Monocular Human Motion Estimation

James Waguespack, Linköping University
Computer Vision Lab, Part pf ISY at LiU
Supervisors: Bastian Wandt, Michael Felsberg (co-supervisor)

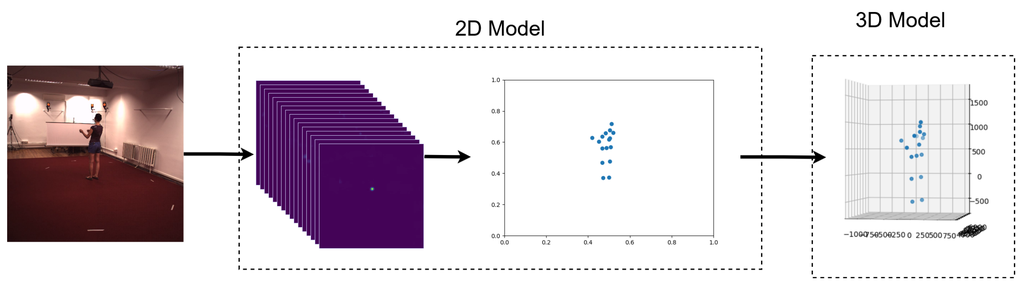
LINKÖPINGS UNIVERSITET

Motivation & Research Goals

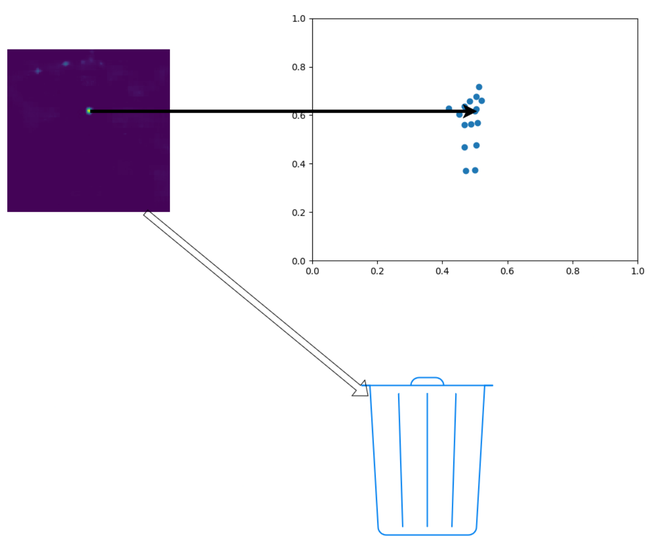
Human motion estimation has the potential to greatly improve fields such as healthcare, sports analysis, and motion-driven animation. While multi-camera systems — sometimes combined with purpose-built outerwear — offer high accuracy, they require complex, cost-prohibitive setups, making them impractical for everyday use. This research aims to advance monocular human motion estimation, improving its robustness and accessibility. The goal is to create a solution that provides reliable, accurate motion estimation while remaining affordable and easily deployable for a wide range of applications.

Intro & Method

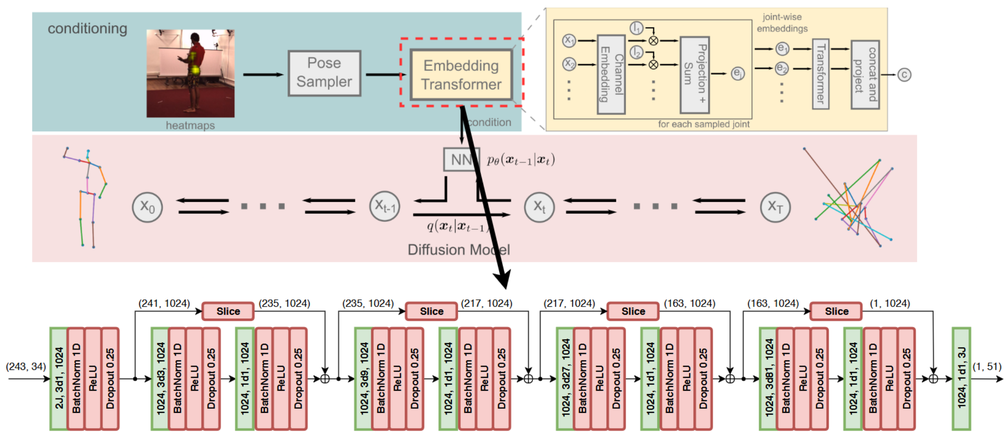
The current research focuses on the problem of occlusion. When joints are occluded, they are more difficult to predict accurately. We are attempting to address this by propagating probabilistic information further through the system.
Traditionally, HME systems are comprised of 2 models: a 2D joint estimator and a 3D inflation model. The 2D estimator will generate a probability map for the position of each joint, but only the position with highest likelihood is chosen for the 2D estimation. This estimation is then given as input to a 3D inflation model, which attempts to add a 3rd dimension to the estimation.



