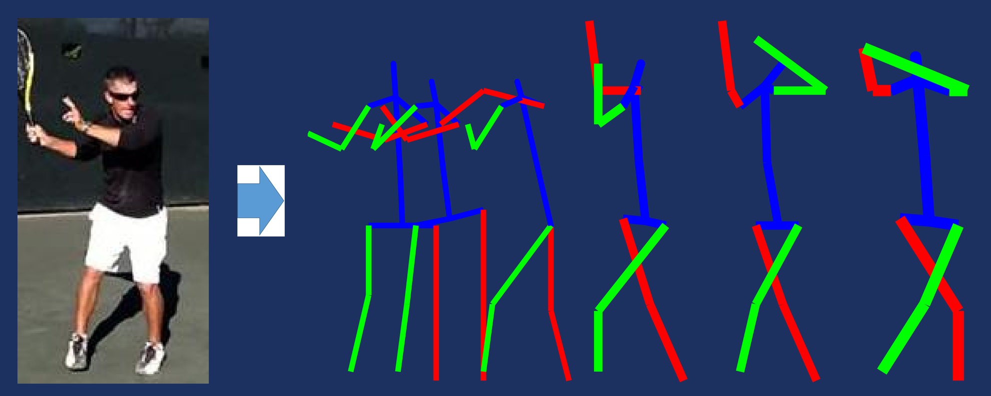
Forecasting Human Dynamics from Static Images

Yu-Wei Chao[1], Jimei Yang[2], Brian Price[2], Scott Cohen[2], Jia Deng[1]

[1]University of Michigan, Ann Arbor [2]Adobe Research



Human Pose Forecasting

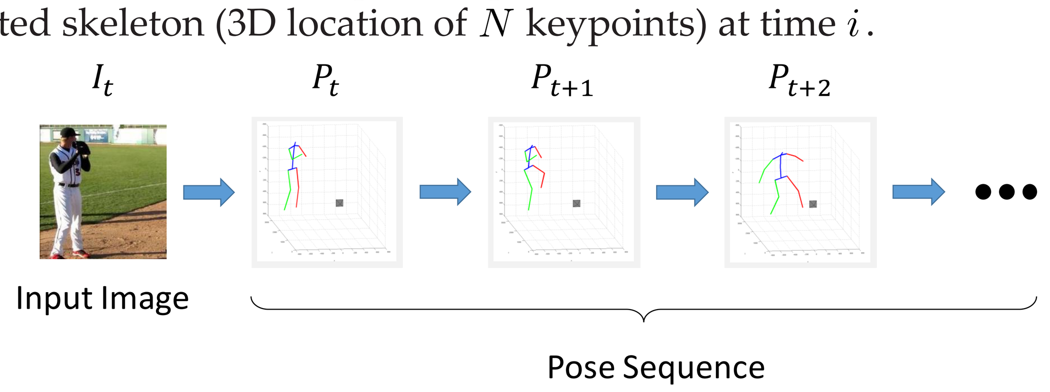
Why Important?

1. The ability of forecasting reflects a higher-level intelligence beyond perception and recognition.
2. For robot assistants, action forecasting is particularly crucial since the capability of action recognition (i.e. identifying action categories after a complete obseravtion) is not sufficient for providing timely response.

Problem Statement

Input I_t : A single image captured at time t.

Output $\{P_t, \ldots, P_{t+T}\}$: A pose sequence, where $P_i \in \mathbb{R}^{3 \times N}$ denotes the predicted skeleton (3D location of N keypoints) at time i.



