Stacked Hourglass Networks for Human Pose Estimation

Alejandro Newell, Kaiyu Yang, Jia Deng
University of Michigan, Ann Arbor

1 Introduction

A key step toward understanding people in images and videos is accurate pose estimation, which precisely localizes keypoints of the body. Methods based on Convolutional Neural Networks (ConvNets) [2, 8, 9, 11], have greatly replaced classical methods and yielded drastic improvements on standard benchmarks [1, 6]. We introduce a novel 'stacked hourglass' network design capturing and consolidating information across all scales. A hourglass module pools down to a very low resolution, then uses a symmetric topology to upsample and combine features across multiple resolutions. Two hourglass modules are placed together end-to-end with intermediate supervision, allowing for repeated bottom-up, top-down inference The final model achieves an improvement on the state-of-the-art for two standard benchmarks (FLIC [6] and MPII Human Pose [1]).

Our method for combining features across different resolutions resembles Tompson et al. [9]. But they use a deep ConvNet and a graphical model jointly, whereas we achieve superior performance without a graphical model or any explicit enforcement of human body structure. Other attempts to improve pose estimation performance include iterative or multi-stage methods [2, 11], cascaded refinement of predictions [8], and leveraging of additional information such as depth or motion cues [4, 7]. Our approach shares some features with work featuring intermediate supervision, but offers a different building block (the hourglass).

Our work is closely connected to fully convolutional networks [5] and other designs processing spatial information in multiple scales for dense prediction. Our hourglass module (without being stacked) differs from these designs in its symmetric topology and roughly equal distribution of model capacity between bottom-up processing (from high resolutions to low resolutions) and top-down processing (from low resolutions to high resolutions). For example, fully convolutional networks [5] are heavy in bottom-up processing but light in top-down processing, as they only perform a weighted merging of predictions across multiple scales. Another major difference is that they perform a single pass of bottom-up, top-down inference, whereas we perform repeated bottom-up, top-down inference by stacking two hourglass modules.

The hourglass module is also related to convolution-deconvolutional architectures, which deploys a DeconvNet to do pixel-wise prediction. The symmetric topology is similar to our hourglass, but the nature of the operations is quite different in that we do not use unpooling or deconvolutional layers. Instead, we rely on simple nearest neighbor upsampling and skip connections for top-down processing.

2 Network Architecture

Figure 1 shows the overall architecture of two hourglasses stacked together. The hourglass design (without being stacked) is motivated by the necessity to capture information at every scale. It branches off at each resolution and combines information across multiple resolutions by nearest neighbor upsampling followed by elementwise addition. After reaching the output resolution, three consecutive 1x1 convolutions are applied to produce the final predictions, which is a set of heatmaps indicating probabilities of each joint's presence at every pixel.

While maintaining the overall hourglass shape, we explore several options in the specific implementation of layers and choose the residual learning modules. The first layer is a standard 7x7 convolution with stride 2, and anywhere else that the resolution drops implies max pooling with a 2x2 window and stride 2. All residual modules output 256 features except for layers right before upsampling where there are 512.

Two hourglasses are stacked end-to-end with intermediate loss, providing repeated bottom-up, top-down inference and reevaluation of initial estimates across the whole image. The first hourglass predicts an initial set of heatmaps upon which we apply a loss. Then, the second hourglass processes these high level features again across all scales to further capture higher order spatial relationships, which is critical to the final performance.

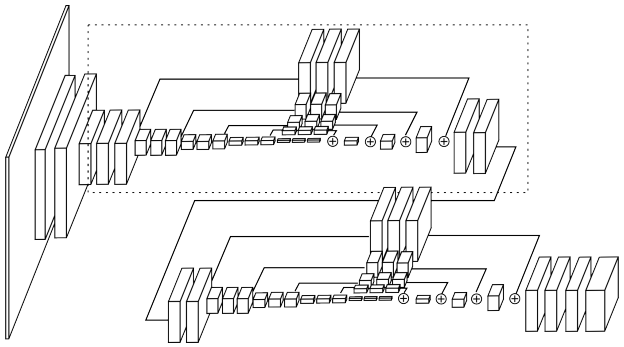
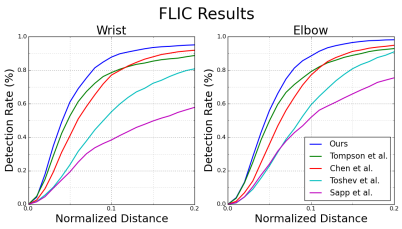


Figure 1: The stacked hourglass design. The dashed lines surround one "hourglass" module. Each box is a residual module

3 Results

We evaluate our network on FLIC [6] and MPII Human Pose [1]. For each sample in MPII , we crop around the target person and resize it to 256x256. Data augmentation includes flipping, rotation, and scaling. The supervision is the same as Tompson et al [9].

FLIC: Our results on FLIC (Figure 2) are very competitive reaching almost perfect performance on the shoulder and elbow (observer-centric Percentage of Correct Keypoints(PCK) metric at a normalized distance threshold of .2), and 95.2% on the wrist. We run on only one hourglass module without intermediate supervision.

