離れている場合には画像からの特徴点マッチングが機能しなくなる, 最適化が困難になる, という問題がある. 特に本論で対象としている楽器演奏時には前景となる手が激しく動き, かつ背景となる鍵盤は繰り返しを多く含んでいるため特徴点マッチングが機能しない. また, 楽器周辺に密にカメラを配置することは現実的ではない. このような状況に従来法では対応が困難である.

3 提案手法

3.1 逆投影問題

複数視点のカメラ画像において 2 次元の手関節位置 $f_i^j \in \mathbb{R}^2$ が事前学習済みの機械学習モデルによって得られたとする. これを特徴点として用いると, 関節それぞれが一意に定まるため, 画像からの特徴点抽出で問題となる画像間でのマッチングエラーは回避できる. 本論では片手につき 21 点の関節点を使う. ここから従来のバンドル調整では, 世界座標系での未知の 3 次元手関節点 $\mathbf{P}_i \in \mathbb{R}^3 = [x_i, y_i, z_i]^T$ をそれぞれのカメラへ投影 $\mathcal{P}()$ したときの画像上の点と各特徴点 f_i^j との再投影誤差を最小化することが一般的である.

$$\arg\min_{P,W,K} \sum_{i\in\mathbb{C}} \sum_{j\in\mathbb{F}_i} \| K\left(\mathcal{P}(\mathbf{W}_i P^j)\right) - f_i^j \|_2 \tag{1}$$

ここで, $\mathbb{C}$, $\mathbb{F}$, $\mathbf{W}_i$, $\mathbf{K}()$ はそれぞれ手を認識したカメラ集合, カメラ画像 i での特徴点集合. カメラ i に対する原点からの剛体変換, カメラ内部パラメータに依存する関数である. しかし, この最適化はカメラ視点が極端にスパースであったり, 特徴点が局所的 (本論の場合は手) である場合に解くことが困難となり, 実際に筆者らの環境で実験したところ正常な値へ収束することはなかった. この問題を解決するため本論では, 以下のように光線軌跡をカメラ側から世界座標系へ逆に辿る逆投影問題として定式化する.

$$\arg\min_{z,\mathbf{W},\theta^k} \sum_{\substack{i\neq j \\ i,j\in\mathbb{C}}} \sum_{k\in\mathbb{F}} \| \mathcal{P}^{-1}\left(K_{\theta^k}^{-1}(f_j^k), z_j^k\right) - \mathbf{W}_{ij}\mathcal{P}^{-1}\left(K_{\theta^k}^{-1}(f_i^k), z_i^k\right) \|_2 \tag{2}$$

ここで $\mathbf{W}_{ij} \in \mathbb{R}^{4\times4}$ はカメラ i から j への回転 $\mathbf{R}_{ij} \in \mathbb{R}^{3\times3}$ と並進 $\mathbf{T}_{ij} \in \mathbb{R}^3$ を表す行列,

$$\mathbf{W}_{ij} = \begin{bmatrix} \mathbf{R}_{ij} & \mathbf{T}_{ij} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \tag{3}$$

$\mathcal{P}^{-1}(u_i^k, v_i^k, z_i^k) = [u_i^k z_i^k,\ v_i^k z_i^k,\ z_i^k, 1]^T$ は逆投影関数である. また, 関節の深度 z_i^k については直接最適化するのではなく, $[z_i^1, ..., z_i^{21}] = z_i^{scale} \cdot LiftNet(f_i^1, ..., f_i^{21}) + z_i^{wrist}$ とモデル化する. $LiftNet()$ は片手すべての関節の 2 次元座標を入力とし, 手首を原点としたそれぞれの関節の深度を出力する事前学習済みニューラルネットワークである. これによって本来の $3 \times 2 \times 21$ 次元の関節推定問題は $2 \times 2 \times |\mathbb{C}|$ 次元の最適化問題にまで簡略化される. 我々はこの問題を Levenberg-Marquardt 法 (LM 法) [4] で解く. ここで, カメラ内部逆関数 $K_{\theta^k}^{-1}()$ についてはレンズ中心からの距離に対する多項式関数としてモデル化した.