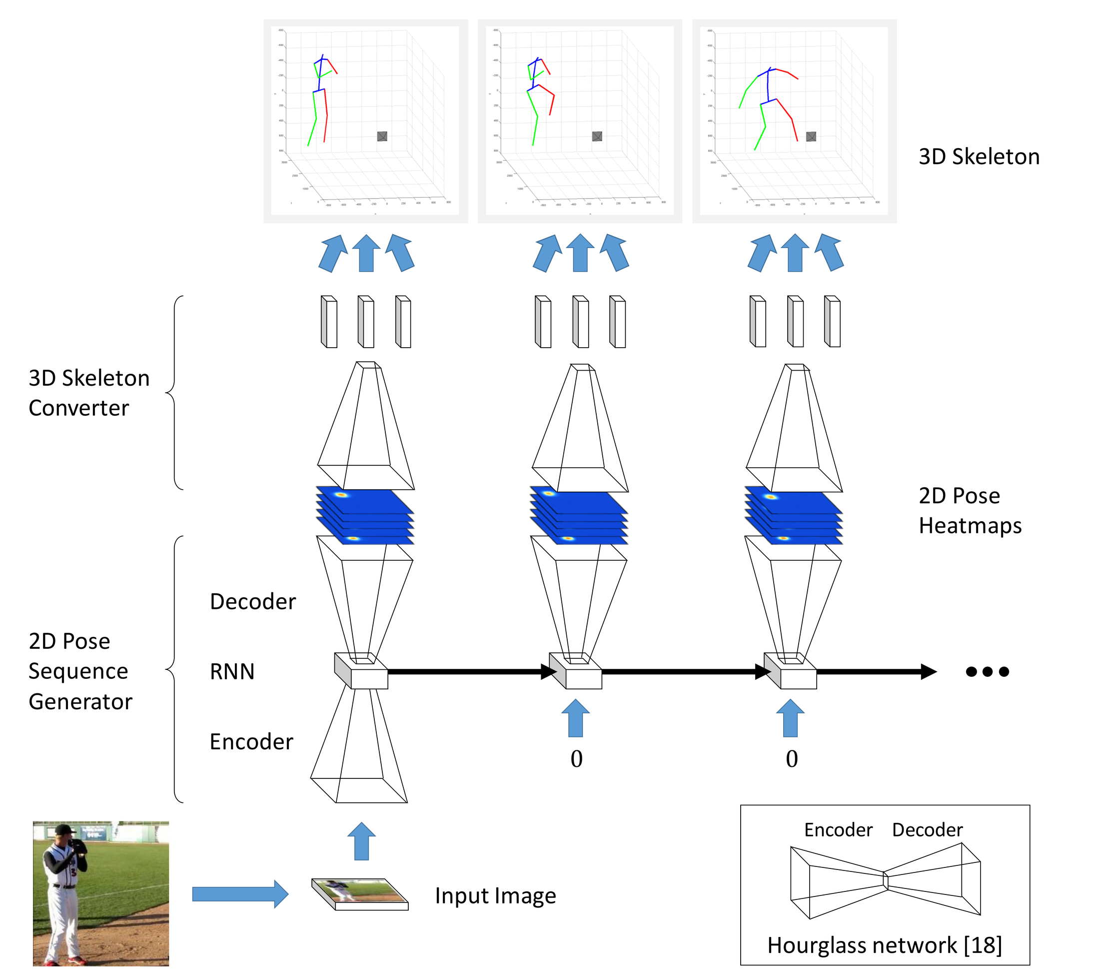
Contributions

1. The first study on single-frame human pose forecasting.
2. A novel DNN-based approach for 2D pose forecasting and 3D pose recovery.
3. Outperforming performance over strong baselines on 2D forecasting and over two state-of-the-art methods on 3D pose recovery.

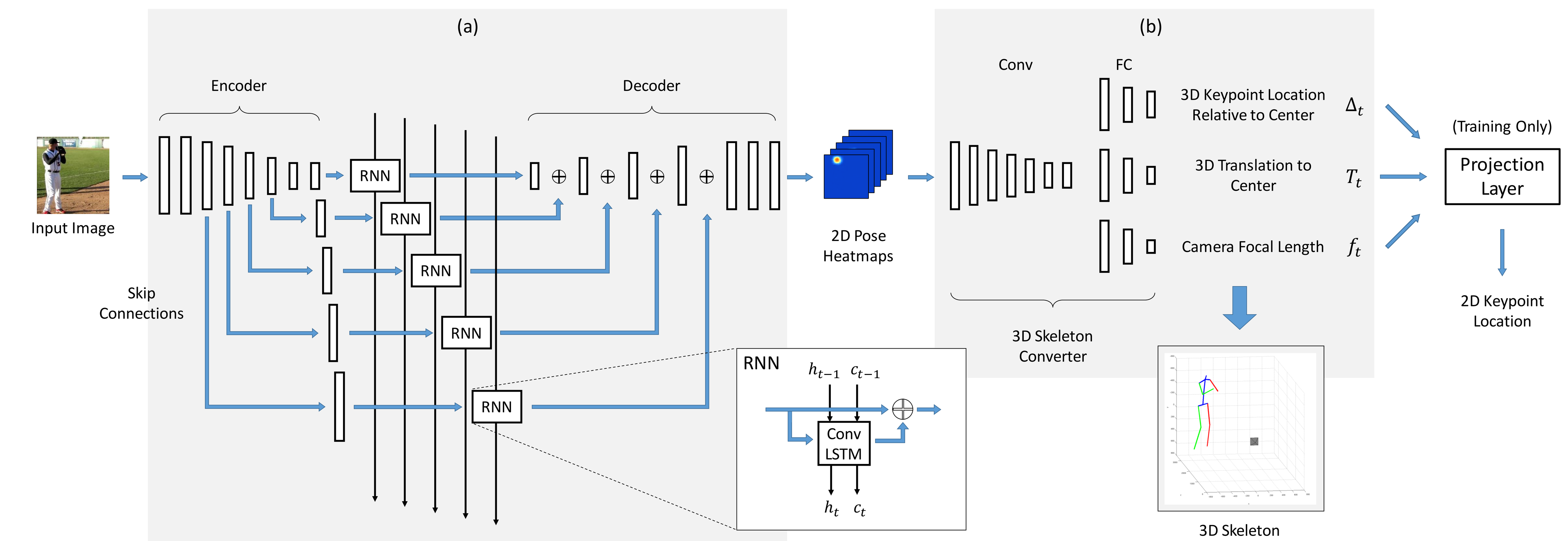
Approach

3D Pose Forecasting Network (3D-PFNet)

