ハンドトラッキングにおける オクルージョン解決のための手画像復元

川上童夢[1]　竹内康太[1]　岡夏樹[2]　田中一晶[1]

概要：機械学習技術の進歩により，カメラ画像のみを用いた姿勢推定の精度が高くなっている．しかしながら，物体によって推定対象となる部分が隠れてしまうとその姿勢推定精度が著しく低下してしまう．特に手の姿勢推定においては，ものを扱うなど動きが多いことから，扱う対象で手が隠れてしまうことが頻繁に発生する．そこで，手を遮蔽する物体に対して隠消現実感技術を用いることで画面内から物体を削除し，完全な手の画像を生成することができれば，オクルージョンが発生している状況でも既存の姿勢推定モデルをそのまま使用できると考えた．この目標に向け，本研究では，物体を削除して手の画像を復元する機械学習モデルの構築を行った．手の画像に対して 9 パターンのマスクを適用させ，作成した画像復元モデルを用いてマスク部分を復元することで，疑似的にオクルージョンの解消を行った．その結果，すべてのマスクパターンにおいて，既存の姿勢推定モデルを用いた手の姿勢推定の精度が向上し，元の手の画像と近い姿勢推定を行えるようになった．また，手を遮蔽する物体に対して自動でマスク処理を行い，手の画像を復元するシステムの試作を行った．

1. 緒言

人の身体の姿勢を推定する技術は，機械学習技術の進展により手軽に利用できるようになってきた．この技術は，スポーツ科学での動作解析[1]，エンターテインメント分野では CG アニメーション[2]，骨格推定を用いて不自然な動きを検出するセキュリティシステム[3]，ロボットハンドの遠隔操作[4]など，様々な分野で活用されている．このように多岐にわたって利用される中で，正確な姿勢推定が行えることと同様に，特殊なデバイスを装着せず，手軽に使用できることも重要視されている．

姿勢推定の一般的なシステムとして，多数の再帰性反射材のマーカを身体に取り付け，環境に配置した複数の赤外線カメラでそれらを追従するモーショントラッカーが挙げられる．特に，手の姿勢推定は複雑であるため，静電容量式近接センサ[7]やグローブ型触覚デバイス[4][8]など，専用のハードウェアを使用する姿勢推定方法が提案されている．装着デバイスを使用する姿勢推定システムでは，高精度の推定が可能であるが，専用のハードウェアが高価でコストがかかることや，デバイス装着する手間があり，手軽に利用できるとは言い難い．

そこで，より手軽に姿勢推定を行えるようにする方法が研究されてきた．例えば，カメラ画像から姿勢推定を行うモデルが提案されており[5][6]，Web カメラ等の安価なデバイスがあれば姿勢推定を行うことができる．最近では，機械学習技術の進歩により，簡単な装置でも高い精度で姿勢推定を行えるモデルが公開されている[6]．しかしながら，このようなカメラ画像を用いた姿勢推定では，オクルージョンが発生すると（身体の撮影が物体に阻まれると）姿勢推定の精度が大きく低下し，場合によっては推定が途切れてしまうという問題が発生する．

そこで本研究では，カメラ画像を用いて手の姿勢推定を行う際にオクルージョンが発生する状況を想定し，オクルージョンの原因となっている物体を画像内から消去し，物体によって隠された手の画像を復元する手画像復元モデルを開発する．このモデルを使用することで，オクルージョンが発生してもカメラ画像ベースの姿勢推定モデルをそのまま使用することが可能になる．

2. 手法

2.1 手画像の復元

本実験では，手の画像におけるオクルージョンの原因となる物体に対してマスクを施し，画像復元モデルを使用してマスク部分を補完する．これにより，元画像からオクルージョンが生じていない自然な手画像の生成を目指す．この画像の復元モデルを構築するにあたり，顔画像を復元する学習済みモデル（以下，顔画像復元モデルと呼ぶ）が画像復元モデルのライブラリ[9]で提供されており，これを手画像でファインチューニングすることとした．人により異なる手に対してこのモデルを使用できるようにするため，様々な人の手の画像で学習を行う必要がある．以下ではそのデータセットの収集方法とデータセットへのマスクの付与方法について説明する．

