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.

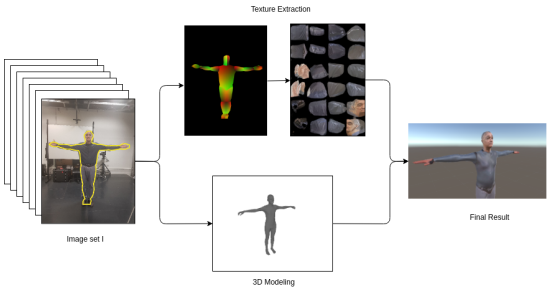

Figure 1: *Our capture pipeline. At first we segment the image, detecting the person and his position. After we generate our texture atlas using the DensePose model [AGNK18].Also we generate our 3D model based on the HMR [KBJM18].*

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