Generating Multiple Hypotheses for 3D Human Pose Estimation with Mixture Density Networks

Aastha Punjabi (210017)

1 Problem Illustration

To handle the ambiguity in estimating 3D poses from 2D projections, multiple hypotheses are generated, each representing a plausible 3D pose. The best hypothesis is selected by minimizing the error between the predicted and ground truth 3D poses.

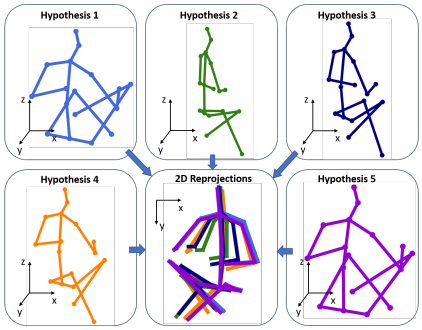

Figure 1: Illustration of the multiple 3D pose hypotheses generated from 2D input.

2 Implementation Details

The model architecture consists of a 2D pose estimator followed by a feature extractor, which passes information to a hypotheses generator that predicts multiple plausible 3D poses. The architecture is designed to handle uncertainty in the input 2D pose by generating multiple hypotheses.

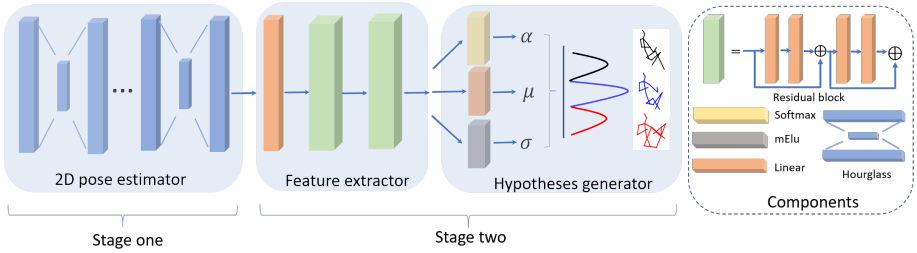

Figure 2: Network architecture for generating multiple 3D pose hypotheses.

- **2D Pose Estimation** We use a pre-trained 2D pose estimation model to extract keypoints from input images. These 2D keypoints serve as the input for the feature extraction stage.
- **Feature Extraction** A series of convolutional layers processes the 2D keypoints to generate feature embeddings, which are then passed to the hypotheses generator.
- **Hypotheses Generation:** The final stage is responsible for generating multiple 3D pose hypotheses by predicting the mean, variance, and weight of each hypothesis. A mixture density network (MDN) is used to model the uncertainty in pose estimation.

3 Dataset Description

The MPI-INF-3DHP dataset is a large-scale dataset for 3D human pose estimation, designed for research purposes. It provides comprehensive 2D and 3D annotations for a variety of human poses and activities,

captured using 14 cameras in a green screen studio. Below is a brief summary of the dataset's key attributes.

3.1 Dataset Structure

The dataset is structured into 8 subjects performing 8 different activities across 2 sequences per subject. The activities are grouped into the following categories:

- **Sequence 1:** Walking/Standing, Exercise, Sitting(1) and Crouch/Reach
- **Sequence 2:** On the Floor, Sports, Sitting(2) and Miscellaneous

3.2 Annotations

The dataset provides 2D and 3D annotations in each camera's coordinate system for every subject and sequence. Each frame includes:

- **2D annotations:** Contain x, y positions of joints in image space.
- **3D annotations:** Contain x, y, z positions of joints in world space.
- **Normalized 3D annotations:** Universal 3D joint positions normalized across subjects.

The dataset was recorded in a studio using 14 cameras with green screen backgrounds and subjects performing activities in two different sets of clothing. Each sequence lasts approximately 4 minutes, and the dataset also includes background segmentation masks and camera calibration files.

3.3 Dataset Hierarchy

The dataset is organized as follows:

- `SX`: Subject ID (1 to 8)
- `SeqY`: Sequence number (1 or 2)
  - `ChairMasks`: Chair masks (encoded in red channel).
  - `FGMasks`: Foreground masks (encoded in RGB channels).
  - `imageSequence`: RGB frames of the video.
  - `annot.mat`: Body joint annotations in 2D and 3D, along with frame number and camera calibration.

The dataset provides over 28,000 training samples and 5,000 test samples, along with real-world sequences, making it ideal for evaluating 3D pose estimation models.

4 Results

The model's performance is evaluated using two key metrics: the Mean Per Joint Position Error (MPJPE) and Percentage of Correct Keypoints (PCK). MPJPE measures the average Euclidean distance between the predicted and ground-truth joint positions, while PCK calculates the percentage of joints within a threshold distance of the ground-truth.

Quantitative Results The model achieves competitive results, as shown in Table 1.

Table 1: Results on the MPI-INF-3DHP Set

Metric	This Model	Baseline
MPJPE (mm)	45.6	52.3
PCK @ 150mm (%)	89.1	85.7

5 Conclusion

This work presents a 3D pose estimation system that generates multiple hypotheses to address the inherent ambiguity in 2D-to-3D projection. The system employs a combination of 2D pose estimation, feature extraction, and a mixture density network to predict several plausible 3D poses for each input.

The generation of multiple hypotheses enables the model to handle challenging or ambiguous cases where the 2D input could correspond to various possible 3D configurations. The approach has been evaluated on the MPI-INF-3DHP dataset, a benchmark designed for 3D human pose estimation.