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

Group 14

Group Member:
Fengkai Chen
Feiyu Zhang
Han Zheng
Zhuoting Han

1.Introduction

Human Pose estimation and reconstruction is a widely researched topic in the recent decades. Its main idea is detecting location of people's joints which form a skeleton, and to estimate the posture and movement of human body. Estimating the pose of a human in 3D given an image or a video has recently received significant attention from the scientific community. The main reasons for this trend are the ever increasing new range of applications (e.g., human-robot interaction, gaming, sports performance analysis) which are driven by current technological advances [1].

Although recent approaches have reported remarkable results in 3D pose estimation from static images, it remains an unsolved problem in continues-time videos. This is because the time-varying overlaps of human bodies in consecutive video frames imposes several challenges in detecting the joints from human bodies, which are not fully addressed by existing methods. The objective of this project is to propose a 3D Pose Estimation paradigm for video setting via leveraging machine learning and optimization technique. Overall speaking, we will first use a deep Convolutional Neural Network (CNN) to detect the human body (pose) from the surroundings in video clips captured by our multi-view camera system. The detected poses are indicated by a group of boxes (bounding boxes). Then we apply multi-way matching algorithm to cluster the detected 2D poses in the resulting bounding boxes, and reconstruct the 3D pose associated to each person.

The multi-way matching algorithm aims at finding 2D poses of the same person in a group of videos clips captured by several cameras (our camera system). For example, there are five people (labeled in 1 to 5) in the room and we have three cameras shooting from different directions. Then the multi-way matching algorithm matches the 2D pose of person 1 in video from camera 1 to his 2D poses in videos from camera 2 and camera 3. The matched 2D poses of person 1 are categorized by bounding boxes of a specific color (a color corresponds to a labeled person), and 2D poses inside the bounding boxes with same color will be cut-out from the video for 3D pose reconstruction. The 2D to 3D pose reconstruction is done by some well-developed approach such as 3D pictorial structure (3DPS) based model [2].

Our fixed cameras will generate real-time videos of 24 fps, and the expected 3D Pose Estimation paradigm can process faster than 20 fps. Assume 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.

2 High-level Requirements

- We expect a reconstruction error (MPJPE) to be 48($\pm$5).
- The whole system should be able to process the locally stored video clips.
- We expect the system to process at the speed of 2s per frame.

3 Block Diagram

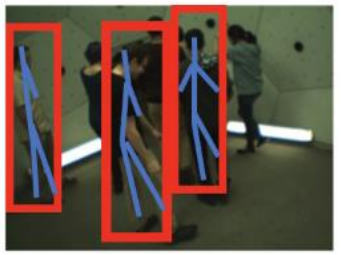
Human motion
in a room setting

Camera System

Multiview shooting

Videos by camera
system (multi-views)

Feature extraction by
Neural Network (NN)

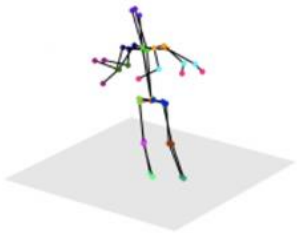


2D Poses indicated
in bounding boxes

Multi-view
matching algorithm

Categorized poses
matched with
corresponding
person

3DPS model



4 Requirements & Verification Tables

Camera Input System

Requirements	Verifications
1. Provide 24±2 fps frames video data. 2. Work under 11.5V-15.5V DC voltage supply 3. Maintain normal working status between 0℃ to 40℃	1A.Measure the real time output data using a laptop, ensuring that the transmitted video are between 22 and 26 fps. 2A.Connect the camera system to a 11.5V DC power supply, measured by a voltage meter. 2B. Measure the real time output data using a laptop, ensuring that the transmitted video are between 23 and 25 fps. 2C. Repeat above process while adjusting the power supply until 15.5V DC. 3A. While verification for Requirement 1 and 2, use a thermometer to ensure that the temperature of working space is between 0℃ and 40℃.

5 Safety & Ethics

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.