2.2 手画像データセットの収集

データセットの収集は図 1 の環境で行った．中心に穴を開けた白のスチレンボードで上下左右を囲み，穴から覗くように Web カメラを設置している．この空間に手を差し込むことで，4 方向から対面のスチレンボードを背景とした

1 京都工芸繊維大学
2 宮崎産業経営大学

手の画像を撮影する．

撮影する際の手の姿勢（各指の屈曲伸展等）としてひらがな 46 種類, アルファベット 26 種類の指文字を採用した．ひらがなとアルファベットでは同じ形の指文字が存在することから，重複が無いように選出した結果，44 種類の指文字を使用することとした．それらをランダムに並び替え，32 通りの指文字列を設定した．この指文字列は，可能な限り同じ並び方にならないように設定した．

様々な手の形状について学習を行うため，本学学生である 16 人（男性 12 人, 女性 4 人）の実験参加者を採用した．各参加者は左右の手で 1 通りの指文字列の順番で手を動かし，その間の画像を撮影した．このデータ収集実験では，手を動かしている途中のフレームなど，モーションブラーが発生して姿勢推定が行えない場合の画像を保存してしまうことを防ぐため，姿勢推定モデルである MediaPipe[6]を使用して画像内に手を認識できる画像のみを保存した．また，同じような姿勢の手画像を収集し過ぎてしまうことを防ぐため，姿勢推定結果であるランドマークの位置座標の変位（前に保存した画像におけるランドマークの位置座標とのユークリット距離）の合計値が閾値を超えた場合のみその画像を保存することとした．閾値は画像の画素数（640×480）の 0.05%である 153.6 とした．その結果，33,934 枚の手画像が得られた．また，手画像を学習させる元のモデルである顔画像復元モデルの入力は 160×160 であったため，中央部分の 480×480 を残してトリミングした後，160×160 に縮小した．その際にも MediaPipe を使用して，姿勢推定が可能な画像のみを採用した．その結果，手画像データセットに含まれる画像は 33,584 枚となった．この手画像データセットからランダムに取り出した 60%を訓練データ, 20%を検証データ, 残りの 20%をテストデータとした．

2.3 手画像へのマスクの付与

使用した手の姿勢推定モデルでは 21 個のランドマークで手の姿勢の推定結果が得られる．そのランドマークが重心となるようにマスクを描画することで，画像内の手の領域に重なるようにマスクを付与することができる．この画像を以下ではマスク画像と呼ぶ．

描画するマスクの形状として四角，丸，三角の 3 種類を設定した．また，そのサイズとして，各形状で画像内の手の領域に基づいて小，中，大の 3 サイズ用意した．マスクの大きさを決定する基準として，図 2 に示されるように，認識された手のランドマークを包含する長方形の面積の 40%を小，60%を中，80%を大とする．したがって，同一の手画像に対して生成される小，中，大のマスクの面積は四角，丸，三角のいずれの形状であっても同じ面積となる．このようにして決定した形状・サイズのマスクを全てのランドマーク上に描画するため，1 枚の手画像において 189 枚（形状 3×サイズ 3×ランドマーク 21）のマスク画像が得られる．

図 1　手画像を収集した環境

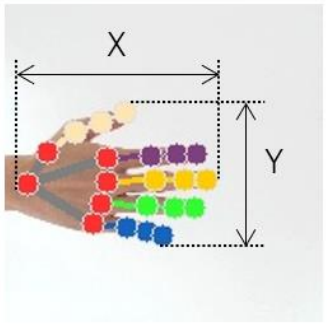

図 2 マスクサイズを決定するための手のサイズ基準

3. 実験

3.1 手画像復元モデルの学習

顔画像復元モデルに対して，作成した手画像データセットを用いてファインチューニングを行う．学習は 3 つのステージに分割されており, ステージ 1 は Generator の学習，ステージ 2 は Generator の重みを固定した Discriminator の学習，ステージ 3 は Generator と Discriminator を並行して学習を行う. 各ステージのトレーニングステップについて，ミニバッチのサイズを 128 とした学習を，ステージ 1 では 100000 ステップ，ステージ 2 は 10000 ステップ，ステージ 3 は 400000 ステップ行った．学習は，1 台の GPU（NVIDIA A100）を用いて 1930 時間実施した．

