Martinez, Hossain, Romero, Little

A simple yet effective baseline for 3d human pose estimation

1 Summary

In this paper, the authors report that lifting 2D joint locations (either ground truth or outputted by a pre-existing detector) into 3D via a simple feedforward network outperforms existing end-to-end (pixel-to-pose) systems.

2 Method

The main contribution of this paper is a 3D pose estimation network which takes 2D joint locations and predicts 3D joint locations. The 2D input can be obtained as ground truth (by projecting ground truth 3D points onto 2D with known camera parameters) or via a 2D pose estimation module (necessary for evaluation on new data as ground-truth will not be available).

The 2D and 3D points are normalized by subtracting the mean of all of the training data and dividing by the standard deviation. This means that the output of the network must be un-normalized according to these statistics before being considered "final."

The 3D points are predicted (i.e. supervised) as being relative to the root joint, which ends up being the hip in the paper's implementation. They are also predicted in the camera coordinate frame, which helps standardize the problem across different cameras and prevents overfitting to a global coordinate frame.

The architecture is simple, typically consisting of a linear layer which takes the input dimensionality to 1024, a series of two residual blocks (which themselves involve two linear layers with subsequent batch norms, ReLUs, and dropout, and a residual connection from the block input to block output), and finally a linear layer mapping to the output dimensionality. In total there are six linear layers in this standard configuration.

Residual connections are said to facilitate the training of deep networks, improve generalization performance (*the case here; a 10% error reduction was reported*), and reduce training time. Likewise, batch normalization and dropout are reported to have improved the performance of the system.

A max norm constraint is also imposed, which ensures that the maximum norm of the weights of each layer is less than or equal to 1. This, alongside batch normalization, seems to stabilize training and improve generalization when the distribution between training data and test data is different.

3 Training

The authors train for 200 epochs with a starting learning rate of 0.001 and a batch size of 64. The weights of the linear layers are Kaiming-initialized. A forward-backward pass takes about 5 ms and a forward pass takes about 2 ms (i.e. real-time).

4 Results

The system is evaluated on three 3D human pose datasets, HumanEva, H36M, and MPII (qualitatively, because the ground-truth 3D is not available for this one). To quote the paper, H36M contains "3.6 million

images featuring 7 professional actors performing 15 everyday activities." It includes 2D and 3D ground truth.

Protocol #1: On H36M the authors follow standard protocol and train on subjects 1, 5, 6, 7, and 8 while testing on subjects 9 and 11. They report average millimeter error between predicted and ground truth values across all joints and cameras, after aligning the hip joint. They also train action-specific models. With protocol #1, average error across all actions is 62.9 mm.

Protocol #2: builds on protocol #1 by further aligning the 3D prediction with the ground truth using a rigid transformation. With protocol #2, average error across all actions is 47.7 mm.

When trained and tested on ground truth 2D (no noise), the system achieves a peak performance of 37.1 mm error (under protocol #2, i.e. *with Procrustes alignment*). When trained on ground truth 2D and tested on predicted 2D, the system achieves a peak performance of 60.52 mm error. I assume that this is average error across all joints and actions.

4.1 Analysis

The design choices with the biggest impact on performance are reportedly (1) preprocessing the data to be in camera coordinates (without this, error ends up being above 100 mm), (2) residual connections (removing them increases error by 8 mm), and (3) layer sizes (e.g. using 512 units worsened the result, and using 2048 units increased training time and didn't improve the result).

Qualitatively, the system cannot recover from poor 2D detections, struggles with poses substantially different from those seen in the H36M dataset, and can only deal with full-body poses.

The network also does not have access to visual evidence, which might help in some cases.

References

[1] Martinez, Julieta and Hossain, Rayat and Romero, Javier and Little, James J. A simple yet effective baseline for 3d human pose estimation. ICCV 2017.