第一组：王一林, 侯哲宇, 胡太穆, HOKHY TANN

Video Based Reconstruction of 3D People Models Paper

Summarization and Discussion on comprehensively difficult concepts and methods

Written by HOKHY TANN

1. Introduction

A personalized realistic and animatable 3D model of a human is required for many applications, including virtual and augmented reality, human tracking for surveillance, gaming, or biometrics.

In this work, we estimate the shape of people in clothing from a single video in which the person moves. The key idea of our approach is to generalize visual hull methods to monocular videos of people in motion. To make visual hulls work for monocular video of a moving person it is necessary to “undo” the human motion and bring it to a canonical frame of reference.

To understand the concept mentioned above, we need to know what visual hull technique. A visual hull is a geometric entity created by shape-from-silhouette 3D reconstruction technique introduced by A. Laurentini. This technique assumes the foreground object in an image can be separated from the background.

2. Method

Given a single monocular RGB video depicting a moving person, our goal is to generate a personalized 3D model of the subject, which consists of the shape of body, hair and clothing, a personalized texture map, and an underlying skeleton rigged to the surface.

The method proposed by the paper consists of 3 steps: 1) pose reconstruction 2) consensus shape estimation and 3) frame refinement and texture map generation. As mentioned in the paper, their main contribution is in step 2), the consensus shape estimation.

2.1 SMPL Body Model with Offsets

SMPL is a parameterized model of naked humans that takes 72 poses and 10 shape parameters and returns a triangulated mesh with $N = 6890$ vertices. It takes quite a few parameters like, shape β and pose θ deformations. And then, they add these parameters up with the training scan $T\mu$ to create the basic template.

However, the Principal Component shape space of SMPL was learned from scans of naked humans, so clothing and other personal surface details cannot be modeled. In order to personalize the SMPL model, we simply add a set of auxiliary variables or offsets $D \in R\ 3N$ from the template:

$$T(\theta, \beta, D) = T\mu + Bs(\beta) + Bp(\theta) + D$$

2.2 Pose Reconstruction

In the first step, they primarily use SMPL to create the basic poses with some optimizations like, trying to optimize P value, setting constrains if the height of the person is known, using 4 different levels of a Gaussian Pyramid G, using state of the art of 2D joint detection, enforcing a temporal smoothness and initializing the pose in a new frame with the estimated pose θ in the previous frame. However, when optimizing over $P = 5$ poses the scale ambiguity prevails.

2.3 Consensus Shape

This part is the most difficult to understand for me since they used many different techniques to project all the poses to create 3D model. As much as I could summarize, below are the parts I find new and hard to understand:

1. Plucker coordinates
2. Jacobians
3. Laplacian mesh regularizer

After doing research on the subjects above, I would like to briefly talk about each of them as below:

1. Plucker coordinates are a way to assign six homogeneous coordinates to each line in projective 3-space, $\mathbf{P}^3$. Because they satisfy a quadratic constraint, they establish a one-to-one correspondence between the 4-dimensional space of lines in $\mathbf{P}^3$ and points on a quadric in $\mathbf{P}^5$(projective 5-space).
2. In vector calculus, the Jacobian matrix is the matrix of all first-order partial derivatives of a vector-valued function. When the matrix is a square matrix, both the matrix and its determinant are referred to as the Jacobian in literature.

2.4 Frame Refinement and Texture Generation

The optimization is initialized with the preceding frame and regularized with neighboring frames

To create the texture, the paper warps their estimated canonical model back to each frame, back-project the image color to all visible vertices, and finally generate a texture image by calculating the median of the most orthogonal texels from all views.