3.2 拘束

深度の時間的な一貫性を保つために以下の拘束を導入して最小化する. t はフレーム番号である.

$$\alpha \cdot \| z_i^t - 0.5 \cdot (z_i^{t-1} + z_i^{t+1}) \|_2 \tag{4}$$

さらに, 関節の長さが大きく変わらないようにするために As-Rigid-As-Possible 拘束 [5] を導入する.

$$\beta \cdot \sum_{lk\in\mathbb{B}} \Big\| \left(\mathcal{P}^{-1}(K^{-1}(f_i^l)) - \mathcal{P}^{-1}(K^{-1}(f_i^k))\right)^2 - \left(\mathcal{P}^{-1}(K^{-1}(f_j^l)) - \mathcal{P}^{-1}(K^{-1}(f_j^k))\right)^2 \Big\|_2 \tag{5}$$

$\mathbb{B}$ はボーン集合である. また, 隣り合う指の関節は同じ方向にしか稼働できず, これは速度が大きいほど顕著である.

$$\gamma \cdot \sum_{lk\in\mathbb{B}^f} \left(\overline{(\vec{P}_i^{l,t+1} - \vec{P}_i^{l,t})} \cdot \overline{(\vec{P}_i^{k,t+1} - \vec{P}_i^{k,t})} - 1 \right)^2 \cdot \left| (\vec{P}_i^{l,t+1} - \vec{P}_i^{l,t}) \right| \left| (\vec{P}_i^{k,t+1} - \vec{P}_i^{k,t}) \right| \tag{6}$$

ここで上線はベクトルの正規化, $\mathbb{B}^f$ は同じ方向にしか動かないボーンの集合である. 最後に当然ではあるが, 提案手法ではカメラ位置がほぼ固定のため以下のようにカメラ姿勢変換を一周すると元の姿勢に帰還する.

$$\mathbf{R}^{\mathbf{T}}{}_{1C} \prod_{C-1\geq i,j=i+1\geq 1} \mathbf{R}_{ij} = \mathbf{I}, \qquad \mathbf{T}_{C1} + \sum_{i=1,j=i+1}^{C} \mathbf{T}_{ij} = \mathbf{0} \tag{7}$$

最後に, スケールを一意に決定するため, $|\mathbf{T}_{1C}| = 1$ とする.

3.3 ファインチューニング

上記の逆投影問題を LM 法で解き, カメラ姿勢とそれぞれのカメラからの手関節の深度を求めたのち, 最後にすべての手関節 3 次元座標の微修正をおこなう. それぞれの時間

フレームにおいて最も逆投影誤差の小さいカメラペア ij を選び, それぞれのカメラから逆投影した 3 次元関節位置を $P'_i \approx P_i + \delta_i,\ \delta_i \in \mathbb{R}^3$ とおいて以下の最小二乗問題を解く.

$$\arg\min_{\delta_i,\delta_j} \left\|P'_j - P'_i\right\|_2 + \alpha \cdot (|\delta_i| + |\delta_j|) + \beta \cdot \sum_{lk\in\mathbb{B}} \left\|(P_i^{'l} - P_i^{'k})^2 - (P_j^{'l} - P_j^{'k})^2\right\|_2 \quad (8)$$

α, β は重みである. このとき δ 以外のパラメータは固定する. また, δ の大きさは高々隣り合う関節の長さの半分を超えないようにバウンドする. 上式の第二項は正則化項, 第三項は As-Rigid-As-Possible 拘束である. 最終的に得られた P'_i が手関節の 3 次元座標となる.

3.4 絶対スケールと鍵盤平面の決定

RGB 画像のみによるバンドル調整には絶対スケールが不定となる問題がある. これを決定するために鍵盤楽器から出力される MIDI ノート (音符情報) $o \in [0, ..., 127]$ を利用する. MIDI の出力できない楽器の場合は音を録音し, MIDI ノート化する (ここでは多分にエラーを許容する). ここで前提として, カメラ画像上での鍵盤の位置は未知であり, また鍵盤はオクターブごとに繰り返されており, さらに楽器によって何オクターブの鍵盤があるのか不明であるため特徴点マッチングは困難である. 現在どのキーが押されているかは得られているが, どの指がどのキーを押さえているかも未知である.

まず, それぞれの鍵盤の数分だけカメラ画像と同じ解像度のグレースケール画像 $I_k \in \mathbb{R}^{H\times W}, k = [0, ..., 127]$ を用意し, これを鍵盤分布画像と呼ぶ. フレーム $t-1$ から t 間において新しく押された鍵盤の集合を $\mathbb{K}^t$ とし, t において手の平と直交する内側方向へ動いたすべての指先の位置のピクセル位置を平均 μ として適当な分散 σ をもつガウス分布 $\mathcal{N}(\mu, \sigma)$ を $I_k^t \leftarrow I_k^{t-1} + v \cdot a \cdot \mathcal{N}(\mu, \sigma), k \in \mathbb{K}^t$ のようにそれぞれの鍵盤分布画像に足す. ここで v, a は指先の速度と音の強さである. このとき $t \to \infty$ において, それぞれの鍵盤分布画像の極値の位置 p_k を求め, 隣り合う鍵盤の鍵盤分布画像の値を引いたもの $I_k(p_k) - I_{k-1}(p_k) - I_{k+1}(p_k)$ が正の値となる視点が 2 つ以上存在しているとき, この点を鍵盤 k の表面特徴点候補とする. 1 分間鍵盤を適当に演奏した後に得られたこの候補点を鍵盤番号と色相を対応させて可視化したのが図 2 上段である. 繰り返しの影響の無い特徴点が得られていることがわかる.

