2D-to-3D Pose Estimation in Real-Time Motion Capture

Jaee Oh

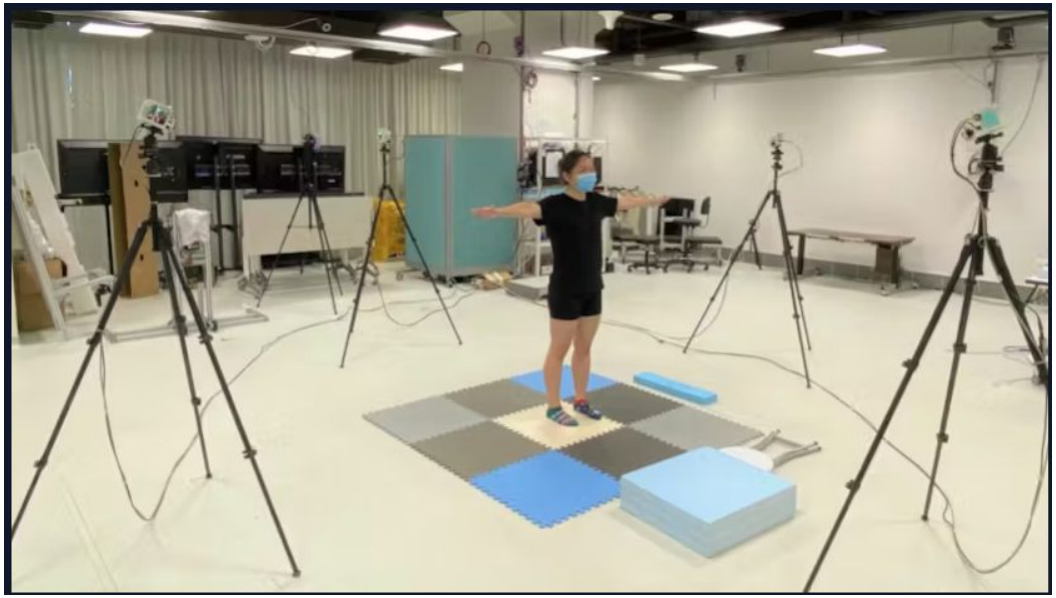
Motion Capture

Image from https://axisxr.gg/common-problems-in-motion-capture/

Traditional MoCap

- Require markers and cameras.
- Used for recorded images/videos (movie, game, etc.)
- Optical, Inertial, …

Motion Capture - Markerless



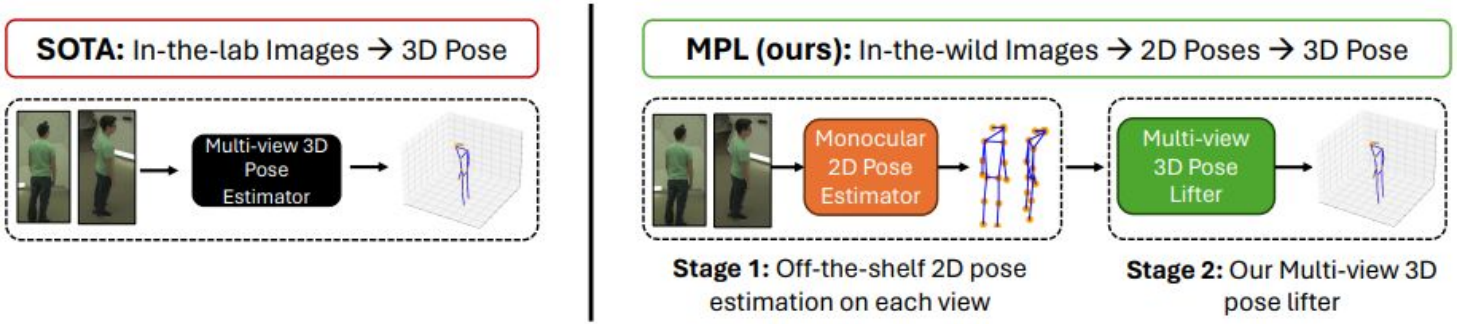
Markerless MoCap

- Rise with the development of machine learning
- No need for markers
- Health care, VR/MR, biomedical research, …

Image from https://www.hackster.io/news/precise-marker-less-motion-capture-system-can-track-patients-movements-to-0-4-accuracy-f9431b96b256

Markerless MoCap Process

- 2D Pose Estimator: Full image -> 2D Positional Data for Joints
- 3D Pose Lifter: 2D Positional Data for Joints -> 3D Positional Data for Joints
- Better result than triangulation.