3.2 評価指標

学習時における学習データと検証データとは異なるテストデータを用いて，マスクされた画像から手の画像を復元する精度を評価した．その評価指標として認識率とハンドモデルの変化量という 2 種類を設定した．これら 2 つの評価指標について，テストデータにおけるマスク画像，顔画像復元モデルを用いた復元画像，手画像でファインチューニングを行った手画像復元モデルでの復元画像に対し，形状 3×サイズ 3 の 9 パターンに分類して比較する．

認識率：マスク画像では手の領域がマスクされることで

4. 結果

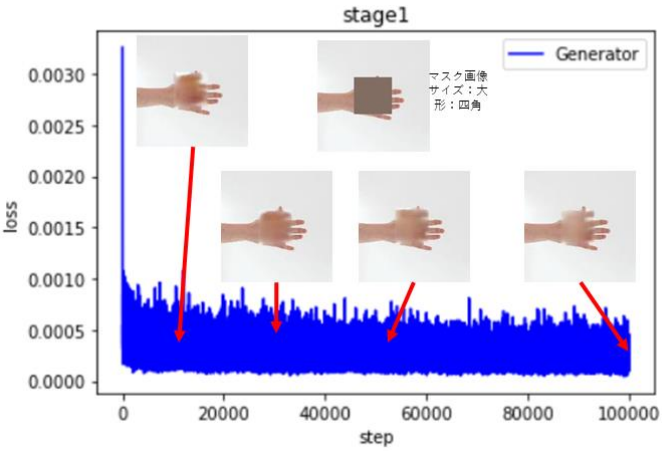

図 3 ステージ 1 における Generator の学習の損失と
各ステップにおける復元画像の例

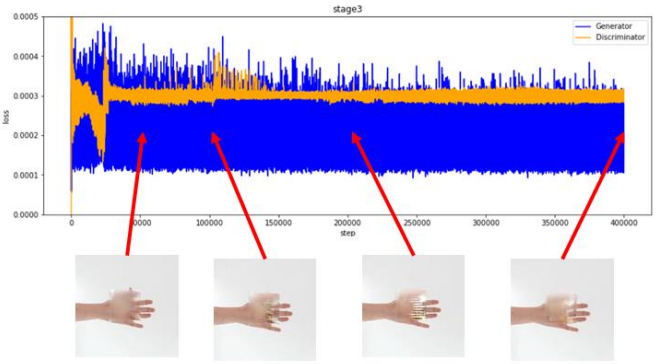

図 4 ステージ 3 における Generator（青）と
Discriminator（橙）の学習の損失と
各ステップにおける復元画像の例

表 1 テストデータにおける手の認識率（高いほど良い）

マスクサイズ	小			中			大		
マスク形状	○	△	□	○	△	□	○	△	□
復元なしマスク画像	94.60%	92.85%	93.37%	84.51%	78.04%	78.65%	66.07%	58.22%	57.65%
顔画像復元モデルによる復元画像	97.97%	97.71%	97.95%	94.77%	94.31%	95.39%	82.39%	77.66%	83.30%
手画像復元モデルによる復元画像	98.02%	97.94%	98.16%	95.42%	95.65%	95.59%	85.68%	85.68%	85.97%

表 2 テストデータにおける元画像からの姿勢推定の変化量（少ないほど良い）

マスクサイズ	小			中			大		
マスク形状	○	△	□	○	△	□	○	△	□
復元なしマスク画像	123.01	137.20	132.98	232.65	239.15	237.39	323.93	323.11	328.37
顔画像復元モデルによる復元画像	88.88	89.90	84.73	139.63	140.61	132.57	208.05	203.68	192.29
手画像復元モデルによる復元画像	85.20	86.00	81.77	128.76	130.28	125.25	192.96	189.64	180.39

手の姿勢推定モデルでは手が認識できなくなる場合がある．手画像の復元によってそれをどれだけ改善できたか評価するため，手の姿勢推定モデルにおいて評価対象となる画像群で手が存在すると認識できた数をカウントし，テストデータ数で割った割合を認識率として算出した．

