4ZE-01

密結合畳み込みネットワークを用いた手背画像で隠れている指先の3次元位置推定

阿部　竜弥†　梅澤　猛‡　大澤　範高‡

千葉大学工学部総合工学科情報工学コース†　千葉大学大学院工学研究院‡

1. はじめに

仮想現実空間において指先を使った操作をする場合，手の画像を基に指先の位置情報を取得することができる．画像を取得するカメラをヘッドマウントディスプレイに装着することを想定すると，手を撮影するときに自身の手の甲などで指先が隠れてしまうセルフオクルージョンが生じる場合がある．このとき，セルフオクルージョンを含んだ画像から指先位置を高精度に推定することが重要となる．

本研究では，指先がセルフオクルージョンによって隠された手背画像から，隠された指先の3次元位置推定を行うモデルを構築する．より高精度な位置推定ができるようにDenseNet[1]などの密結合畳み込みネットワークを基に，推定モデルを構築する手法を評価する．

趙らは，手背画像から指先位置を推定する手法を提案し，推定モデルを構築した[2]．推定モデルはResNet18，ResNet50，VGG16を基に構築し，手背画像から指先の3次元座標を推定したときの二乗平均平方根誤差（RMSE）はすべて5.0mm以下となった．これにより指先の3次元座標を推定する際に，手背画像を基に指先位置を推定できることが示された．しかし指先を使った操作をするためにはさらに高い位置推定精度が必要である．

2. DenseNetによるモデル構築

DenseNetは，各層の入力が直接接続しているDense Blockと，チャンネル数を圧縮し解像度を下げるTransition層を交互に組み込んだConvolutional Neural Network（CNN）である．Dense Blockは図1のような構造をしており，フィルター数nの層から1×1畳み込み層へ分岐する．この層はボトルネック層と呼ばれており，3×3畳み込み層へ移行する前にチャンネル数の増加に制限をかけ，パラメータ数を削減する効果がある．残差を伝搬させる構造はResNetと同様であるが，ResNetで加算していた残差を，DenseNetでは連結する．これにより各層が連結し，勾配の伝搬するルートを確保しているため，勾配消失問題を削減することができる．kは成長率と呼ばれるハイパーパラメータであり，増加させるフィルター数を調整することができる．

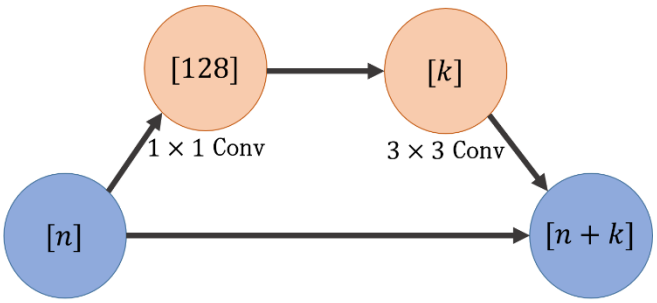

図1 Dense Blockの構造

DenseNetとResNetのパラメータ数と層数，層あたりのパラメータ数の比較を表1に示す．DenseNetは同程度の深さのResNetと比較すると，表1のように層あたりのパラメータ数が少ないため，訓練時間の短縮が期待できる．

表1 各CNNのパラメータ数と層数

	パラメータ数	層数	パラメータ/層
ResNet18	11,689,512	54	216472.4
ResNet50	25,636,712	168	152599.5
DenseNet121	8,062,504	121	66632.3
DenseNet169	14,307,880	169	84662.0

3. データ収集

3.1. 使用機材

撮影時のRGBカメラおよびモーショントラッカに，Intel社のRealSense D435iを使用する．

3.2. データセット構築

手掌側から1台のRGBカメラと1台のモーショントラッカを，手背側からRGBカメラを使用し，手の動作を同時撮影する．モーショントラッカ

Position Estimation of Fingertips Occluded on Dorsal Hand Image with Densely Connected Convolutional Networks
†Tatsuya Abe, Department of Information Engineering, Faculty of Engineering, Chiba University
‡ Takeshi Umezawa, Noritaka Osawa, Graduate School of Engineering, Chiba University

で得られた深度情報は，MediaPipe Hands[3]を使用して取得した手掌画像による指先の3次元座標に変換し，これを正解ラベルとする．

次に趙らの提案した転移学習を利用する手法[2]を採用して 画像から手部のみを抽出する．

4. 推定モデル構築

本研究では ImageNet によって事前学習済みの DenseNet121 ， DenseNet169 ， ResNet18 ， ResNet50 を基に収集したデータで転移学習することで推定モデルを構築した．DenseNet121，DenseNet169 の全体構造を表 2 に示す．