最終的に残った表面特徴点候補群から, 鍵盤表面の 3 次元位置を前段階までで得られたカメラ姿勢とパラメータを用いて逆投影問題を最小化して求める. このとき押下された鍵盤位置の集合 (表面特徴点) は平面となることを拘束条件に含める. 最終的な鍵盤表面位置は有効な特徴点マッチングを RANSAC [6] で選択することによって決定する.

こうして求まった鍵盤平面に対して各特徴点を射影し, 主成分分析をおこなう. この主軸にさらに各点を射影したとき, 最も距離の大きい 2 点間の距離 D に対応している音程 H から, ピアノ鍵盤の幅の規格が 1 オクターブ 165mm であることを使い, このバンドル調整の絶対スケールが $165H/12D$ と算出できる. 図 2 下段は赤点がカメラ位置, 赤線がカメラ角度, 青線が推定された鍵盤平面の主軸と副軸, 緑点が最終的に選択された鍵盤表面特徴点を示す. なおここでは鍵盤真上のカメラ位置を原点としている.

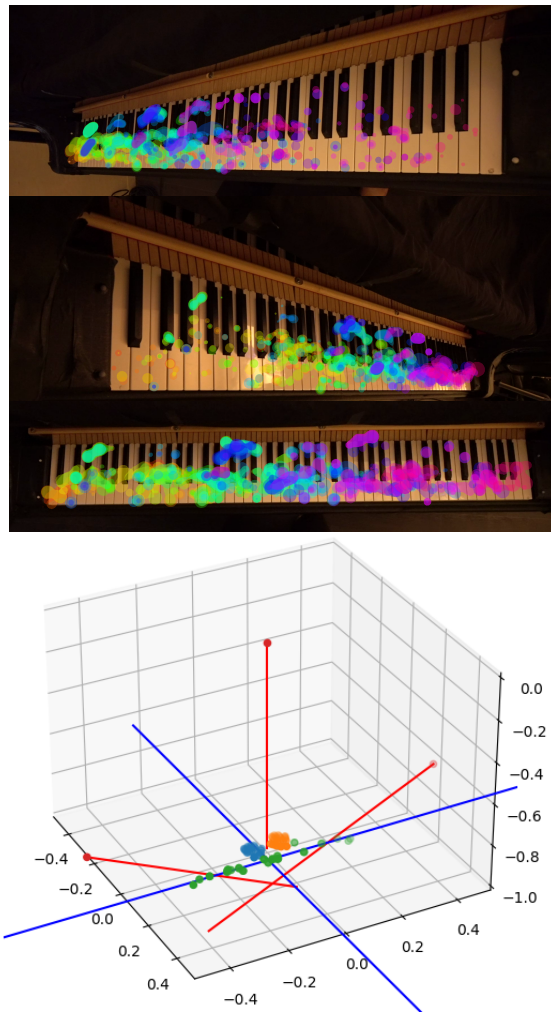

図 2　上段: 1 分間演奏後の鍵盤特徴量の収束結果. 音高が色相に対応している. 下段: 最適化された鍵盤表面特徴点 (緑点) とその主軸と副軸 (青線). 赤点はカメラ位置, 赤線はカメラ角度, 青点とオレンジ点は手を示す.

図 3　実験のセットアップ. 手と全身に光学マーカーを貼り付けて提案手法と同時収録をおこなった. カメラは鍵盤真上と左右の 3 台を設置した.

4　実験

4.1　実装

本システムは, マルチカメラでの撮影システム, 撮影動画からの 2 次元手関節推定システム, 提案手法による 3 次元ハンドトラッキングシステムの 3 つから構成される. 画像からの 2 次元関節位置推定には MediaPipe [1], 3 次元化をおこなう $LiftNet()$ の構造には Graph U-Net [7] を採用し, 事前学習のためには複数視点の手関節データセット [8, 9] を使用した. 実装は C++ でおこない, 実行は Apple M1 チップ上最大 8 並列でおこなった. また, カメラには Sony RX01 II (フル HD, 60fps) を 3 台. ピアノには YAMAHA C3 を使用した.

