Image (video) using deep learning technology, advanced research/ development of body recognition technology

Anh-Quang Duong, Artificical Intelligence Convergence

주식회사 지니소프트

Code: https://drive.google.com/file/d/1jCyxozHYeUsCLeZZwWi3v8rTFPVKwaa0/view?usp=sharing

A group of people dancing

Description automatically generatedA group of people dancing

Description automatically generated with medium confidence

*Abstract*—3D Human mesh reconstruction from video, which is useful for many applications. It can be used for creating animated films, motion graphics, interactive 3D application, computer game. In this project, I present a web application for 3D human mesh reconstruction from video, in which, I used the current state of the art deep learning techniques as well as web development tools to caried my ideas

Keywords—3D human mesh reconstruction, deep learning, web application

# Introduction

Human motion is fundamental to understanding behavior. With the application of 3D mesh reconstruction, we can reduce a lot of time for building human motion, the work, in which we have to build the skeleton and also the position of skeleton for each frame for the making of motion. In the application, we use the VIBE model (Video Inference for Body Pose and Shape Estimation) as a currently video-based state-of-the-art method, to render 3D mesh reconstruction from video. The application is a web-based application, in which, I used flask as a framework for building web server, mysql as a data storage and celery for asynchronous task

# Proposed application

## 3D human mesh reconstruction for multi-person in video:

* First extract frame in video
* Run Multi Person Tracker.
* Run pre-trained models on each extracted person.

The number of frames in video will be determined based on framerate of that video, if framerate is high then the number of extracted frame is high also, and so on. After that, I used Yolov3 for multi-person tracker, to extract each frame of each person in the video. Then, the Vibe [1] will be used to detect 3D mesh reconstruction for each people in the video. The overall architechture for 3D human mesh reconstruction for multipersion is shown in Figure 1.

A picture containing text

Description automatically generated

Fig. 1: 3D human mesh reconstruction for multi-person

## VIBE (Video Inference for Body Pose and Shape Esimation)

VIBE estimate SMPL body hyperparameter for each frame in a video sequence using a temporal generation network, which is trained together with a motion discriminator. The motion discriminator implicitly learns to account for the statics, physics of the human body in motion using the ground-truth motion-capture (mocap) data. The Vibe model architechture is summarized in Figure 2.

Parametric 3D human body models are widely used as the output target for human pose estimation because they capture the statistics of human shape and provide a 3D mesh that can be used for many tasks. SMPL represemt the body pose and shape, which consist of the pose and shape parameter and respectively. The pose parameters include the global body rotation and the relative rotation of 23 joints in axis-angle format, the shape parameters are the first 10 coefficients of a PCA shape space.

Diagram

Description automatically generated with medium confidence

Fig. 2: Vibe Model

## Web Application based on Flask Framework

Flask is a framework for building Web Server using python, it is easy to develop and operation. Also, both Website and Mobile Application can communicate with Web Server through HTTP protocol. But there is 1 problem with webserver is that video processing take time and effort, we can not process in HTTP request (time between request and response is limited). The solution for this challenge is asynchronous task, by using asynchronous, we don’t have to wait for server to do its jobs, the server will send this task to a queue and return the status to client. Now, the queue has to process the task, when the task is finish, it will update the status in database and trigger the client to get the result of that task. To do that, I use celery, which is simple, flexible and reliable distributed system to process vast amounts of messages.

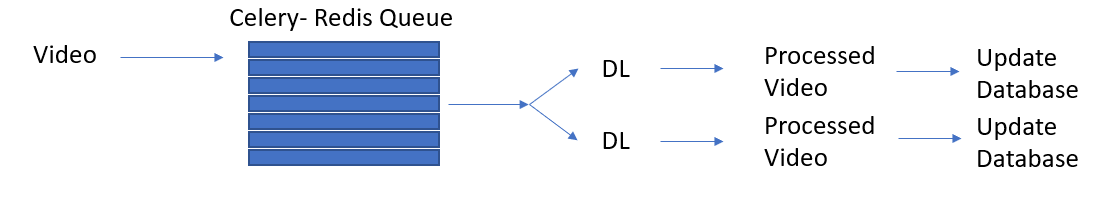


Fig. 3: Celery task distribution.

For storage, I used mysql as a database, there are 2 main table, the first is user and the second is video. Redis is used as a queue for celery. The frond-end is made by html, javascript and jquery. The web application will have some features like log-in, log-out, register, upload video.

# Result & Discussion

For the Vibe model, the pretrained weight is achieved by training on 3DPW dataset with score is 56.5, the evaluation metric is Procrusters-aligned mean per joint position error (PA-MPJPE).

For the web application, figure 4,5,6 show a few examples of the applications.

A picture containing text, indoor, person

Description automatically generated

Fig. 4: Log-in

A picture containing text, indoor

Description automatically generated

Fig. 5: Register

Graphical user interface, application

Description automatically generated

Fig. 6: Homepage

# conclusions

3D human mesh reconstruction is useful for many applications. It can be use in programs such as blender to render human motion, from that, reduce time to process instead of building from scratch. In this project, I combine many techniques to build a complete application, including back-end, frond-end and also deep learning techniques.

##### References

1. Kocabas, Muhammed, Nikos Athanasiou, and Michael J. Black. "Vibe: Video inference for human body pose and shape estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.