2ZB-01

手背画像に基づく隠れた指先の位置推定

趙　政†　　梅澤　猛‡　　大澤　範高‡

千葉大学融合理工学府†　　千葉大学大学院工学研究院‡

1. はじめに

ヘッドマウントディスプレイ(HMD)に映し出された仮想環境上の物体を手指で直接インタラクティブに操作する手法において、HMD の外装前面部に設置したカメラを使って手指を認識することで、外部環境に機器を設置する必要をなくすことができる[1]。しかし、この手法では、手の甲をカメラに向けることが多く、手指の一部がカメラから隠れるセルフオクルージョンが発生した際に指先が認識できない問題がある。

ここで、手の甲側の画像(手背画像)から指先の位置を推定することができれば、セルフオクルージョンに頑健な推定が可能となる。

そこで本研究では、RGB カメラからの手背画像を基に、手背画像のテクスチャやシルエットなどの特徴を分析することで、指先座標を推定する方法を提案する。指先が隠れていても高精度で座標を推定するために深層学習モデルを構築する。

2. 関連研究

清水目らは、手の甲の RGB 画像から得られる手のシルエットや皮膚のテクスチャを特徴量として、畳み込みニューラルネットワーク(CNN)モデルを構築し、親指と人差し指の指先間距離を推定する手法を提案している[2]。シルエットや指先間の距離の変化による MP 関節(指の第 3 関節)近くの皮膚のシワの変化を特徴量とし、モデルを構築することで、指先間距離を推定できることが示されている。

3. 研究目的

本研究では、RGB カメラで撮影した手背画像を用いて、指先が隠れている場合でも、手の甲のテクスチャなどの特徴を分析し、指先の座標を推定する深層学習モデルを構築し、高精度の指先座標推定を実現することを目的とする。

4. 提案手法

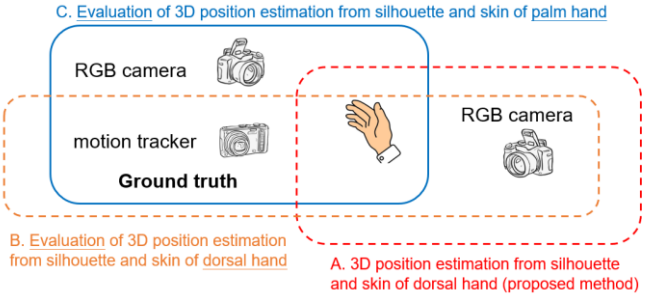

図 1　提案手法

提案手法の目的は、図 1 のケース A に示すように、指先が隠れる場合に、RGB カメラから撮影した手背画像を、深層学習モデルに入力し、すべての指先の 3 次元座標を推定することである。

推定精度を評価するために、ケース B に示すように、手のひら側にモーショントラッカを設置して指先の座標を記録し、正解とする。

ケース C はケース B の前段階として、指先が見えている手掌画像から、深層学習による指先座標の推定精度を評価し、ケース A の評価のベースラインとなる。

5. データ収集

深層学習モデルの訓練データとして、指先の三次元座標を正解ラベルとして対応付けた RGB 画像が必要である。データ収集のパイプラインを図２に示す。

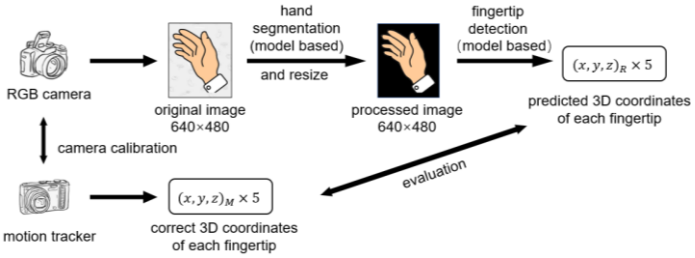

図 2　データ収集のパイプライン

5. 1 実験用のカメラ

実験で使用したカメラは 2 台の RealSense D435i である。一台は RGB カメラとして、もう一台はモーショントラッカとして利用する。モーショントラッカ用 RealSense の深度センサから得ら

3D position estimation of fingertip based on an RGB image of dorsal hand
†Zhao Zheng, Graduate School of Science and Engineering, Chiba University
‡Takeshi Umezawa, Noritaka Osawa, Faculty of Engineering, Chiba University

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.

