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.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Computer Vision/Texture Generation/3D modeling—Computer Animation

1. 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. [KBJM18] 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 [LMR*15], 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. When we consider volumetric capture in a studio, this is not only a costly technology but also depends on specialized hardware. Moreover, it is far from being accessible to most producers. We can find solutions that present alternative ways to reduce the cost and computational processing for this kind of application. For example, Pandey et al. . [PTY*19] 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. [SHN*19] 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.

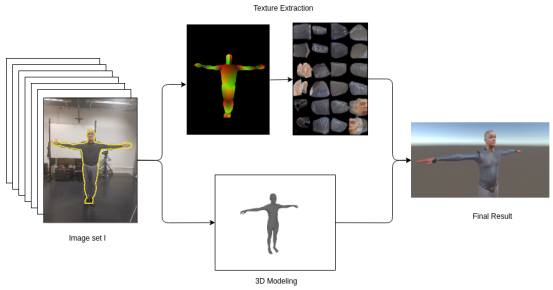

Figure 1: *Our capture pipeline. At first we segment the image, detecting the person and his position. After we generate our texture atlas using the DensePose model [AGNK18].Also we generate our 3D model based on the HMR [KBJM18].*

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 [KBJM18] 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.

2. Our Approach

We divided our method into two stages: texture extraction from a set of images I and the 3D modeling. For texture extraction, we

DOI: 10.2312/egp.20201034

use the DensePose model [AGNK18] to generate our texture atlas.They propose a variant of a Mask-RCNN, to densely regress part-specific UV coordinates within every human region at images or videos. Since we have I images with different views as input, we produce N partial atlas and after we compose them. Using this model, we can extract each pixel that relates to a specific body part of a person detected in each image. While the original article, seeks to provide a texture transfer, i.e., mapping textures previously provided to image pixels based on estimated correspondences, we aim to do the inverse. We map each pixel of an estimated coordinate into a correspondent pixel in a texture atlas. In other words, DensePose predicts 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 [LMR*15].

We also use partial convolutions [LRS*18] 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 renormalized to be conditioned on only valid pixels. We use the textures provided by the SURREAL dataset [VRM*17] 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. This discontinuity is caused by different illumination conditions while capturing the images due to the fact the pictures are taken by modifying the relative position between the camera and light source. To solve this problem, we subdivide the atlas following the part division of DensePose and use a method similar to presented by Junior [Jun06]. 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. Let r be the radius distance considering each frontier edge; for each point in the line, we calculate the color difference between corresponding texels and, after with these correction factors, perform a diffusion of them over the whole texture space. As proposed by junior et al. [Jun06], 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 [SdSR*19]. 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 2 show the results our approach.

3. Results and Conclusion

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 and a post-processing step to solve minor errors in the transition of body parts.

Figure 2: *3D reconstructed models. Here we apply the texture captured in the first step over the mesh gerenerated by the second.*

References

[AGNK18] ALP GÜLER R., NEVEROVA N., KOKKINOS I.: Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 7297–7306. 1, 2

[Jun06] JUNIOR J. S.: *Variational Texture Atlas Construction and Applications*. PhD thesis, IMPA, 2006. 2

[KBJM18] KANAZAWA A., BLACK M. J., JACOBS D. W., MALIK J.: End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 7122–7131. 1

[LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG) 34*, 6 (2015), 248. 1, 2

[LRS*18] LIU G., REDA F. A., SHIH K. J., WANG T.-C., TAO A., CATANZARO B.: Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 85–100. 2

[PTY*19] PANDEY R., TKACH A., YANG S., PIDLYPENSKYI P., TAYLOR J., MARTIN-BRUALLA R., TAGLIASACCHI A., PAPANDREOU G., DAVIDSON P., KESKIN C., ET AL.: Volumetric capture of humans with a single rgbd camera via semi-parametric learning. *arXiv preprint arXiv:1905.12162* (2019). 1

[SdSR*19] SILVA L. J. S., DA SILVA D. L. S., RAPOSO A. B., VELHO L., LOPES H. C. V.: Tensorpose: Real-time pose estimation for interactive applications. *Computers & Graphics 85* (2019), 1–14. 2

[SHN*19] SAITO S., HUANG Z., NATSUME R., MORISHIMA S., KANAZAWA A., LI H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172* (2019). 1

[VRM*17] VAROL G., ROMERO J., MARTIN X., MAHMOOD N., BLACK M. J., LAPTEV I., SCHMID C.: Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 109–117. 2