**3D Human pose estimation in video using attention and deep interpolation**

Alex Yudashkin

[alexyuda@gmail.com](mailto:alexyuda@gmail.com) / [yudashkin@mail.tau.ac.il](mailto:yudashkin@mail.tau.ac.il)

304325426

**Abstract**

This project is based on [[1]](#sm3p8vyf12zw). the task is to estimate 3D human pose using a fully convolutional model based on a dilated temporal network over 2d keypoints (skeletons) and a semi-supervised training model.

In [[1]](#sm3p8vyf12zw) a set of N skeletons is used to estimate a 3d human pose, inspired by attention in NLP, the first objective of this project is to show that some skeletons (in an N sequence of skeletons) are more “important” than others in the task of 3d human pose estimation by adding a shallow network that will predict a value of 0 to 1 for each skeleton and act as an attention mechanism.

The second objective of this project is to learn deep 3d pose interpolation by using the learned attention mechanism to “remove” (assigning a 0 value to a skeleton attention confidence coefficient) samples from the N sequence of skeletons and validate the ability of the architecture to interpolate the correct 3d pose while samples are “missing”.

The second objective can be useful in real life scenarios where occlusion can lead to missing skeletons samples.



**Architecture inputs**

1. (N,3J) - where N (N = 243) is the number of skeletons, J (J = 17) is the number of skeleton joints.

Each joint is described by x location,y location and **confidence** (confidence is added as part of this project).

* In some 2d keypoints estimators as in [[3]](#89s9sfqquhn7), joint confidence is part of the network output. For this type of 2d key points estimators we can exploit the additional data of joint confidence.
* When joint confidence is not available as in [[2]](#j1sy80xil7cv) joint confidence value will be initialized as 1.

1. Attention network input is the same as in 1.

**Architecture outputs**

1. (1,51) - 3d pose estimation for the 17 joints.
2. Attention output is a (243,1) vector, each element in the vector denotes the attention confidence coefficient for the appropriate input skeleton.

Each attention confidence coefficient is multiplied with the third channel of the appropriate input skeleton which is the confidence channel.

**Relevant work**

1. The project is based on [[1]](#sm3p8vyf12zw).
2. CPN - pre trained 2d key point estimator [[2]](#j1sy80xil7cv).

**Existing code**

<https://github.com/facebookresearch/VideoPose3D> - code for [[1]](#sm3p8vyf12zw)

**Proposed method**

The proposed architecture can be seen in Figure 1.

Pre trained CPN will be used as 2d key points estimator.

Training method:

The full architecture will be trained together end to end.

1. First iteration - all training data will be used to train the attention and the 3d pose estimation network.
2. Second iteration - after attention was learned, center samples will be “removed” from training data (assigning a 0 value to a skeleton attention confidence coefficient) , for example from a 243 sequence of skeletons we can “remove” 50 samples (e.g samples 96-145).

Loss: the loss is as described in [[1]](#sm3p8vyf12zw):

1. 3d reconstruction loss for labeled data.
2. 3d Trajectory loss for labeled data.
3. Bone length loss for labeled data.
4. 2d backprojection loss from the estimated 3d pose - is used for labeled and unlabeled data.

**Novelty**

1. Adding attention to 3d pose estimation scheme.
2. Using attention to learn 3d pose interpolation.

**Data set**

1. Human3.6M  [[4]](#4tnt3vm8g8i)
2. HumanEva-I [[5]](#k6ofwtgvefw4)

**References**

[1] Pavllo, Dario, et al. "3D human pose estimation in video with temporal convolutions and semi-supervised training." *arXiv preprint arXiv:1811.11742* (2018).

[2] Chen, Yilun, et al. "Cascaded pyramid network for multi-person pose estimation." *arXiv preprint arXiv:1711.07319*(2017).

[3] Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." *arXiv preprint arXiv:1611.08050*(2016).

[4] Ionescu, Catalin, et al. "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 (2014): 1325-1339.

[5] Sigal, Leonid, Alexandru O. Balan, and Michael J. Black. "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion." International journal of computer vision 87.1-2 (2010): 4.