4R-01

単一画像からの上半身3次元モデル生成の検討

武藤　凌†　木全　英明†
工学院大学　情報学部†

1. はじめに

近年，5G をはじめとする通信技術の発展によるデータの大容量・高速・低遅延・多数同時接続通信の実現，加えてスマートフォン，HMD，モーションキャプチャ等のデバイスの小型化や，映像・音響ソフトウェアの高性能化により、現実世界と仮想世界をつなぐ XR 技術が注目を集めている．仮想空間で用いる 3 次元アバターの生成に関する研究が盛んに行われている一方，大規模な撮影スタジオを用いる手法や複数の撮影画像を用いる手法が多く，少量のデータや身体の一部分を用いた，限られた撮影条件での研究例は少ない．本研究では，深層学習を用いて単一画像から人間の全身の 3 次元モデル生成が可能な PIFu[1]のネットワークモデルを用いて，上半身の 3 次元モデルの生成を検討する．

2. 提案手法

実生活の中で大規模な撮影スタジオを用いる手法や，複数枚の撮影画像を用いる手法は非実用的である．また，全身の画像・動画を用いるケースは少なく，証明写真や Web 会議のように上半身のみで十分なケースが多い．

そこで，本研究では上半身の単一画像が与えられたとき、深層学習を用いて 3 次元モデルの形状とテクスチャを生成することを目標とする．

今回用いる PIFu[1]のアーキテクチャを以下の図 1 に示す．アーキテクチャは大きく表面形状推定を行う PIFu とテクスチャ推定を行う Tex-PIFu に分類される．まず PIFu は 512×512 の背景を削除した RGB 画像を入力とし，積層砂時計型画像エンコーダ[2]で形状特徴量 Fv を抽出する．その後，多層パーセプトロン(MLP)をベースとした陰関数 fv に特徴量 Fv と奥行き z を入力し，3 次元の点 X の確率場を予測する．次に，Tex-PIFu は先程の入力画像と PIFu で求めた画像特徴量 Fv を入力とし，6 つの残差ブロックから

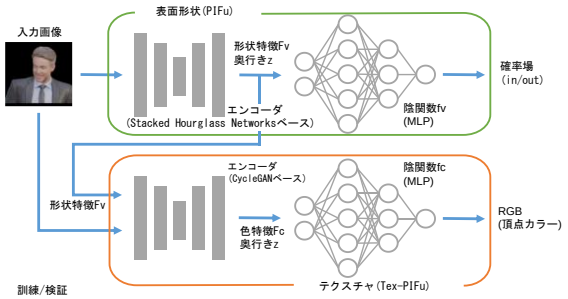

図 1 ネットワークアーキテクチャ

図 2 訓練画像生成時のモデル回転方向

なる CycleGAN[3]のアーキテクチャを用いた画像エンコーダで特徴量 Fc を抽出する．その後先程と同様に MLP をベースとした陰関数 fc に色特徴 Fc と奥行き z を入力し，頂点カラーを予測する．

本研究では PIFu の事前学習済みモデルに対して，自身が用意したデータで再学習を行う．その際，訓練データに関して工夫を行う．まず，先行研究では訓練画像生成時に図 2 左側のように 1 方向のみの回転であるが、本研究では，画像が人物の正面以外，特に上方や斜方からも撮影されることを考慮し，元の回転方向に加え，図 2 右側のように 3 方向の回転を行う．加えて，先行研究では全身画像を用いて訓練を行うが，実生活では人物の全身が写っている写真が少なく，また頭部のみでは情報量が少ないことを考慮し，上半身の画像を用いて学習を実施する．

3. 実験

3.1. 実験環境

本実験の実験環境を以下の表 1 に示す．

Study on Generation of a Upper Body 3D Model from a Single Shot Image
†Ryo Muto and Hideaki Kimata,
Faculty of Informatics, Kogakuin University

表 1 実験環境

プロセッサ	Intel(R) Core(TM) i5-10400 CPU @ 2.90GHz (12CPUs)
RAM	16.0 GB
GPU	NVIDIA GeForce RTX 3060(12GB)
仮想環境	docker 20.10.21
実行環境	Ubuntu 18.04.6 LTS/Python 3.8.8/Pytorch1.8.0