表 2 DenseNet の全体構造

Layers	Output Size	DenseNet 121	DenseNet 169
Convolution	112×112	7×7 conv, stride 2	
Pooling	56×56	3×3 max pool, stride 2	
Dense Block	56×56	[1×1 conv, 3×3 conv]×6	[1×1 conv, 3×3 conv]×6
Transition Layer	56×56	1×1 conv	
	28×28	2×2 average pool, stride 2	
Dense Block	28×28	[1×1 conv, 3×3 conv]×12	[1×1 conv, 3×3 conv]×12
Transition Layer	28×28	1×1 conv	
	14×14	2×2 average pool, stride 2	
Dense Block	14×14	[1×1 conv, 3×3 conv]×24	[1×1 conv, 3×3 conv]×32
Transition Layer	14×14	1×1 conv	
	7×7	2×2 average pool, stride 2	
Dense Block	7×7	[1×1 conv, 3×3 conv]×16	[1×1 conv, 3×3 conv]×32
Classification Layer	1×1	7×7 global average pool	
		1000D fully-connected, softmax	

各モデルの入力サイズは 640×480 とした．各モデルの出力はそれぞれ 1024 次元，1664 次元，512 次元となっているため，出力が片手の 5 本の指先 3 次元座標となるように，表 1 における Classification 層を変更し，全結合層を用いて 15 次元に線型変換した．構築したデータセット（11558 件）を 8:2 に分割し，訓練データ（9246 件）とテストデータ（2312 件）とした．Epoch 数を 80，学習率を 0.0001，DenseNet の成長率を 32 と設定した．損失関数には RMSE を用いた．データの総数をn，i番目のデータから得られた指先の 3 次元座標の正解値と推定値をそれぞれ$\boldsymbol{y}_i$，$\hat{\boldsymbol{y}}_i$，とすると，各指の二乗平均平方根誤差を RMSE’とすると，RMSE’は式（1）で表せる．式（1）に従って 5 本の指先の RMSE’をそれぞれ求め，それらを 5 で割った平均値を RMSE とする．

$$RMSE' = \sqrt{\frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i|^2} \qquad (1)$$

5. 評価

手掌画像と手背画像からそれぞれ位置推定を行い，その誤差を評価する．各推定モデルをテストデータによって評価したときの各指の平均 RMSE を図 2 に示す．棒グラフの系列は，それぞれ

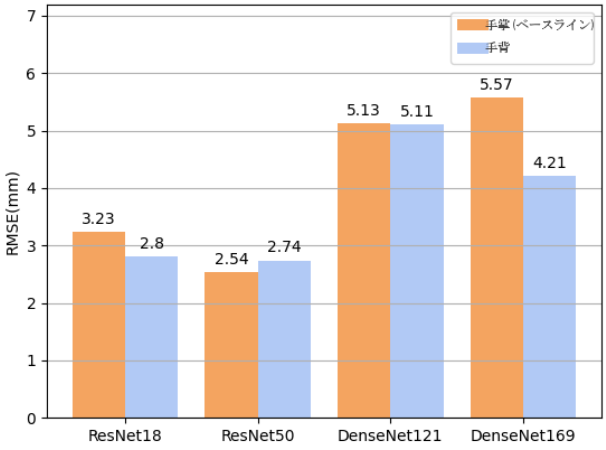

図 2 各モデルによる指先位置推定の RMSE

手背画像と手掌画像から構築した推定モデルの RMSE の平均である．各推定モデルの手掌画像と手背画像の訓練時間（秒）を表 3 に示す．

表 3 各モデルの訓練時間

推定モデル	訓練時間（手掌）[s]	訓練時間（手背）[s]
ResNet18	13328	14248
ResNet50	33988	34051
DenseNet121	30561	31144
DenseNet169	32902	32970

DenseNet121 より層の多い DesneNet169 の方が手掌画像と手背画像の RMSE の平均にばらつきが生じた．手背画像に基づく推定モデルの方が，ベースラインである手背画像に基づく推定モデルよりも誤差が小さい場合があり，他の研究では高い性能を示すことの多い DenseNet の方が ResNet よりも誤差が大きい．これらの結果の原因として，訓練データが不足していることが考えられる．今後は，訓練データを充実させるとともに推定精度と訓練時間が適切なモデルを検討し，評価を行う予定である．

参考文献

[1] Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger. Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp.2261-2269

[2] Zheng Zhao, Takeshi Umezawa, Noritaka Osawa. Position Estimation of Occluded Fingertip Based on Image of Dorsal Hand from RGB Camera. In:International Conference on Human-Computer Interaction HCII 2021: Virtual, Augmented and Mixed Reality pp.259-271. https://arxiv.org/abs/1608.06993

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