1. Integrate recent advances on (1) single-image human pose estimation (*hourglass networks* [19]) and (2) sequence prediction (*recurrent neural networks*).
2. Convert 2D predictions into 3D space.



Architecture of the 3D-PFNet



(a) Recurrent hourglass architecture for 2D pose forecasting (b) 3D skeleton converter

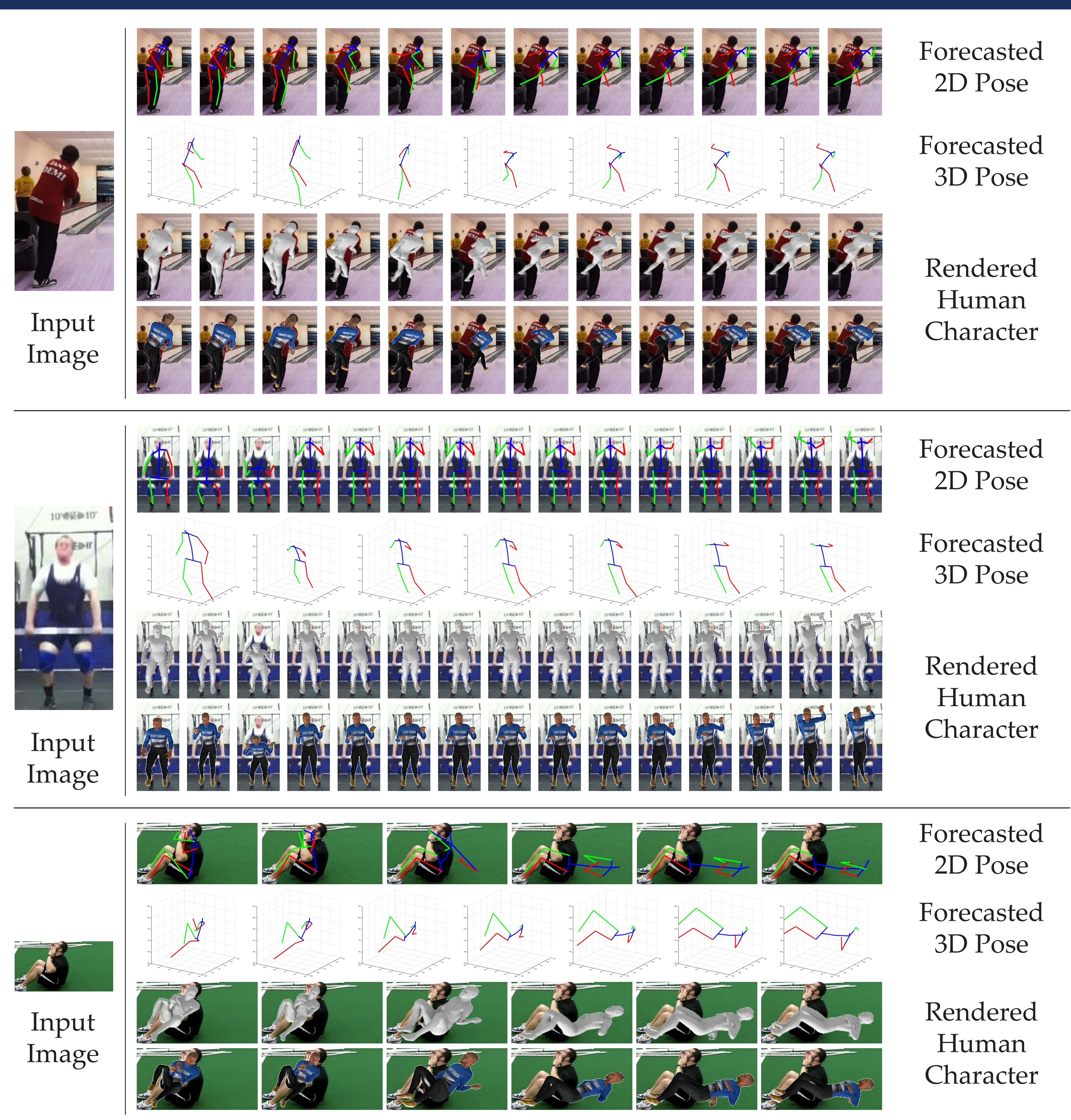
Three-Step Training Strategy

1) Hourglass
Pre-train the hourglass network on the MPII Human Pose datase and fine-tune on the Penn Action dataset.