変化量：手画像の復元によって画像内に手があると認識できるようになったとしても，正しく復元できていなければ元画像と同様の姿勢推定が行えない場合がある．手画像の復元によって，どれだけ元画像に近い姿勢推定が行えるようになったか評価するため，評価対象となる画像群において姿勢推定モデルで手が認識できた画像を対象とし，推定されたランドマーク座標と，元画像で推定されたランドマーク座標とのユークリッド距離を算出した．その平均値が変化量である．

4.1 学習の損失と復元画像

図 3 と図 4 は Generator の学習を行うステージ 1 とステージ 3 における学習の損失であり，青色が Generator，橙色が Discriminator の損失を表している．同時に，あるマスク画像における各学習段階での復元例を示す．学習が進むごとに，マスク部分の復元が鮮明に変化している．また，図 5 に，元画像，マスク画像，顔画像復元モデルを用いた復元画像，手画像復元モデルによる復元画像のセットを示す．同時に，手の姿勢推定システムを用いた場合の認識率と変化量を示す．

図 5 では，マスク画像においても領域内に手を認識できたものを例として挙げているが，いずれもマスクによって

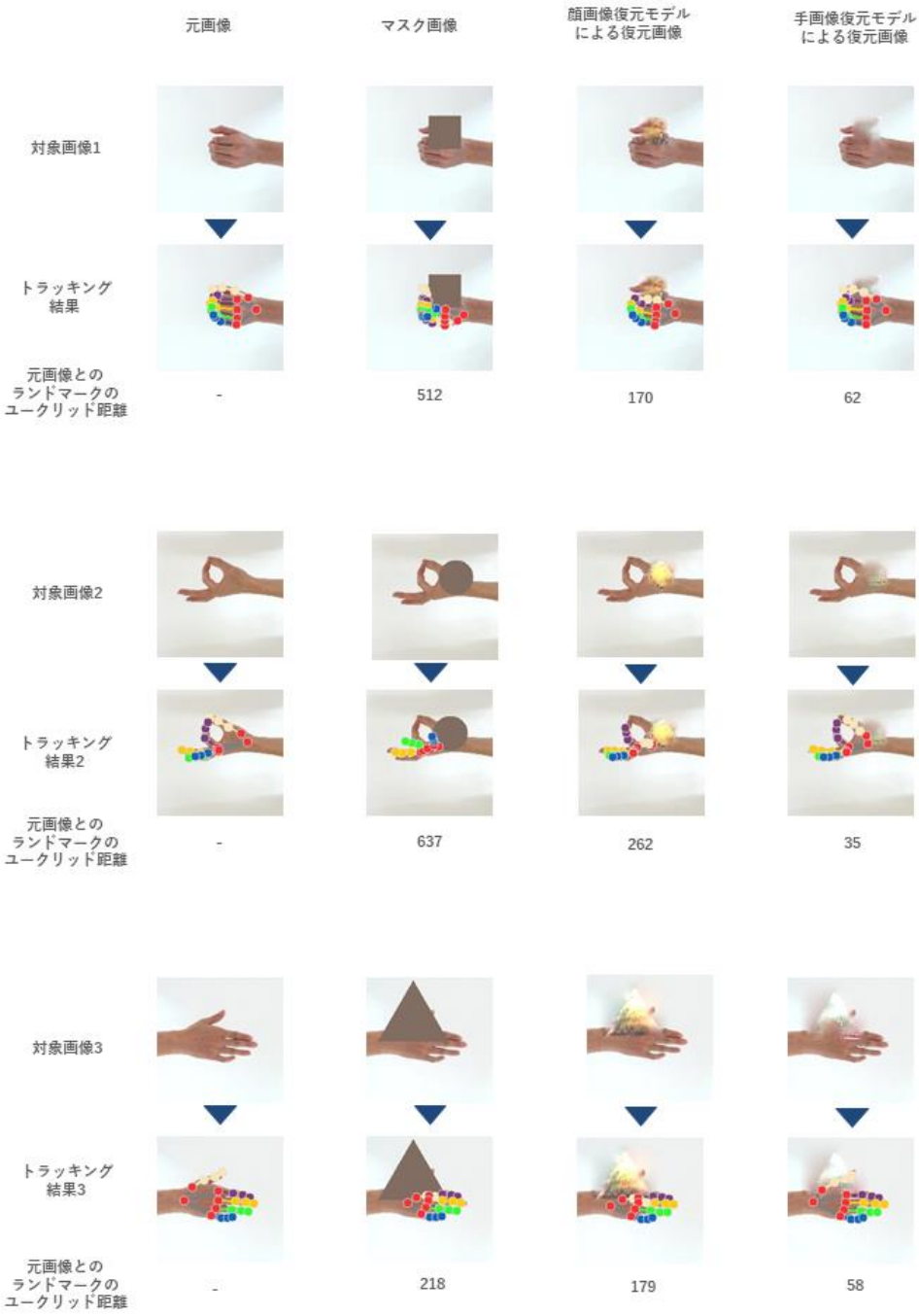

図 5 手の画像復元及び姿勢推定の例

正しい姿勢推定が行えなくなっていることが見受けられ，変化量が高くなっている．顔画像復元モデルを用いた復元画像において，ある程度の改善が見られるものの，元画像での姿勢推定とは依然として大きく異なっている．これらに対し，手画像復元モデルによる復元画像では元画像に近い姿勢推定を行えており，変化量が小さくなっている．このことから，手画像によるファインチューニングによって構築した手画像復元モデルにはオクルージョンを解決する効果があると考えられる．以下では，テストデータ全体に対する認識率と変化量を示す．

4.2 認識率

復元無しマスク画像，顔画像復元モデルを用いた復元画像，手画像復元モデルによる復元画像における認識率を表 1 に示す．認識率は 3 種類のマスクの形状と 3 種類の大きさに分けて算出している．全てのパターンで，手画像復元モデルによる復元画像において，手の認識率が向上していることが分かる．また，この結果には若干劣るが，顔画像

復元モデルでの復元画像においても認識率の向上が見られた．

4.3 変化量

