3D Human Pose Estimation from 2D images

Stanford University

Bhargav R Reddy | brkreddy@Stanford.edu

Introduction

- Since most depictions of humans are 2D e.g. images, videos etc., many systems have been designed to estimate from these 2D images the 3D skeleton of the depicted person.
- Many applications would benefit from accurate 3D human pose estimation for e.g. virtual reality, augmented reality and autonomous driving.
- The most challenging part of this problem is to infer the depth information given only a 2D image.

Data

- The dataset chosen was a subset of the Human3.6M dataset with 312188 labeled images of size 256x256x3 and a test set of 10987 unlabeled images on which to perform predictions. The labels provided are the 2D pixel positions of the joints and 3D joint positions in mm w.r.t the root joint (pelvis) which is always at (0, 0, 0).
- 2188 samples are held out for validation to estimate the validation loss and MPJPE and training is done on on the 310k samples left. The data is shuffled and batched with batch size 4. Trained each model for 4 epochs and predicted pose estimates for test data after every epoch.

Model

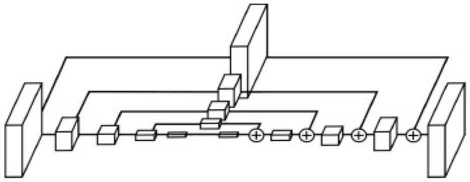

Fig 1 - Hourglass module architecture

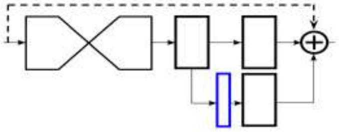

Fig 2 – 1 Hourglass block

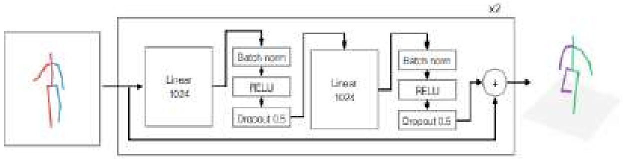

Fig 3 – Network to lift 2D Pose to 3D Pose

- The Hourglass (HG) block (fig 1 and 2 from [1]) forms the central component of both the models under consideration.

Model 1 – 2D HG Model with Feed Forward NN

- Succession of hourglass blocks. Produce heatmaps. These heatmaps are then passed into a feed forward neural network (FFNN) which outputs 17*3 dimensions (17 joints – x, y and z)

Model 2 – 3D HG Model

- In order to predict the 3rd dimension with resolution zi for the i'th HG block, we simply modify the i'th HG block to output 17*zi channels instead of 17 channels (so we get zi channels per joint) We follow the coarse-to-fine approach described in [3] where the resolution zi is increased from HG block to another until a final output resolution zout of the last HG block is reached (choose zout = 64 to match the x and y resolution of the output).

Results

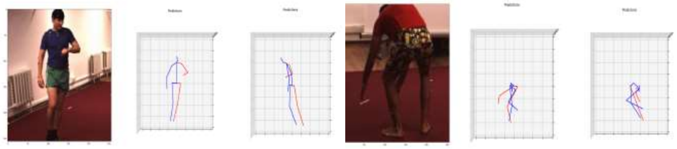

Fig 4 – Success Case (Left) Failure Case (Right) for best pred

The failure case is probably due to the training set having few poses as in this image since most images in the pose depict a person is a Standing or Sitting position, also this image contains many occlusions making it harder for our model to accurately predict the pose.

Model	Learning Rate	HG Blocks	Data Aug	Training Epochs	MPJPE (mm)
2D HG + FFNN	2.50E-04	8	Yes	2	**109.02**
2D HG + FFNN	2.5E-04 + Decay 0.001	4	Yes	2	**110.34**
3D HG	2.50E-04	8	No	4	**99.17**
Per Joint Mean	NA	NA	NA	NA	**89.69**

Although the 2D HG model seemed to benefit from the data augmentation and learning rate decay, the 3D HG model seemed to suffer from both of them since adding them gave a best result of 118.28 after 4 epochs and leaning rate starting at 2.5e-4. Our best score was 89.69 and achieved by taking a per-coordinate average of the predictions with best score for each of the above models.

Conclusion and Future Work

We show the effectiveness of hourglass based models in estimating 3d pose from 2D images. We develop a simple and effective way of training these models and achieve a mean per-joint prediction error (MPJPE) score of 89.69 mm using a combination of both the 3D hourglass model and the 2D hourglass + FFNN. Further experimentation with the base architecture and training over more epochs would be the next steps.