上記の実験環境において，エポック数 50，バッチサイズ 6，サンプルポイント数 5,000 で表面形状とテクスチャの学習にそれぞれ約 3 日を要した．

3.2. データセット

本実験では先行研究で用いられていた Renderpeople のサンプルに加え，CGTrader で公開されている無料モデルを合わせた，計 28 人の 3D モデルを用いた．具体的にはこれらの obj ファイルを Blender を用いて上半身を切り抜き，x, y, z 軸周りに 1°ずつ 360°回転させ，28 人×3 軸×360° = 30,240 枚の画像を生成した．これらの画像のうち、全体の約 80%にあたる 24,120 枚を訓練データ，約 20%にあたる 6,120 枚を検証データとした．

3.3. 評価実験

検証データの MSE，IoU，precision, recall を求め，定量評価を行った．また，DeepFashion データセット[4]をテストデータとし，データセット画像から生成した 3 次元モデルに関して，20 代の男女 14 人に以下のアンケート評価を行い，定性評価を行った．

(1)．提示画像：4 種類
(2)．角度：3 方向(正面，側面，背面)
(3)．再学習有無：2(有り，無し)
(4)．繰り返し：2 回

計 48 枚の画像に関して，形状とテクスチャの再現性についてそれぞれ 5 段階評価を行った．

4. 実験結果

実験結果を以下の表 2，3，図 3 に示す．なお，テクスチャに関して再学習モデルと比較し事前学習済みモデルの方が性能が良かったため，そちらを使用した．そのため，形状は再学習モデル，テクスチャは事前学習済みモデルを使用した．

表 2 生成モデルに関する定量評価

	MSE ↓	IoU ↑	precision ↑	recall ↑
学習	0.26	0.51	0.79	0.59
再学習	**0.20**	**0.60**	**0.85**	**0.67**

表 3 生成モデルに関するアンケート評価

	ベースライン		再学習	
	形状	テクスチャ	形状	テクスチャ
正面	4.39	4.07	4.38	3.92
側面	**1.45**	1.44	**2.09**	1.50
背面	2.40	1.61	2.60	1.65

図 3 入力画像と生成した 3 次元モデル例

まず表 2 より，事前学習済みモデルを用いない 0 からの学習と再学習を比較し，全ての評価項目において正の転移であることが確認できた．次に表 3 と図 3 より，画像正面の再現性は高い一方，側面や背面の再現性が低いことが分かった．また先行研究のベースラインと比較し，特に側面の形状について再学習モデルが優位であると分かった．

5. 考察

実験結果より，入力画像の正面以外の方向に関して形状，テクスチャ共に再現性が低く，モデルの方向を考慮せずに学習していることが原因であると考えられる．この課題を改善するためには，入力画像の前後左右を推定する，3 次元モデル生成時に方向別に処理を行う等，追加の処理を行う必要があると考えられる．

6. まとめ

本論文では，深層学習を用いて単一画像から人間の全身の 3 次元モデル生成が可能な PIFu のネットワークモデルを用いて，上半身の 3 次元モデルの生成を検討した．

実験結果より，側面形状の再現性が向上した反面，形状とテクスチャ共に正面と比較すると側面と背面の再現が困難であることが分かった．

今後の課題として，人物の前後左右の判別により形状の再現精度を向上させること，生成後のモデル形状に合わせたテクスチャ生成によりモデルの忠実性を向上させることが挙げられる．

参考文献

[1]Shunsuke Saito, et al, “PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization”, ICCV2019.
[2]Alejandro Newell, et al, “Stacked Hourglass Networks for Human Pose Estimation”, ECCV2016.
[3]Jun-Yan Zhu, et al, “Unpaired image-to-image translation using cycle-consistent adversarial networks”, ICCV2017.
[4]Ziwei Liu, et al, “Deepfashion: Powering robust clothes recognition and retrieval with rich annotations”, ICCV2016.