An end-to-end framework for 3D capture and human digitization with a single RGB camera

Luiz José Schirmer Silva[1], Djalma S. da Silva[1],Luiz Velho[2] and Hélio Lopes[1]

[1]PUC-Rio - Pontifícia Universidade Católica do Rio de Janeiro,
[2]IMPA - Instituto Nacional de Matemática Pura e Aplicada

Abstract

We present a low cost and accessible end-to-end framework for 3D modeling and texture capture of Humans using deep neural networks and a single RGB camera. We generate a texture atlas considering a set of multi-view images. We also capture data to generate 3D shape models and finally combine it with the generated textures to obtain a full 3D reconstruction of the human body that can be used in a game engine.

Introduction

3D motion capture, pose estimation, texture capture, and volumetric capture are important tasks to generate content for computer animation, in particular for human Digitization. Kanazawa et al. [1] show an end-to-end framework for reconstructing a full 3D mesh of a human body from a single RGB image. They used the generative human body model, SMPL[2], which parameterizes the mesh by 3D joint angles and low-dimensional linear shape space. An image is passed through a convolutional encoder and sent to a 3D regression module that infers the 3D representation of the human. This mesh can be useful to generate humanoid animations, which could immediately be used by animators.
Pandey et al. .[3] proposed a method to synthesize free-viewpoint renderings using a single RGBD camera. Besides the impressive results, it seems to be far to be applicable in real situations. Also, considering texture capture, Saito et al. [4] introduce Pixel-aligned Implicit Function (PIFu), which is an implicit representation that locally aligns pixels of 2D images with the global context of their corresponding 3D object, although their techniques seem to be computationally intensive.
Despite the independent advances in each area, there is still no proposal that uses texture capturing, pose estimation, and mesh recovery in a unified way to generate virtual characters using an accessible and low-cost architecture. In this work, we present a low cost and accessible end-to-end framework for 3D modeling and texture capture of Humans using deep neural networks and a single RGB camera. Our main contribution is an end-to-end approach to generate virtual characters based on image segmentation, pose estimation, and human mesh recovery. We apply the HMR method [1] to the captured data to generate 3D shape models and finally combine with the generated textures to obtain a full 3D reconstruction of the human body that can be used in a game engine.

Technical Details

We divided our method into two stages: texture extraction from a set of images I and the 3D modeling. For texture extraction, we use the DensePose model [5] to generate our texture atlas. We use DensePose to predict the UV coordinates of 24 body parts, and we compute a look-up table to convert the DensePose UV maps to the SMPL UV parameterization [2].
We also use partial convolutions[6] to fill small gaps in the texture since using the previous method is not possible to fill the entire UV maps. In this model, the convolution is masked and re-normalized to be conditioned on only valid pixels. We use the textures provided by the SURREAL dataset [7] in the training process, where we create patches of size 32×32 for each image, and also do data augmentation by rotating and using noise to create different masks.
Another problem is the color discontinuity between the parts of the mapped textures. To solve this problem, we subdivide the atlas following the part division of DensePose and use a method similar to presented by Junior[8]. Considering the frontier between the parts, we use a method that diffuses the color difference between the frontier zone of adjacent areas for each part. As proposed by junior et al. [8], we consider sparsely-defined texels as heat sources and solve the problem applying the diffusion equation on each heat source, which represents the flow of heat from that source. The factors between frontier edges remain fixed, and other values are relaxed across the image. In our second stage, considering the 3D model, we aim to generate the model from an initial pose. We use a factorized form of the original OpenPose model to improve performance considering the frame rate, where we use a streamlined architecture based on tensor decomposition [9]. In this initial step, with this approximation, we boosted the initial inference time by 30% concerning the original OpenPose. Subsequently, the obtained data is sent to the HMR method to infer the 3D mesh adapted from the captured person. Figure 1 show the results our approach.

Results and Conclusion

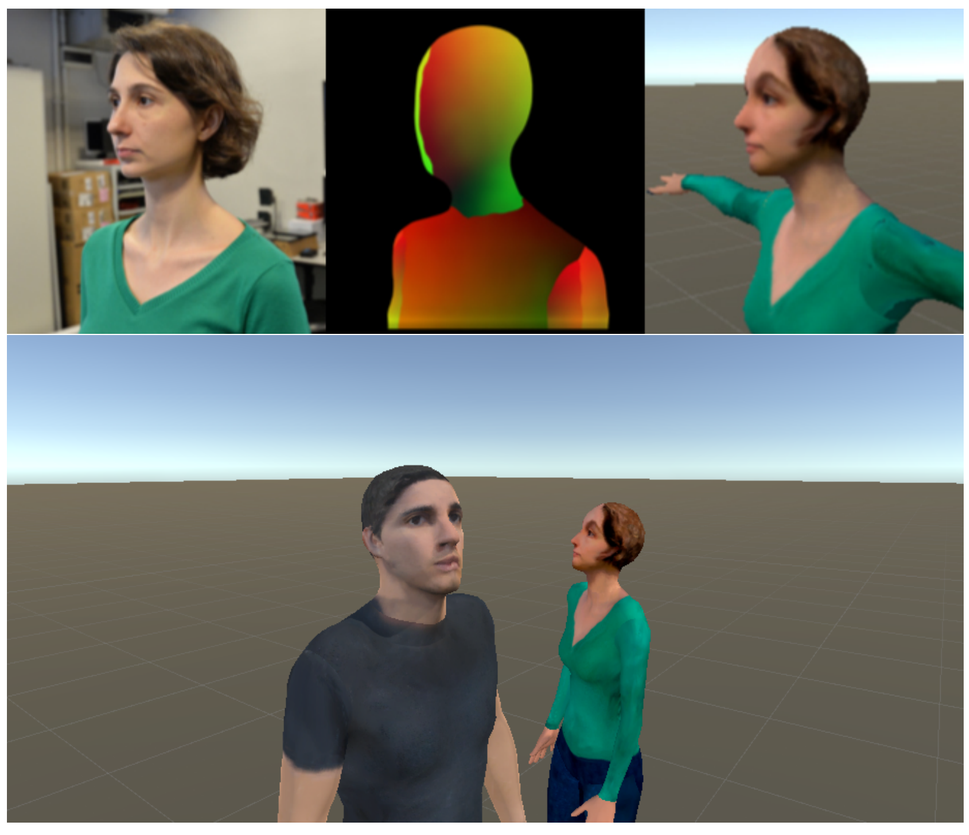

Figura 1:3D reconstructed models. Here we apply the texture captured in the first step over the mesh gerenerated by the second.

The experiments were performed in a controlled environment, with 34 photos captured for texture generation. Here we present a low cost and accessible framework that can be easily used to generated 3D human animations and other applications. As future work, we intend to create a complete model for motion and texture capture, not only depending on the HMR model and allowing the simultaneous capture of several users. Also, we intend to develop a post-processing step to solve minor errors in the transition of body parts, considering the captured texture.

References

[1] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik.
End-to-end recovery of human shape and pose.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[2] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black.
Smpl: A skinned multi-person linear model.
ACM transactions on graphics (TOG), 34(6):248, 2015.

[3] Rohit Pandey, Anastasia Tkach, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Ricardo Martin-Brualla, Andrea Tagliasacchi, George Papandreou, Philip Davidson, Cem Keskin, et al.
Volumetric capture of humans with a single rgbd camera via semi-parametric learning.
arXiv preprint arXiv:1905.12162, 2019.

[4] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li.
Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization.
arXiv preprint arXiv:1905.05172, 2019.

[5] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos.
Densepose: Dense human pose estimation in the wild.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

[6] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro.
Image inpainting for irregular holes using partial convolutions.
In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[7] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid.
Learning from synthetic humans.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2017.

[8] Jonas Sossai Junior.
Variational Texture Atlas Construction and Applications.
PhD thesis, IMPA, 2006.

[9] Luiz José Schirmer Silva, Djalma Lúcio Soares da Silva, Alberto Barbosa Raposo, Luiz Velho, and Hélio Côrtes Vieira Lopes.
Tensorpose: Real-time pose estimation for interactive applications.
Computers & Graphics, 85:1–14, 2019.