Because the 2D positions are naively predicted by the argmax of the probability map, the probability information is lost entirely and the 3D inflation model never has access to it.

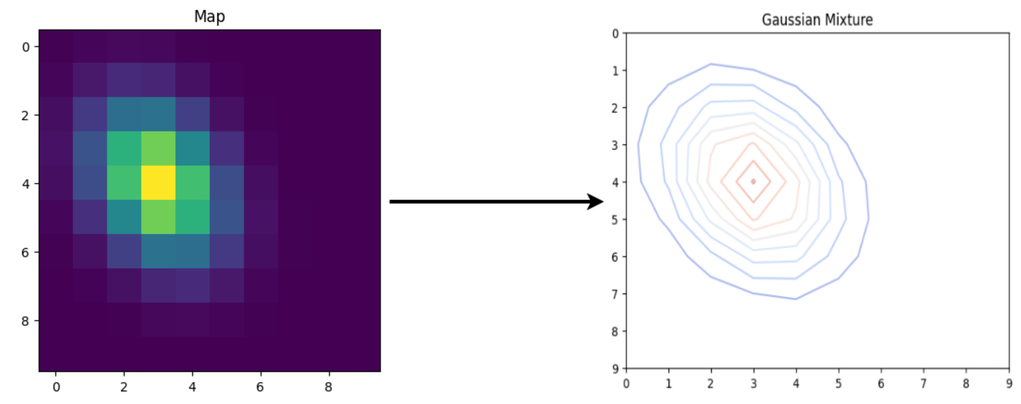


To avoid this loss of information, we employ the embedding transformer from Diffpose[1]. Many samples are taken from the heatmap for an accurate representation, and these samples are used as input to the transformer, which can embed the distribution information. The embeddings are then used as input for VideoPose3D. Video pose takes information from 243 frames to predict the pose of the middle frame.



Current Issues

Originally, DiffPose used cropped portions of the image to predict local pose information. However, our wish is to predict global positions. This means that we cannot crop the input images. The embedding transformer can handle any image size, so that is not an issue. However, without cropping, the size of the probability maps increases from 64x64 pixels to 254x254 pixels. The probability maps were originally over 30GB, and the lack of cropping makes the new dataset prohibitively large to store in its raw form. To address this problem, we are implementing a Gaussian mixture to capture the information of the probability distribution while drastically reducing the storage size.



Each component of the mixture only requires storage of 7 floating point numbers: the mean (2), covariance matrix (4), and the weight (1).

Future Exploration

In the future, we would like to add more features to increase the system's robustness.

1. Implement this method for multiple established HME systems. This will allow for robust comparison to see if this method improves accuracy.
2. Add functionality for a moving camera. This will likely require a significant overhaul of the current system but, once functioning, will greatly increase the applicable use cases. We will begin by exploring optical flow systems to see if they can be incorporated into the current pipeline
3. Account for the possibility of multiple people in frame

References

[1] DiffPose: Multi-hypothesis Human Pose Estimation using Diffusion models
Karl Holmquist and Bastian Wandt, ICCV, 2023

[1] 3D human pose estimation in video with temporal convolutions and semi-supervised training
Dario Pavllo and Christoph Feichtenhofer and David Grangier and Michael Auli, CVPR, 2018

[1] 3D Human Pose Estimation with Spatial and Temporal Transformers
DCe Zheng and Sijie Zhu and Matias Mendieta and Taojiannan Yang and Chen Chen and Zhengming Ding, ICCV, 2021

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM