Human Pose Estimation - from 2D to 3D

Siddharth Seth [1]

Abstract

The problem of human pose estimation has been extensively studied during the last few years and especially since the advent of deep learning era. Using deep learning algorithms, 2D pose estimation from images and videos gave state of the art results and was thought to be almost a saturated area. Our results however, state otherwise. We extend a 2D to 3D pose estimation architecture, wherein we experiment with two different approaches - Generative Adversarial network and a Deep Residual network with hidden layer loss. Results demonstrate that our approaches outperform baseline predictions with a significant margin on the Human3.6M dataset.

1. Introduction

Today, vast amount of data that we have in the form of images or videos is two dimensional. Even though humans can understand context of a scene from 2D information, it is an arduous task to train machines with such capabilities. Applications include video surveillance, autonomous driving, teaching a robot to interact with the environment etc.

The problem that we aim to tackle in this work is estimating 3D projections of a human pose given its 2D projections. Methods proposed to solve this task include estimating 3D pose from 2D heatmaps, directly estimating 3D joint locations from given 2D joint locations, using depth information with 2D pose to predict 3D pose and several others. In order to train the model in an end-to-end fashion for 3D pose estimation, factors such as camera orientation, background illumination, skin colour, clothing, occlusion need to be accounted for by the algorithm.

Building upon the algorithm put forward by (Martinez et al., 2017), we propose two approaches for lifting 2D projections to 3D projections. First, using a deeper residual network architecture, we achieve superior results than the baseline with a significant margin. Second, using Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) we achieve slightly better results than the baseline. All experiments have been performed on the Human3.6M dataset (Ionescu et al., 2014).

[1]Department of Computational and Data Sciences, Indian Institute of Science, Bengaluru, India. Correspondence to: Siddharth Seth <siddharthseth@iisc.ac.in>.

Proceedings of Machine Learning for Signal Processing Project 2018

2. Previous Work

Pose estimation has been studied from different perspectives according to the need of an application as mentioned before.

Stacked Hourglass: Work of (Newell et al., 2016) is one of the state of the art methods for 2D pose estimation from images. The architecture consists of successive pooling and upsampling to give the final predictions, hence the name Stacked Hourglass. We use this work to take the 2D predictions as set of inputs to our networks.

Coarse to fine: (Pavlakos et al., 2017) estimate 3D pose from 2D predictions of a stacked hourglass architecture. They predict the likelihood of each joint location per voxel by discretizing the 3D space around a subject thereby creating a volumetric output.

We consider the work of (Martinez et al., 2017) using a residual network for predicting 3D joint locations given 2D joint locations either via ground truth images or predictions made by a stacked hourglass architecture as our baseline. They demonstrate that their lightweight model is easy to train and outperforms other state of the art methods with closest competitor being the coarse to fine volumetric prediction approach (Pavlakos et al., 2017).

3. Solution Methodology

As we try to estimate 3D pose from 2D information, we assume that we already have the 2D joint locations. We use ground truth images and stacked hourglass predictions as our input for 2D pose. We thus aim to minimize the error between the predicted 3D pose from our network and the corresponding ground truth 3D pose. More formally, for N poses, we try to learn a function $f^* : R^{2n} \rightarrow R^{3n}$ that minimizes the error given by:

$$f^* = \min_f \frac{1}{N} \sum_{i=1}^{N} L(f(x_i) - y_i). \quad (1)$$

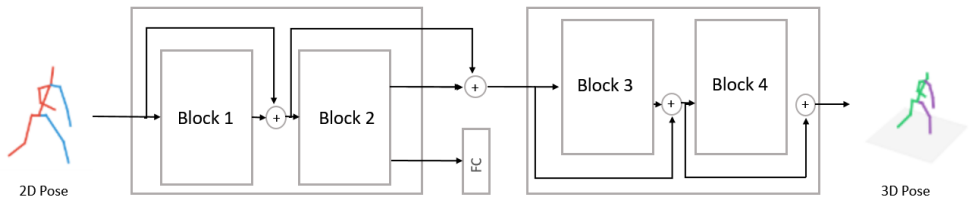


Figure 1. DeepBase2 - Deep Baseline Model with an FC layer between 2 consecutive blocks for loss calculation.

