IMAGEimate - An End-to-End Pipeline to Create Realistic Animatable 3D Avatars from a Single Image Using Neural Networks

Suriya Dakshina Murthy
University of California, Santa Barbara
USA
suriya@ucsb.edu

Tobias Höllerer
University of California, Santa Barbara
USA
holl@cs.ucsb.edu

Misha Sra
University of California, Santa Barbara
USA
sra@cs.ucsb.edu

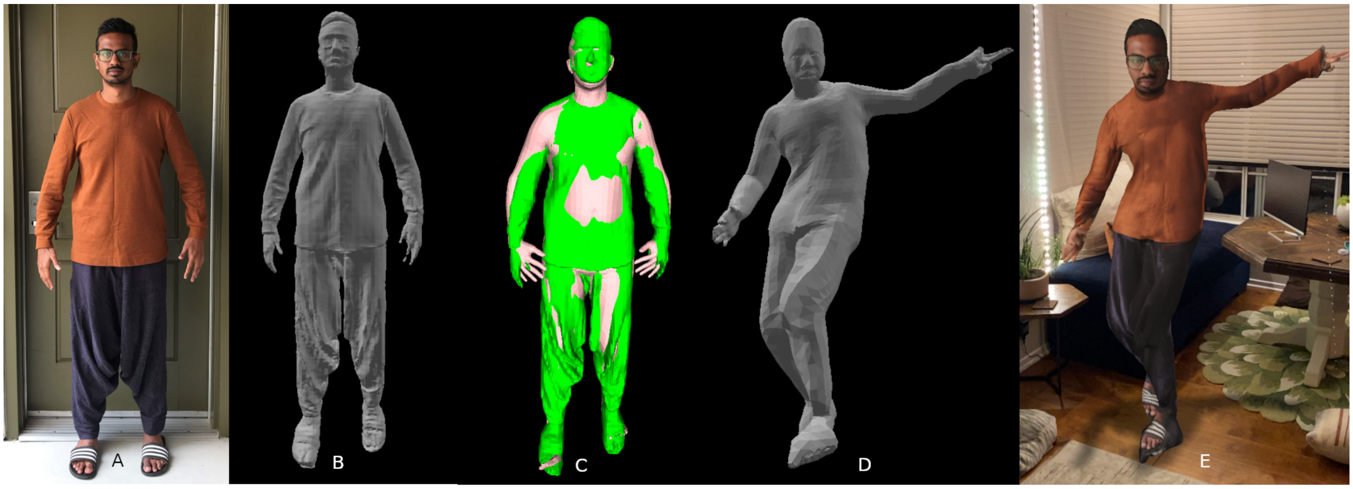

Figure 1: Results of each stage in the pipeline: A) Input image B) Predicted 3D mesh C) Parametric model fit D) Reposed mesh E) Augmented using Adobe Aero* and textured using Substance Painter†

ABSTRACT

Current advances in image based 3D human shape estimation and parametric human models enable creating realistic 3D virtual humans. We present a pipeline which takes advantage of these models and takes a single input image to create realistic 3D animatable avatars. The pipeline extracts shape and pose parameters from the input image and builds an implicit surface representation, which is then fitted onto a parametric human model. This fitted human model is animated to new and novel poses extracting pose parameters from a motion capture dataset. We extend the pipeline showcasing realism and interaction by texture painting it using Substance Painter and embedding it in an AR scene using Adobe Aero respectively.

ACM Reference Format:
Suriya Dakshina Murthy, Tobias Höllerer, and Misha Sra. 2021. IMAGEimate - An End-to-End Pipeline to Create Realistic Animatable 3D Avatars from a Single Image Using Neural Networks. In *27th ACM Symposium on Virtual Reality Software and Technology (VRST '21), December 8–10, 2021, Osaka, Japan.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3489849.3489922

1 INTRODUCTION

Creating realistic 3D avatars lies at the center of building immersive virtual environments. To create plausible human avatars using the traditional techniques of 3D modelling, character rigging and animation is time consuming and requires expertise with interactive 3D software tools. An efficient and robust alternative approach is to extract features from images and build an automated pipeline for virtual human creation. To foster such a pipeline, we identify and adapt neural-network based solutions to solve the following main sub-problems: 1) Image to mesh, 2) Mesh rigging and skinning, and 3) Re-posing mesh and applying animations.

Current pipelines such as Tex-An Mesh [2] are not automated end-to-end and require manual intervention using web-based application like Mixamo[1] in their intermediate stages. In this work, we present a fully automated pipeline, IMAGEimate, which takes an input RGB image of a human and creates a 3D animatable character.

VRST '21, December 8–10, 2021, Osaka, Japan
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9092-7/21/12.
https://doi.org/10.1145/3489849.3489922

*https://www.adobe.com/products/aero.html
†https://www.substance3d.com/
[1] https://www.mixamo.com/

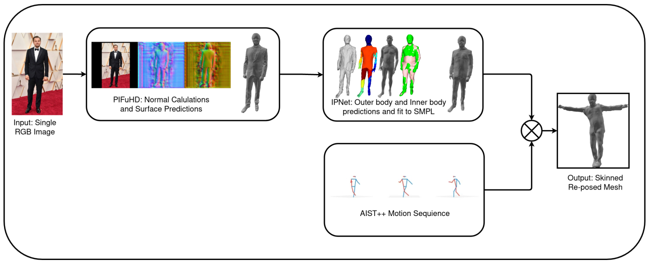
Input: Single RGB Image
PIFuHD: Normal Calulations and Surface Predictions
IPNet: Outer body and Inner body predictions and fit to SMPL
AIST++ Motion Sequence
Output: Skinned Re-posed Mesh

Figure 2: Complete System Overview

2 SYSTEM OVERVIEW

Our pipeline links together existing pre-trained deep learning networks in a scripted end-to-end Python application taking a single input image of a person as input and generating an animated articulated 3D model of that person as output.

