ECE445/ME470 Project Proposal

A Deep Learning Based Paradigm in 3D Human Pose Detection and Estimation in Multi-View Videos

Team members: Fengkai Chen

Feiyu Zhang

Zhuoting Han

Han Zheng

Project Advisor: Gaoang Wang

Course Instructor: Mark Butala

1.Introduction

1.1 Objective

The main objective of this project is to propose a 3D Pose Estimation paradigm which solves the noise-sensitive problem in 3D pose capture and reconstruction via leveraging machine learning and optimization technique. Using the video captured by multi-view camera system, we will apply multi-way matching algorithm to cluster the detected 2D poses in all views, and reconstruct the 3D pose of each person from the corresponding bounding boxes and associated 2D poses. Our fixed cameras generate real-time videos of 24 fps, and the expected 3D Pose Estimation paradigm can process faster than 20 fps. Assuming that we only need to work on the key frames (frame which contains informative features compared to adjacent frames), the expected running time can successfully support real-time 3D Pose Estimation. The overall process of this project consists of deep neural network design and architecture validation, optimization algorithm formulation and implementation, and real-time experiment and demonstration.

1.2 Background

Recovering 3D human pose and motion from videos has been a long-standing problem in computer vision, which has a variety of applications such as human-computer interaction, video surveillance and sports broadcasting [2]. In recent decades, animation and science fiction films have come to dominate the film market. One essential procedure in science fiction films production is 3D motion capture and reconstruction of actor movements.

Past 3D motion capture and reconstruction paradigms are usually noise-sensitive and fail to recognize and reconstruct human pose in multiple-demonstrator scenario due to the overlap of features from different human bodies, which is the obstacle we would like to tackle.

1.3 Physical Diagram

Camera System

Raw images shot by camera system (multi-views)

PC

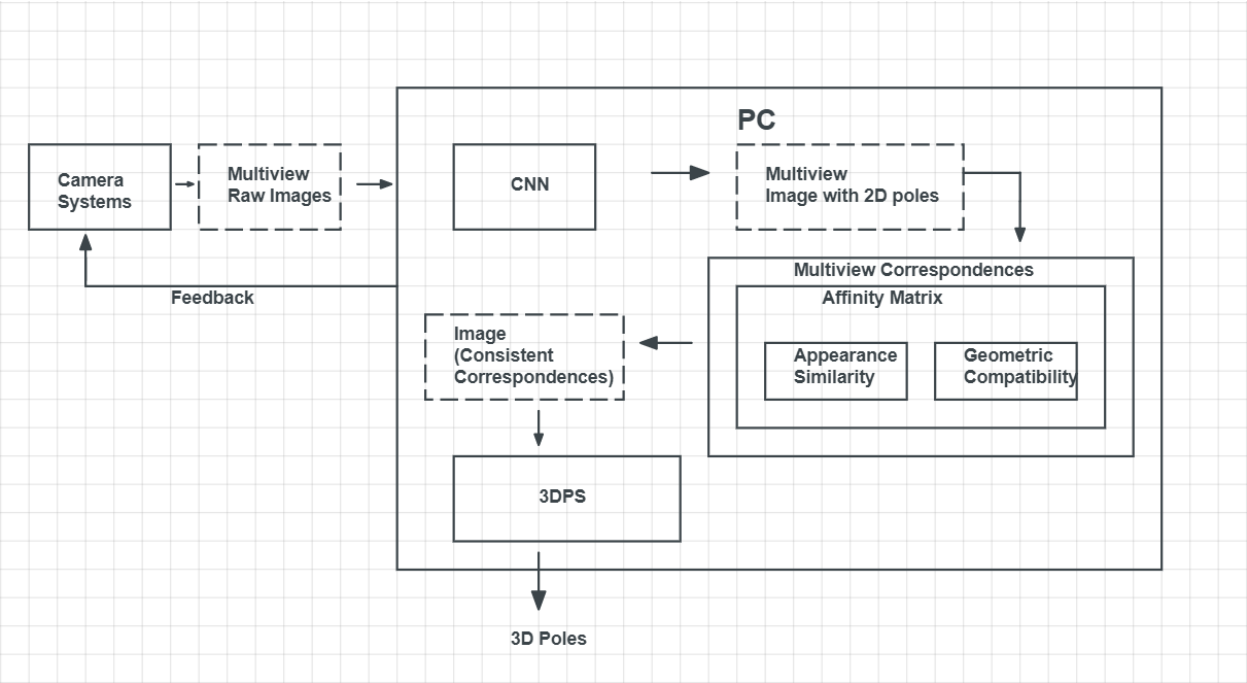
3D Pose estimation

1.4 High-level Requirements List

- As we found in the paper 3D human pose estimation in video with temporal convolutions and semi-supervised training [5], we can judge by how low the reconstruction error (MPJPE) is. As in the paper mentioned above, the MPJPE is as low as 46.8 (the lower the better) on the dataset Human3.6M. Considering we will implement a model that can detect videos instead of photos, the error will be higher. We plan to have similar reconstruction error on the promising datasets like Human3.6M and TotalCapture. We expect a reconstruction error to be 48(±5).
- The whole system, consisting of video capture, 2D pose detection, multi-view correspondence and 3D post reconstruction, should be able to process real time captured videos at runtime, and process the locally stored data.
- We intend to propose a solution with self-explanatory question modeling accompanied with solid ground of mathematics. The solution should be mathematically rigorously defined.

2.Design

- 2.1 Block Diagram



- 2.2 Functional Overview

2.2.1 Camara system

The camera system contains several cameras with different positions and projection angles. The camera system provides us a multi-frame real-time video of human movement, which will be used as raw data for subsequent software processing.