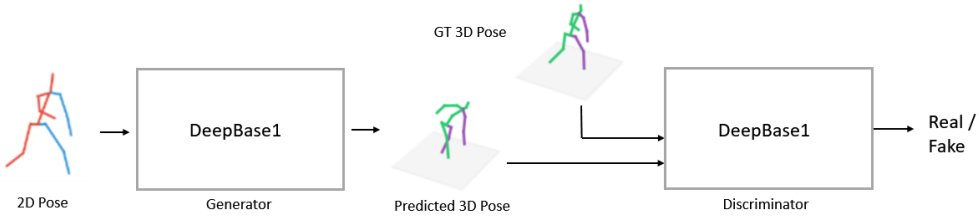


Figure 2. Deep Baseline based GAN architecture

Considering this paper as baseline we propose two approaches, described in the following subsections. Baseline network comprises of two residual blocks with each block composed of 2 fully connected (FC) layers. Two more fully connected layers, one at the input and other at the output result in a total of 6 FC layers with residual connections in between. Each FC layer has a dimensionality of 1024.

3.1. DeepBase Architectures

We experiment with two deeper architectures:

DeepBase1 as we call the first modification, comprises of four residual blocks giving a total of ten fully connected layers. This network performs slightly better than the baseline and going according to the original paper does take more than 200 epochs for training on a batch size of 64. However, our analysis shows this time can be significantly reduced.

DeepBase2 the second proposed network, performs the best of the three methods described here, the analysis for which is presented in the next section. The architecture is an extension to our first proposed model with four residual blocks but an extra fully connected layer after the first two blocks. This FC is used to calculate loss after the first two blocks. Architecture is illustrated in Figure 1.

3.2. Generative Adversarial Network

Our third approach is to use GANs in predicting 3D joint locations. Introduced by (Goodfellow et al., 2014), they have performed satisfactorily in a number of probability density estimation tasks. We use the DeepBase1 architecture for both generator and discriminator. We give 2D joint locations as input to the generator which then tries to predict or generate the corresponding 3D pose. This predicted 3D pose and the 3D ground truth are then fed as input to train the discriminator by teaching it to differentiate between a generated unrealistic pose and a ground truth pose. The generator learns to generate actual 3D poses through this minimax game. The prediction results are better than baseline as well as the DeepBase1 network but second to DeepBase2. Unsurprisingly, the network takes a long time to train with a batch size of 512 running for more than 300 epochs. Architecture is illustrated in Figure 2.

In all our networks, we have used ReLU non-linearity for all but the input and output fully connected layers. We also use batch normalization and dropout as suggested in the baseline for improving the performance. TensorFlow framework has been used for code implementation.

4. Experiments and Analysis

We demonstrate our results on the standard dataset for human pose estimation: Human3.6M (Ionescu et al., 2014), for which both 2D and 3D ground truth joint locations are available. It has 3.6 million images consisting of 15 everyday activities stated in the evaluation table.

DeepBase1's performance was only slightly better than the baseline, improving the average error by a little more than 1mm on ground truth inputs. A detailed analysis showed the exponential decaying of learning rate was not enough for the network to train efficiently. Thus, we had to manually reduce the learning rate as the training started to saturate. However, the major bottleneck to the network's learning was the diminishing gradient flow in the early layers. A comparison between DeepBase1 and DeepBase2's gradient flow is shown in Figure 4. Loss calculated at the hidden layer after the first two residual blocks provided enough gra-

Table 1. Comparison between proposed and baseline architectures on Human3.6M. SH - Stacked Hourglass, GT - Ground truth inputs

Architectures	Direct.	Discuss	Eat	Greet	Phone	Phot	Pose	Purch.	Sitting	SittingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Martinez et al. (GT)	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
DeepBase1 (GT)	36.7	43.8	39.9	41.3	46.5	53.3	43.8	39.4	51.2	57.0	44.0	45.5	47.5	36.4	39.9	44.4
DeepBase2 (GT)	34.4	40.6	37.0	39.4	44.1	51.1	41.3	36.5	48.7	52.1	40.9	42.5	43.3	33.3	35.8	41.4
GANs (GT)	36.2	42.5	38.9	41.9	46.1	55.4	43.2	40.6	52.1	58.2	44.3	42.9	46.4	34.6	36.7	44.0
Martinez et al. (SH)	53.3	60.8	62.9	62.7	86.4	82.4	57.8	58.7	81.9	99.8	69.1	63.9	67.1	50.9	54.8	67.5
DeepBase1 (SH)	54.2	61.1	61.2	63.8	85.0	83.8	57.7	60.9	81.7	105.1	70.4	65.4	69.2	55.2	59.3	68.6
DeepBase2 (SH)	51.6	59.1	58.4	61.0	84.3	79.5	56.8	57.3	80.0	100.3	68.2	62.9	66.1	51.0	54.7	66.1

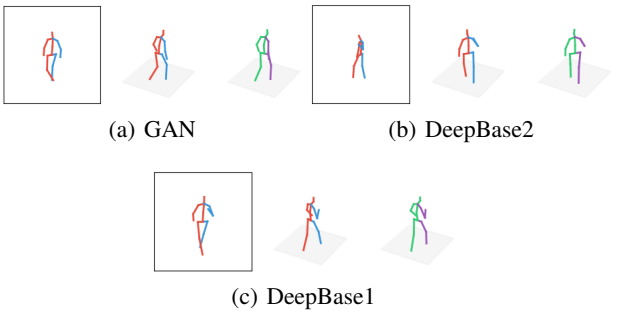
dients to outperform the baseline with a significant margin of 4mm on ground truth 2D joint locations. The results are not so promising when instead stacked hourglass predictions are used as inputs. Using DeepBase1 does not yield any improvement which has been demonstrated by (Martinez et al., 2017) too while DeepBase2 yields an average improvement of about 1.5mm compared to the baseline.

GAN based on DeepBase1 performs better than DeepBase1 but lags behind DeepBase2. It improves upon the baseline by an error of 1.5mm compared to 4mm by DeepBase2. The experiments were carried out only on ground truth 2D joint locations due to their immense training time.

Throughout our experiments, we used exponentially decaying learning rate with at times having to manually reduce it when the training started to saturate until there was completely no effect of changing it. This greatly reduced the training time by a factor of 2 compared to the baseline. A comparison between different networks used and the baseline is shown in Table 1 with all errors being in millimeters. 3D predictions by our networks are illustrated in Figure 3.

5. Conclusion

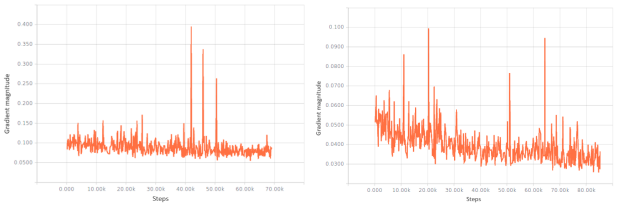
Using the ground truth 2D joint locations as inputs gives a significant bump in improving the networks. Less improvement from stacked hourglass outputs as inputs indicates there is still a lot of scope in improving 2D pose estimation form images which are otherwise noisy compared to ground truth joint locations. Also, using a deeper architecture demands a supervision of gradients as deeper networks suffer from vanishing gradients problem. A thorough analysis of GANs performance needs to be done so as to make any conclusions about their lagging behind DeepBase2. A deeper network is one intuition.



(a) GAN (b) DeepBase2

(c) DeepBase1

Figure 3. 3D Pose predictions. In each figure, left-2D input, middle-GT 3D, right-predicted 3D



(a) DeepBase2 first layer (b) DeepBase1 first layer

Figure 4. Flow of gradients

References

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014.

Ionescu, Catalin, Papava, Dragos, Olaru, Vlad, and Sminchisescu, Cristian. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

Martinez, Julieta, Hossain, Rayat, Romero, Javier, and Little, James J. A simple yet effective baseline for 3d human pose estimation. *arXiv preprint arXiv:1705.03098*, 2017.

Newell, Alejandro, Yang, Kaiyu, and Deng, Jia. Stacked hourglass networks for human pose estimation. In Leibe, Bastian, Matas, Jiri, Sebe, Nicu, and Welling, Max (eds.), *Computer Vision – ECCV 2016*, pp. 483–499, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46484-8.

Pavlakos, Georgios, Zhou, Xiaowei, Derpanis, Konstantinos G, and Daniilidis, Kostas. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.