2) 3D Skeleton Converter
Train on the Human3.6M (MoCap) dataset by synthesizing 2D pose heatmaps from 3D ground truths.

3) Full Network
Train on Penn Action using static images and their corresponding ground-truth pose sequence.

Qualitative Results

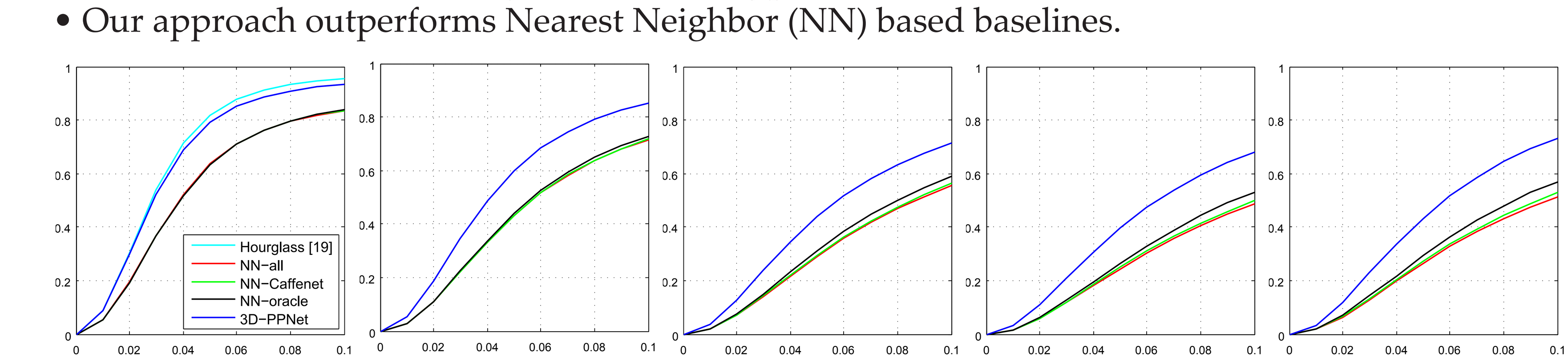


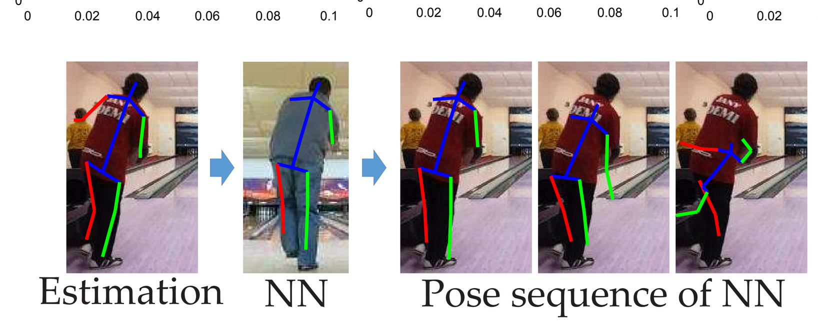
- Human characters are rendered using the public code provided by Chen *et al.* [4].
- See `http://www.umich.edu/~ywchao/image-play/` for more qualitative results.

Quantitative Results

1) 2D Pose Forecasting on Penn Action

- Evaluation metric: Percentage of Correct Keypoints (PCK).
- Our approach outperforms Nearest Neighbor (NN) based baselines.





- Near Neighbor may fail due to a small error in pose alignment.
- The internal feature representation learned by 3D-PFNet can better align human poses in the given context.

2) 3D Pose Recovery on Human3.6M

- Evaluation metric: mean per joint position error (MPJPE) in mm.
- Our approach outperforms two state-of-the-art methods.

	Head	R.Sho	L.Sho	R.Elb	L.Elb	R.Wri	L.Wri	R.Hip	L.Hip	R.Kne	L.Kne	R.Ank	L.Ank	Avg
Convex [40]	145.3	123.5	122.8	139.1	129.5	162.2	153.0	115.2	111.8	172.1	171.7	257.4	258.5	158.6
SMPLify [3]	132.3	117.4	119.3	149.6	149.5	204.3	192.8	140.9	124.0	131.9	135.3	202.3	213.6	154.9
Ours	**72.3**	**64.7**	**63.5**	**93.9**	**88.8**	**135.1**	**124.2**	**59.1**	**57.5**	**75.7**	**76.5**	**113.6**	**113.4**	**87.6**

References

[3] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In ECCV, 2016.

[4] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3d pose estimation. In 3DV, 2016.

[19] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In ECCV, 2016.

[40] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis. 3d shape estimation from 2d landmarks: A convex relaxation approach. In CVPR, 2015.