2.2.2 PC

The PC is our main operation platform for algorithm testing and demonstration. Key joints from human body in separate 2D views will be detected and aggregated in the following stages to reconstruct 3D poses.

- 2.3 Block Requirements

2.3.1 CNN

Convolutional Neural Networks - Pick the CNN model with least test error (Cascaded Pyramid Network is preferred). Feed raw images/videos captured by the camera system to the network and generate output image with 2D pose attached.

2.3.2 Multi-View Correspondences

Match the detected 2D poses across views, i.e. find the 2D bounding boxes belonging to the same person in all views. We will use an affinity metric to measure the likelihood between two 2D bounding boxes which belong to the same person, and a matching algorithm will be developed to establish the correspondences of bounding boxes across multiple views.

2.3.2.1 Affinity Matrix

A) Appearance Similarity – Feed bounding box pairs to the CNN and compute the Euclidean distance between two feature vectors from the output layer. Then normalize this distance using sigmoid function to range (0,1) as the appearance similarity score.

B) Geometric Compatibility – Measured from the point-to-line distance between pose joints in one view to the epipolar line associated with these joints in another view. Then normalize this distance using sigmoid function to range (0,1) as the appearance geometric compatibility score.

2.3.3 3DPS

3DPS is used to match 2D pictures with 3D position. A traditional 3DPS algorithm will take a long time to search in a mega space. However, we use epipolar association to reduce the search time to be linear, which significantly improves the matching efficiency and precision.

- 2.4 Risk Analysis

2.4.1 Fail to complete Real-time 3D pose reconstruction

The real-time videos we used as input are 24fps. Reconstructing 3D poses require heavy calculation, if the processing speed of our model is lower than 24 frame per second, our model won't be able to support real-time 3Dpose reconstruction.

2.4.2 Risk in network transmission

The network transmission could be a possible risk. Our system relies on fast and robust transmission of data at real time. If the data packet has a high drop rate, our network might be in congestion, which causes the whole system to be stuck. And once many pictures/videos are lost, the predicted result will also have a higher reconstruction error. Both aspects could lead to severe negative influence on our result.

3. Ethics & Safety

Our project has several potential safety and ethics concerns. The first concern is network intrusion. Currently we are using campus network to transmit our information and signals. However, every network has a possibility to be attacked, and this rule also applies to our campus network. This is against 7 and #9 of the IEEE Code of Ethics – "the people committing piracy are not properly crediting the work of others, and they could be injuring the copyright holders by sharing content without paying for it." [4] Once the network is controlled, we may lose our control over the whole system, such that our core codes and algorithms may leak. Actually, we do not have a perfect plan for this. Our current solution is that use version control tools, like SVN and git, to store our codes and do not publish it before some sense of agreement is made.

The second concern is the private pictures/video disclosure. The disclosure violates the ACM code of Ethics, #1.6, "Therefore, a computing professional should become conversant in the various definitions and forms of privacy and should understand the rights and responsibilities associated with the collection and use of personal information." [5] Due to the high volume of picture/videos used for network training, saving all data in our personal laptop is not recommended. For convenience in calling data, we plan to store our data on an online server, which may be cyber-attacked and cause data disclosure. To minimize such risk, we suggest shutting down network acceleration software such as Cisco AnyConnect Mobility Client and Express VPN when testing online algorithms.

With the following concerns are fully considered, we still want to make sure that the model will treat everyone equally. If we use a biased training dataset, like some dataset mostly containing videos/pictures of white people, the model may have worse effects on black, Asian and Hispanic people. If we use a training dataset that mostly involves men moving and acting, this model may have worse effects on women. All these violate the #8 of the IEEE Code of Ethics, "to treat fairly all persons and to not engage in acts of discrimination based on race, religion, gender, disability, age, national origin, sexual orientation, gender identity, or gender expression" [4]. To avoid such things, we will carefully choose our dataset, including the percentage of different races, genders, ages and other tags that may divide people into different groups, to ensure an unbiased development process.

References

1. openaccess.org, “.Fast and robust multi-person 3d pose estimation from multiple views”, 2020. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/papers/Dong_Fast_and_Robust_Multi-Person_3D_Pose_Estimation_From_Multiple_Views_CVPR_2019_paper.pdf [Accessed: 28- Feb- 2021].7

2. CVSSP Research. Multi-Person 3D Pose Estimation and Tracking in Sports. Apr.17, 2019. [Online Video]. Available: https://www.youtube.com/watch?v=jLEv14GAcbs
3. eee.org, "IEEE IEEE Code of Ethics", 2020. [Online]. Available: http://www.ieee.org/about/corporate/governance/p7-8.html. [Accessed: 28- Feb- 2021].7
4. ethics.acm.org, “ACM Code of Ethics”, 2020. [Online]. Available: https://ethics.acm.org. [Accessed: 28- Feb- 2021].7
5. openaccess.org, “3D human pose estimation in video with temporal convolutions and semi-supervised training”, 2020. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/papers/Pavllo_3D_Human_Pose_Estimation_in_Video_With_Temporal_Convolutions_and_CVPR_2019_paper.pdf [Accessed: 28- Feb- 2021].7
6. Umar.Iqbal, Pavlo.Molchanov, Jan.Kautz, NVIDIA Research.” Weakly-Supervised 3D Human Pose Learning via Multi-view Images in the Wild” In CVPR 2020 https://research.nvidia.com/sites/default/files/publications/Iqbal_Weakly-Supervised_3D_Human_Pose_Learning_via_Multi-View_Images_in_the_CVPR_2020_paper.pdf