4.2　比較

比較のため, 手の提案手法と同じ関節位置, および手以外の全身に光学式のマーカーを貼り付け, OptiTrack (https://www.optitrack.jp) のシステムでトラッキングをおこない提案手法と同時収録した (図 3). 演奏時には手のマーカーの遮蔽が多分に発生し, 関節のいくつかが認識できないことが多発したため, 手動での修正をおこなった.

まず OptiTrack で得られた手関節の結果と提案手法で得られた関節出力のプロクラステス距離を比較した結果, ショパンのエチュードやスケールなどの 6 分間の演奏で関節位置の最大誤差 20mm 以下となった (図 4). 再投影誤差の最小化をはじめ, 従来手法はすべて最適化が失敗したため比較が困難となったが非接触での手関節認識のひとつのベースラインを得たと言える. この誤差の原因は単純な推定誤差だけではなく, 貼り付けたマーカー位置で 2 次元関節推定座標がそのまま得られるわけではないことにもある, 絶対位置/スケールの評価として, OptiTrack で測定された現実スケールのデータと重ね合わせたものを添付動画に示す. 動画では 2 手法の結果が高い精度で一致していることがわかる. このとき左右の手が最も離れたときの手首位置の距離の誤差が約 30mm となった. 深度情報を一切利用しないバンドル調整でこの精度を達成することは従来法では困難である. 最後に, 逆投影問題の最適化時間は 1 フレームあたり約 0.2 秒となった. 計算が正常な値へ収束しなかったが, 再投影誤差の最小化問題として実行した場合は 1 フレームあたり 3 秒程度かかったことと比較すると大幅に計算が効率化されていることがわかる.

5　まとめ

難易度の高い楽器演奏時の手指の詳細なトラッキングを非接触で実現するために, 画像特徴量ではなく 2 次元手関節推定の結果を使ってその逆投影問題として定式化し, 機械学習とのハイブリッドな最適化によってこれを効率的に解けることを示した. また, 音情報を併用することによってバンドル調整だけでは得られない絶対スケールや鍵盤平面を決定できることを示した.

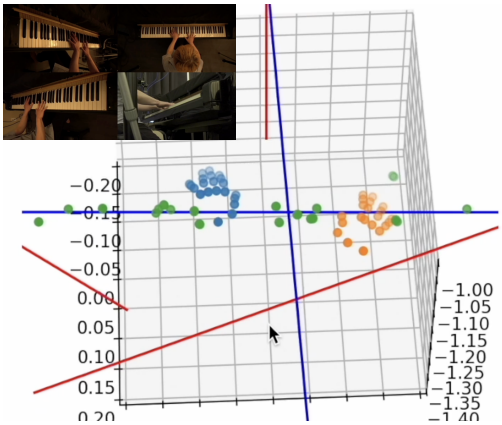

図 4　手関節推定結果.

提案手法にはまだいくつかの問題が残されている. まず, 演奏するフレーズによっては片手が完全にもう片方の手に遮蔽されてしまう. これに対応するためにはカメラを下側からボトムアップアングルで撮影する必要がある. しかしこれを実際に試してみたところ, 従来の画像からの 2 次元手認識手法では左右の手を混同してしまうなどの問題が発生したためかえって精度を悪化させた. これを解決するためには楽器下側からの手の画像を収集しての手認識モデルの学習をおこなう必要があるが今後の課題である.

謝辞

本論での実験収録については, Seoul National University と KAIST 研究グループからの支援を受けた.

参考文献

[1] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmannl, “Mediapipe hands: On-device real-time hand tracking,” in *CVPR*, 2020.

[2] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, “Bundle adjustment — a modern synthesis,” in *ICCV*, 1999.

[3] J. L. Schönberger and J.-M. Frahm, “Structure-from-motion revisited,” in *CVPR*, 2016.

[4] L. Kenneth, “A method for the solution of certain non-linear problems in least squares,” in *Quarterly of Applied Mathematics*, 1944.

[5] T. Igarashi, T. Moscovich, and J. F. Hughes, “As-rigid-as-possible shape manipulation,” in *SIGGRAPH*, 2005.

[6] M. A. Fischler and R. C. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” in *Communications of the ACM*, 1981.

[7] B. Doosti, S. Naha, M. Mirbagheri, and D. J. Crandall, “Hope-net: A graph-based model for hand-object pose estimation,” in *CVPR*, 2020.

[8] F. Gomez-Donoso, S. Orts-Escolano, and M. Cazorla, “Large-scale multiview 3d hand pose dataset,” in *arXiv:1707.03742*, 2017.

[9] Z. Yu, L. Yang, S. Chen, and A. Yao, “Local and global point cloud reconstruction for 3d hand pose estimation,” in *BMVC*, 2021.