The entire pipeline is divided into three stages - first, to extract features from an image and estimate 3D human shapes, second, to fit parametric models to the extracted shape, and finally to re-pose the parameterized model into novel poses which can be interpolated to create animations. The first stage of the pipeline takes a single RGB image and extracts pixel-aligned features which are mapped to an implicit surface representation [7]. This implicit surface represents the estimated 3D human shape from the input image. The implicit surface is then converted into a 3D mesh using the marching cubes algorithm. The goal of the system is to create animatable avatars and one of the major drawbacks of implicit functions is that they can only produce static surfaces that are not re-posable. To solve this problem, we introduce the second stage of the pipeline, which takes as input a static 3D mesh and registers a parametric human model - Skinned Multi-Person Linear Model (SMPL) along with surface deformations (SMPL+D) [1]. This registered SMPL+D mesh allows implicit reconstructions to be re-posed to novel poses. In the final stage of the pipeline, the registed SMPL+D mesh is combined with extracted motion capture pose parameters [3] to produce an animatable 3D virtual human.

3 IMPLEMENTATION

In this section we present implementation details of our pipeline. The system takes a single image as input and produces object meshes for each frame of the animation.

3.1 Image based 3D Human Shape Estimation

Extraction of human shape from images can be achieved using pose estimation based techniques [6], but they suffer from inaccurate depth prediction, especially when provided with only a single image. On the other hand, pixel aligned implicit functions can generate a high resolution mesh from a single image. PIFuHD [7] is one such approach that learns highly detailed implicit models by reprojecting 3D points into a pixel-aligned feature representation. This allows for extracting features from an image and projecting them to predict implicit surfaces. We use it as the first step in our pipeline.

3.2 Fitting Parametric Model to 3D Mesh

The generated mesh from the previous stage is a static mesh, i.e. it cannot be reshaped to new poses. To solve this problem, there are two main approaches - first, predicting the underlying skeleton and estimating skinning weights [8] and second, fitting a skinned parameterized 3D mesh. The first approach does not provide control over the skeletal hierarchy which makes it challenging to automate animation of the predicted skeleton. The second approach works well especially for human shape estimation as it is a blend shapes based skinned parametric model fitting technique. The second stage of the pipeline thus uses Implict Part Network(IPNet) [1] which jointly predicts the outer 3D surface, inner body surface under clothing and semantically matches it to Skinned multi-person linear model [4] with a displacement model for hair, garments and face details(SMPL+D).

3.3 Generating Novel Poses

The SMPL model can be reposed using two parameters - Shape parameters(θ) – 10 scalar values for width and height, and 72 Pose parameters(β) – 24x3 scalar values representing relative rotation vectors. To animate and re-pose the fitted SMPL+D model, we use open motion datasets AIST++ [3] and AMASS [5]. This involves extracting 72 joints rotation vectors for each frame and re-posing the SMPL+D model to the new poses, from which new meshes can be extracted. These meshes then get parsed to animations using the Stop Motion OBJ add-on[2] in Blender[3]. The resulting animation can be exported to desired formats like fbx, which can be directly imported into AR apps like Adobe Aero.

The entire pipeline was tested on Ubuntu 20.04 LTS with PyTorch 1.8.1. The system configuration was Intel® Core™ i7-9750H CPU@2.60GHz×12 64-Bit with 16GB RAM and Nvidia GeForce RTX 2060 Mobile with 6GB of VRAM.

4 CONCLUSION AND FUTURE WORK

In this work we presented an end-to-end pipeline which takes a single input image and creates animatable 3D humanoid avatars that can be used in AR/VR or other applications. The pipeline is modular and can be easily modified as newer techniques become available. In the future, we hope to extend it to better fitting algorithms and newer parametric models, to better highlight high frequency details like contours of the face, hair and clothing. The pipeline currently only supports body poses, but can be easily extended to facial expressions, hand gestures and clothing animations with models like SMPL-Expressive, SCALE and SCANimate respectively. We would also like to natively support textures and export the output to fbx animations for easy import into other applications.

REFERENCES

[1] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. 2020. Combining Implicit Function Learning and Parametric Models for 3D Human Reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer.

[2] Levon Khachatryan. 2020. Tex-An Mesh: Textured and animatable human body mesh reconstruction from a single image. https://github.com/lev1khachatryan/Tex-An_Mesh.

[2] https://github.com/neverhood311/Stop-motion-OBJ
[3] https://www.blender.org/

[3] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. 2021. Learn to Dance with AIST++: Music Conditioned 3D Dance Generation. arXiv:2101.08779 [cs.CV]

[4] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

[5] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *International Conference on Computer Vision*. 5442–5451.

[6] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 10975–10985.

[7] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[8] Zhan Xu, Yang Zhou, Evangelos Kalogerakis, Chris Landreth, and Karan Singh. 2020. RigNet: Neural Rigging for Articulated Characters. *ACM Trans. on Graphics* 39 (2020).