れた奥行距離と画像を MediaPipe[3] によって処理し、推定した二次元座標から三次元座標を計算する。

5.2 座標系変換

モーショントラッカと RGB カメラの画角と座標系が異なるため、座標系変換を行う必要がある。本研究では、Zhang ら[4]の手法に従って、あらかじめ用意したチェスボードを利用して、カメラごとのキャリブレーションを行う。2 台のカメラが同時にチェスボードを撮影し、対応点を一致させることで、2 台のカメラの位置関係のレジストレーションを行う。

実験環境の再投影の二乗平均平方根誤差(RMSE)は 0.274 ピクセルであった。

5.3 手部の抽出

画像から手の特徴を捉えるために、画像から背景を取り除いて手を抽出する処理を行う必要がある。皮膚の色によって手部を抽出する方法は、背景にある皮膚の色に近いピクセルを排除することが困難である。また、前景領域抽出よく使われる GrabCut は、画像の解像度が高くなるほど処理時間が大きくなる問題があるため、本研究の用途には適さない。そこで、深層学習モデルを利用する手法を検討した。

本研究では転移学習手法を用いて、手作業で正解ラベルを付けた画像 80 枚を訓練データに、20 枚をテストデータに分割し、すでに訓練された RefineNet-Res101 モデルに入力し、エポック数は 30、学習率は 0.00001 で訓練を行った。

評価結果では、mPrec(mean Precision)が 98.9%で、mRec(mean Recall)が 98.8%で、mIOU(mean Intersection over Union)が 97.7%であった。

6. モデル構築

同時に撮影した手掌画像と手背画像をそれぞれ 1154 枚取得し、訓練データ 925 件、テストデータ 229 件に分割した。ラベルがユーザ片手すべての指先三次元座標であるため、モデルの出力は 15 次元である。

訓練済の ResNet18、ResNet50、VGG16 を基に推定モデルを構築する。エポックを 50、学習率を 0.0001 に設定して訓練を行った。

モデルは RMSE によって評価した。結果を図 3 に示す。青色とオレンジ色がそれぞれ、手掌画像と手背画像を使用したモデルごとの RMSE である。手掌画像を用いた場合には、3 つのモデルに大きな違いはなかった。それに対して、手背画像を用いた場合に、ResNet18 の RMSE が明らかに大きくなり、ResNet50 と VGG16 の RMSE が 1.5mm 程度増加した。VGG16 より、ResNet50 の画像 1 枚あたりの処理時間が 10ms ほど短く、18.67ms であった。

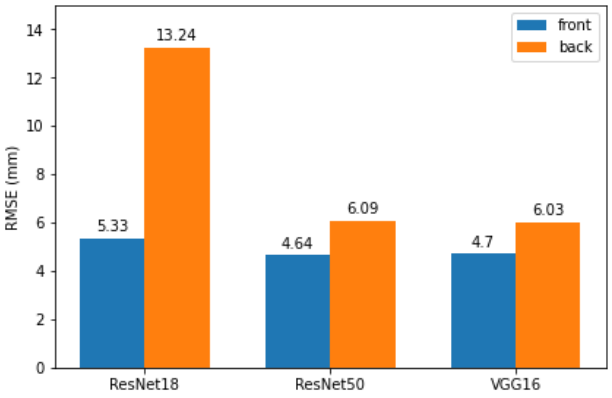

図 3　モデルごとの RMSE

7. まとめ

手掌画像による推定の RMSE と、セルフオクルージョンによって指先が隠れる場合の手背画像による推定の RMSE の差が 1.5mm 以下で、手掌画像に近い精度での推定が可能であることが示唆された。

ResNet50 と VGG16 の精度はほぼ同じで、ResNet50 の方が処理時間が短いため、今後、ResNet50 を用いて、Leap Motion などの既存手法と比較検討を行う予定である。

参考文献

[1] Park, Gangrae, et al. Virtual figure model crafting with VR HMD and Leap Motion. In: The Imaging Science Journal 65.6. 2017.pp. 358-370.

[2] Takuma Shimizume, Takeshi Umezawa, Noritaka Osawa. Estimation of distance between thumb and forefinger from hand dorsal image using deep learning. In: Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology. 2018. pp. 1-2.

[3] Zhang, Fan, et al. MediaPipe Hands: On-device Real-time Hand Tracking. In: arXiv preprint arXiv:2006.10214. 2020. https://arxiv.org/abs/2006.10214

[4] Zhengyou Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In: Proceedings of the Seventh IEEE International Conference on Computer Vision. 1999. Vol. 1. pp. 666-673.

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.