復元無しマスク画像,顔画像復元モデルを用いた復元画像,手画像復元モデルによる復元画像における変化量を 表 2 に示す．変化量は 3 種類のマスクの形状と 3 種類の大きさに分けて算出している．全てのパターンで，手画像復元モデルによる復元画像において変化量が小さくなっており，手画像の復元によってオクルージョンが発生していない状況での姿勢推定に近づけることができたと言える．顔画像復元モデルを用いた復元画像においても同様の効果が見られるが，手画像復元モデルよりもその効果が低いことも見受けられる．

5. オクルージョンの原因物体への自動マスク処理及び手画像の復元

5.1 マスク処理の自動化

提案手法を実際に使用するためには，オクルージョンの原因物体に対し，自動でマスク処理を行う必要がある．そこで，画像内の物体を自動でセグメンテーションする Segment Anything[10]の高速版である FastSAM[11]を用いることで手を遮蔽している物体を特定し，その物体のみをマスク処理する．このセグメンテーションモデルでは，深度画像が必要であるため，深度センサ付きカメラで手の RGDB 映像を撮影した．

最初のフレームでは，追跡対象となる手が完全に映った RGB 画像において MediaPipe を使用して手の位置を特定する. ここで取得した手のランドマークから中心点を決定し，その周りにランドマーク座標の分散と深度センサの値から手の部分を囲んだ正方形の領域（以下，手のバウンディングボックス）を設定する．ここで設定したバウンディングボックス内で FastSAM を実行し，追跡対象となる手の領域を取得する．

2 フレーム目以降では，前フレームの手の領域と深度画像を用いて手のバウンディングボックスを求め，バウンディングボックス内で FastSAM を実行し，手の遮蔽物があればその領域をマスクする．その際，前フレームの手の領域との類似度を算出することで MediaPipe を使用せずに手の領域を追跡する．手の領域の前フレームとの差分領域に被っており，深度画像において手よりもカメラに近い物体を手の遮蔽物として特定し，RGB 画像においてマスク処理する．

5.2 E2FGVI での手画像復元

4 節でファインチューニングした手画像復元モデルの性能は現時点では十分とは言えない．そこで，上述の手法でマスク処理を行った動画を作成し，動画修正の End-to-End フレームワークである E2FGVI[12]を用いることでを動画内のマスク領域（つまり手を遮蔽している領域）の復元を行った．動画として処理することで，一連のフレームの繋がりから手画像の復元がより自然に行える可能性があると考えた．しかしながら，自然な復元が行えたとしても手の映像を動画として扱うため，リアルタイムな姿勢推定に適用することは現時点ではできていない．

5.3 動画における手画像復元結果

手にオクルージョンが発生している動画において手画像復元を行った結果を図 6 に示す．遮蔽物によって MediaPipe の姿勢推定が途切れたシーンを(a)に，その遮蔽物に 5.1 節の手法で手の遮蔽物に自動でマスク処理を行った結果を(b)に, E2FGVI で手画像復元を行った結果を(c)に，その結果において MediaPipe で姿勢推定を行った結果を(d)に示す．この結果の通り，手画像復元を行った動画では途切れることなく安定的に姿勢推定が行われた．

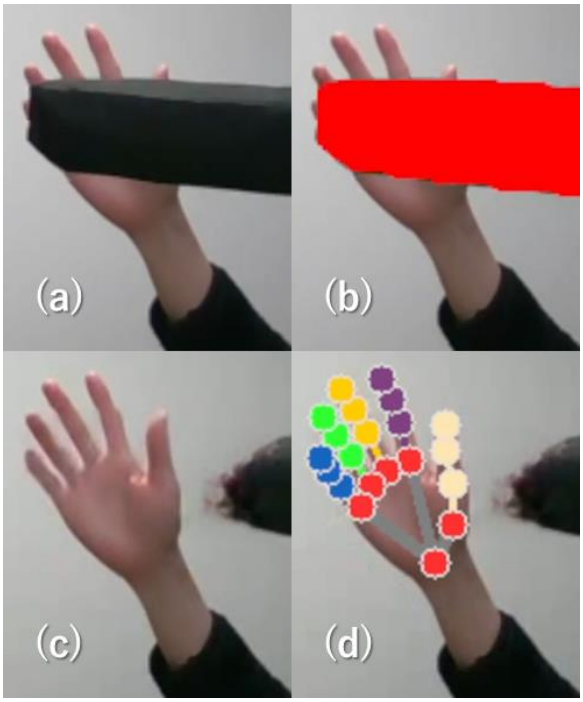

図 6　遮蔽物体への自動マスク処理と手画像復元結果

6. 考察

実験の結果，オクルージョンの発生を想定したマスク画像において，手画像復元モデルを用いてマスク部分の復元変化量のいずれの評価指標においても示された．したがって，手の姿勢推定中にオクルージョンが発生しても本研究で構築した手画像復元モデルを使用することで継続して姿勢推定が行える可能性がある．

ファインチューニング前の顔画像復元モデルによる復元画像と手画像復元モデルによる復元画像の結果を比較すると，マスクのサイズが大の場合にファインチューニングの効果が見られるが，他のサイズでは大きな差は見られなかった．これは作成したデータセットの背景が白であったことに起因すると考えられる．図 5 を見ると，顔画像復元モデルによる復元画像では，明らかに手とは異なる像が生

成されているが，その部分が白い背景とは混ざらずにマスクで途切れた部分を補うため，姿勢推定モデルにおいて手として認識されやすくなったことが考えられる．しかしながら，手の姿勢は元画像とは異なって認識されやすくなっており，変化量にはマスクサイズが大きくなるほど悪影響が表れている．これに対し，手画像復元モデルによる復元画像ではノイズが少なく，比較的自然な色や形の像を生成していることから，元画像により近い姿勢推定が行えたと考えられる．したがって，複雑な背景画像でファインチューニングを行った場合は，その効果がより明確になる可能性がある．

手画像復元モデルを構築するにあたり，ファインチューニングに使用した手画像は 33,584 枚であり，要した時間は 213 時間であった．本研究で構築した手画像復元モデルにおいて 1 枚の手画像を復元するのにかかる時間は約 0.13 秒であった．従って，このモデルで姿勢推定を行った場合，他の処理を無視したとしても高々 8fps の低フレームレートになってしまう．したがって，オクルージョンを解決しながらより高いフレームレートでの円滑な姿勢推定を行うためには，手画像復元にかかる処理時間の短縮が必要となる．また，オクルージョンの原因となる物体を自動で検出してマスクするシステムも必要となる．これらは今後の課題である．

動画における手の画像復元（5 節）では，姿勢推定を行う上で十分な精度で手画像復元が行われた．その結果として，オクルージョンによって姿勢推定が途切れたシーンであっても継続的に姿勢推定が行えることが確認できた．しかしながら，この結果は録画した映像を処理したものである．リアルタイムにこの精度で手画像復元を行い，高いフレームレートで継続的に姿勢推定を行えるようにすることは今後の課題である．

7. 結論

手の姿勢推定において，手にオクルージョンが発生すると姿勢推定の精度が低下する．発生したオクルージョンによっては，手の大部分が隠れてしまい，姿勢推定を行うことが困難になる．この問題に対して，隠消現実感技術を用いることでオクルージョンの原因となる物体を画面内から削除し，可能な限り自然な手の画像を生成することによって，既存の姿勢推定モデルであっても途切れることなく姿勢推定を継続できるようになると考えた．

そこで本研究では，オクルージョンの原因となる物体にマスク領域を描画し，マスクが適用された部分を復元する手画像復元モデルを構築した．このモデルを使用して手の復元画像を生成し，手の姿勢推定モデルを使用してどの程度元画像に近い姿勢推定が行えるようになるか評価した．その結果，マスクによってオクルージョンが生じている画像では画像内の手が認識できなかったり，認識できたとしても元画像よりも大きく手の姿勢推定の結果が変わってしまったりしていたが，本研究の手画像復元モデルを使用するとそれらの結果の改善が見られた．本研究の成果によって，単一の RGB 画像のみの入力であってもオクルージョンに姿勢推定精度が左右されない姿勢推定システムが構築できるようになることが期待される．

さらに動画を用いたリアルタイムのオクルージョンの除去について，トラッキング対象を追跡することで常にオクルージョンを除去する可能性を示した．この技術は現在研究開発中であるため，今後の発展が期待される．

謝辞 本研究は，JSPS 科研費 JP22K12126，JP19K12081 の支援を受けた．

参考文献

[1] Hiroki Ozaki, Minoru Matsumoto, Hideyuki Nagao, Toshiharu Yokozawa, "Potential Use of Deep Learning-Based Pose Estimation in Sports Biomechanics", Journal of High Performance Sport, Vol.10 pp.167-182, 2022. https://www.jstage.jst.go.jp/article/jissjhps/10/0/10_167/_article
[2] SONY, mocopi, https://www.sony.jp/mocopi/ (参照 2023-10-7).
[3] https://prtimes.jp/main/html/rd/p/000000089.000043312.html asilla, AsillaPose, https://jp.asilla.com/post/about-asillapose (参照 2023-10-7).
[4] HaptX, HaptX Gloves G1, https://g1.haptx.com/learnabout (参照 2023-10-7).
[5] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A. Otaduy, Dan Casas, Christian Theobalt, "RGB2Hands: Real-Time Tracking of 3D Hand Interactions from Monocular RGB Video", ACM Transactions on Graphics Volume 39, Issue 6, Article No.:218, pp 1–16.
[6] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, Matthias Grundmann, "MediaPipe Hands: On-device Real-time Hand Tracking", Google Research. 1600 Amphitheatre Pkwy, Mountain View, CA 94043, USA.
[7] Kazuyuki Arimatsu and Hideki Mori, "Evaluation of Machine Learning Techniques of Hand Pose Estimation on Handheld Cevuce with Proximty Sensor", International Conference on Human Factors in Computiong Systems (CHI 2020), April 25–30, 2020, Honolulu, HI, USA.
[8] CyberGlove Systems, Cyberglove III, http://www.cyberglovesystems.com/cyberglove-iii (参照 2023-10-7).
[9] Otenim. 2019. GLCIC-PyTorch. San Francisco (CA): GitHub;https://github.com/otenim/GLCIC-PyTorch (参照 2022-12-22).
[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. arXiv preprint arXiv:2304.02643, 2023
[11] Xu Zhao and Wenchao Ding and Yongqi An and Yinglong Du and Tao Yu and Min Li and Ming Tang and Jinqiao Wang, "Fast Segment Anything", arXiv:2306.12156, 2023
[12] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, Ming-Ming Cheng, "Towards An End-to-End Framework for Flow-Guided Video Inpainting", arXiv:2204.02663, CVPR 2022