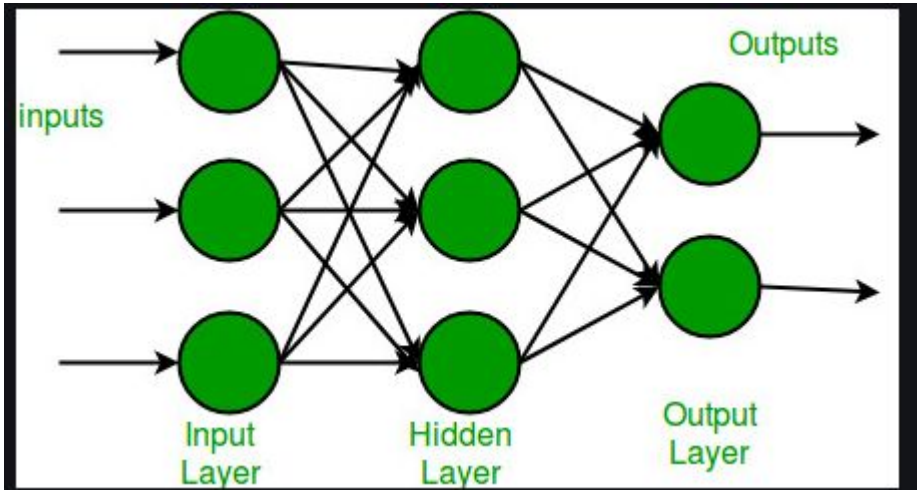


(Seyed Abolfazl Ghasemzadeh, Alexandre Alahi, and Christophe De Vleeschouwer, 2024)

2D-to-3D Pose Estimation with One Webcam

- Use MediaPipe as a 2D Pose estimator/lifter
- 2D-to-3D Pose Estimator: Regressor
- Models tested:
  - Linear Regression, Lasso, Ridge
  - KNN
  - Random Forest
  - XGBoosting
  - MLP (Multi Layer Perceptron)

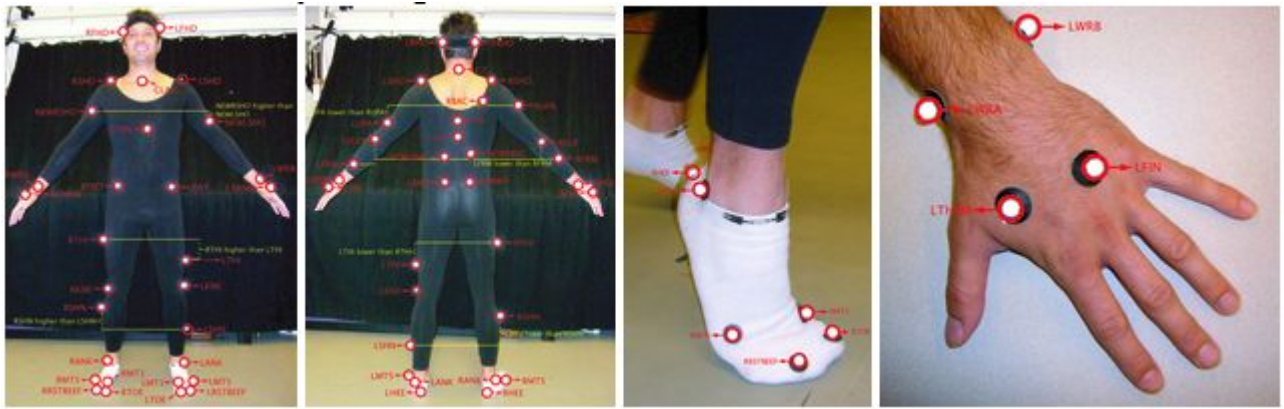
MLP(Multi Layer Perceptron)



- Input layers -> Hidden layers -> Output layers
- Each neuron has
  - Weighted sum
  - Activation function
- Back propagation for minimizing loss function

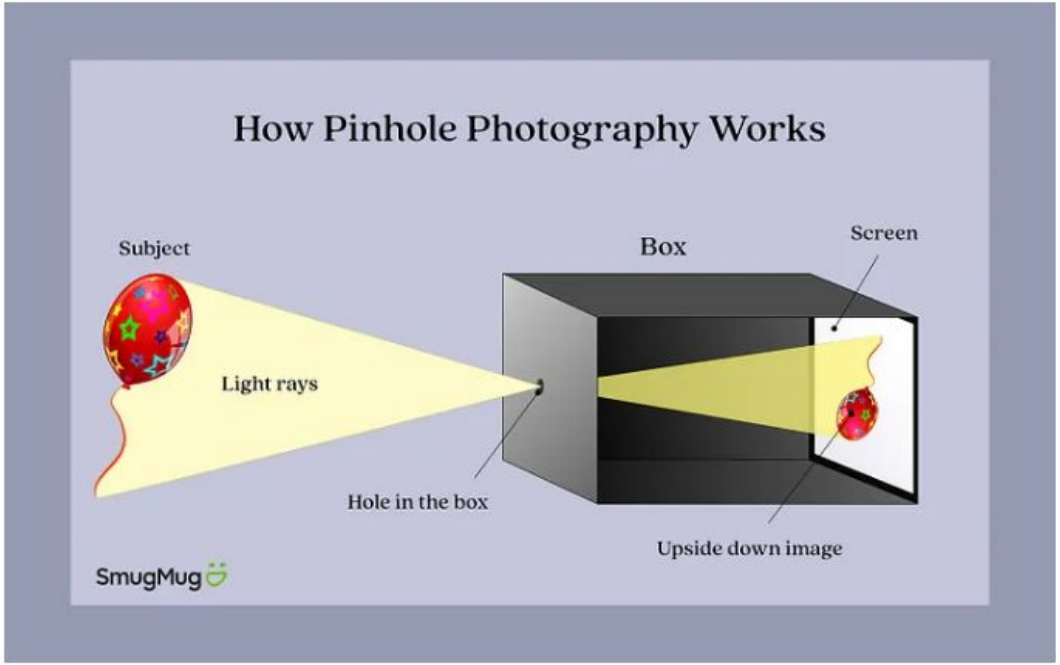
Image from https://www.geeksforgeeks.org/deep-learning/multi-layer-perceptron-learning-in-tensorflow/

CMU MoCap Data



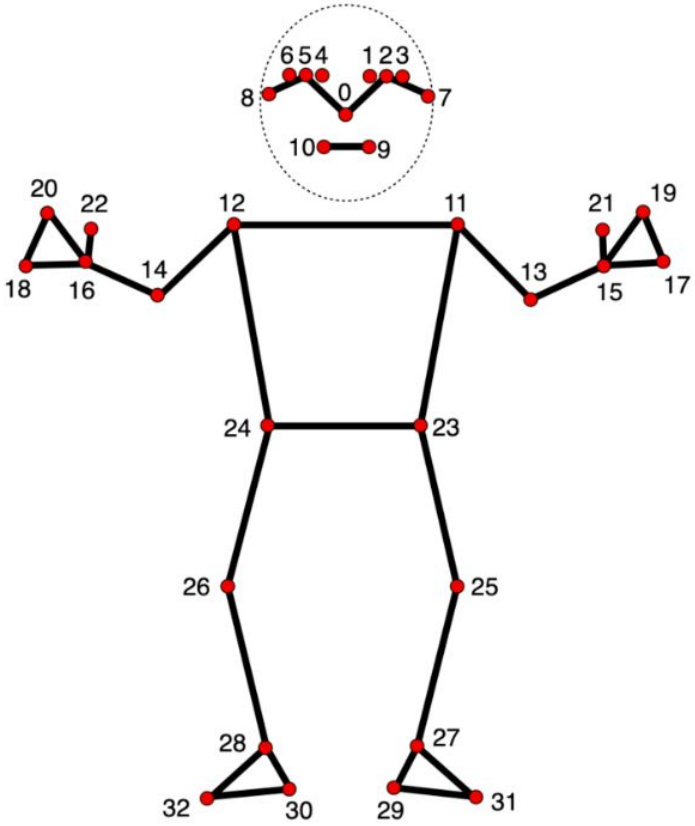
- 31 Joint Data from 41 Markers
- 3D positions + 3D rotations

Generate 2D Input Data Using CMU Mocap 3D Data



- Simple pinhole camera model to generate 2D image from a 3D positional data.
- x_2d = x_3d * f / z_3d
- Normalized to
  - X = 0~1
  - Y = 0~1
- Front view was taken.

MediaPipe Landmarks



- Reduced to 19 Joints.
- Training 19*2 features -> 31*3 features

Result with ~50k datapoints after Gridsearch

- GridsearchCV(5-fold), 0.2 Test-Train Split
- Linear Regression
  - Test R^2 = 0.8615
- Ridge (alpha = 1e-7)
  - Test R^2 = 0.8615
- Lasso (alpha = 0.005)
  - Test R^2 = 0.7520

Result with ~50k datapoints

- KNN
  - Test MSE = 1.1356e-5
- Random Forest
  - Test MSE = 0.007186
- XGBoosting
  - Test MSE = 0.0004715
- MLP ((512, 256, 128), alpha=1e-6, learning_rate = 0.001)
  - Test MSE = 0.004629

Result with ~5M datapoints

- KNN
  - Test MSE = 1.093e-5
- Random Forest
  - Test MSE = 0.0554
- XGBoosting
  - Test MSE = 0.01302
- MLP ((512, 256, 128), alpha=1e-6, learning_rate = 0.001)
  - Test MSE = 0.0301

Result with ~5M datapoints Tested on 10k data

- KNN
  - Test MSE = 1.963e-6
  - Prediction Time = 927.5s -> ~10s latency per 1s (for 100fps)
- Random Forest
  - Test MSE = 0.1837
  - Prediction Time = 0.8513s -> ~0.01s latency per 1s (for 100fps)
- XGBoosting
  - Test MSE = 0.0343
  - Prediction Time = 1.703s -> ~0.02s latency per 1s (for 100fps)
- MLP ((512, 256, 128), alpha=1e-6, learning_rate = 0.001)
  - Test MSE = 0.0808
  - Prediction Time = 0.0476s -> ~0.0005s latency per 1s (for 100fps)

Program Demo

- Z-depth estimation is pretty bad (Unity connection was meaningless).
- XGB seems to work much slower than expected.
- Should aim for much lower error rate in future, and higher estimation for z-depth.

Conclusion

- Only One webcam was not enough for accurate estimation (Especially z). Need at least two cameras for more accurate z-depth estimation.
- Training side views in addition to front view will help.
- 3D Rotational data needs to be added for more accurate estimation.
- Physical constraints will help generating true/false labels on some data.
- Overall, we need larger input data for predicting in real-time, so MLP is still preferred.

References

- Seyed Abolfazl Ghasemzadeh, Alexandre Alahi, and Christophe De Vleeschouwer. “MPL: Lifting 3D Human Pose from Multi-view 2D Poses”. In: arXiv eprints (2024), arXiv–2408.
- Julieta Martinez et al. “A simple yet effective baseline for 3d human pose estimation”. In: Proceedings of the IEEE international conference on computer vision. 2017, pp. 2640–2649.
- Dario Pavllo et al. “3d human pose estimation in video with temporal convolutions and semi-supervised training”. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, pp. 7753–7762.
- CMU MoCap
- amc2bvh
- bvh-toolbox
- usbipd
- MediaPipe
- OpenCV