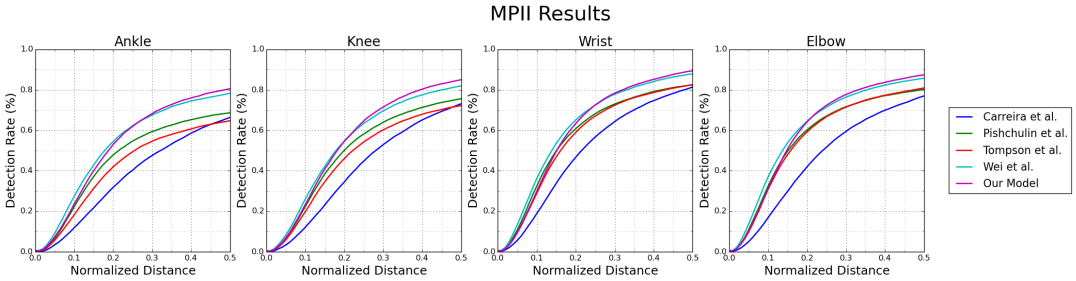


	Elbow	Wrist
Tompson et al.[8]	93.1	89.0
Chen et al.[3]	95.3	92.4
Toshev et al.[10]	92.3	82.0
Sapp et al.[6]	76.5	59.1
Ours	**98.2**	**95.2**

Figure 2: Pose estimation results on FLIC (PCK@0.2)

MPII: We achieve state-of-the-art results on the MPII Human Pose dataset (Figure 3, PCKh metric). On difficult joints like the wrists, elbows, ankles, and knees we improve upon the most recent state-of-the-art results by a margin of 1-2%. For the elbow we reach a final accuracy of 90% and for the wrist an accuracy of 85.2%.

Figure 4 shows abalation experiments exploring the effect of different design choices on performance and training speed. First, for the stacked hourglass design, we compare our stacked hourglass model (HG-Stacked) with a single hourglass (HG) with the same number of layers and approximately the same number of parameters. Next, for intermediate supervision,



	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Tompson et al., NIPS'14	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
Carreira et al., CVPR'16	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Tompson et al., CVPR'15	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Pishchulin et al., CVPR'16	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Wei et al., CVPR'16	**97.8**	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Our model	97.6	**95.4**	**90.0**	**85.2**	**88.7**	**85.0**	**80.6**	**89.4**

Figure 3: Pose estimation results on MPII Human Pose (PCKh@0.5)

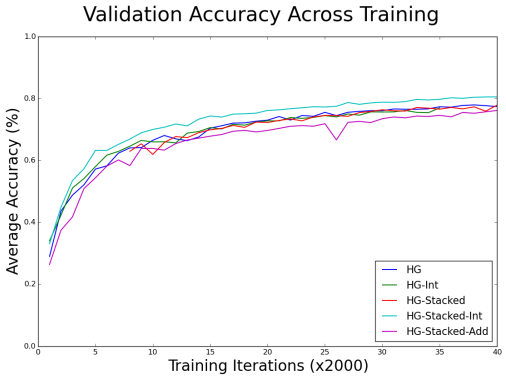
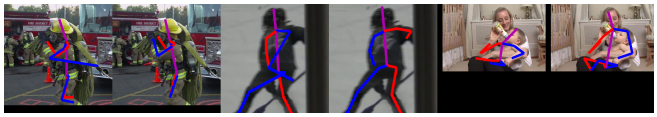


Figure 4: Comparison of validation accuracy as training progresses. The accuracy is the averaged across the wrists, elbows, knees, and ankles.



	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle
Intermediate	96.7	92.9	85.0	79.8	83.7	78.6	74.3
Final	97.7	94.6	88.5	83.3	87.5	83.0	79.0

Figure 5: Examples and validation performance (PCKh) contrasting intermediate and final predictions

we compare HG and HG-Stacked with their intermediately supervised versions: HG-Int and HG-Stacked-Int. Our last experiment HG-Stacked-Add replaces two separate losses with a single loss applied to the sum of intermediate and final heatmaps. The results indicate a dramatic improvement on both training speed and localization performance when both stacking and intermediate supervision are applied.

To see the reevaluation done by the second hourglass, in Figure 5, we compare the final output to the intermediate predictions produced by the first hourglass. We compare the PCKh metric and visualize some qualitative results. We see the network do the job often reserved for a graphical model, enforcing global consistency across predictions.

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3686–3693. IEEE, 2014.

[2] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. *arXiv preprint arXiv:1507.06550*, 2015.

[3] Xianjie Chen and Alan Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[4] Arjun Jain, Jonathan Tompson, Yann LeCun, and Christoph Bregler. Modeep: A deep learning framework using motion features for human pose estimation. In *Computer Vision–ACCV 2014*, pages 302–315. Springer, 2014.

[5] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[6] Brian Sapp and Ben Taskar. Modec: Multimodal decomposable models for human pose estimation. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3674–3681. IEEE, 2013.

[7] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

[8] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 648–656, 2015.

[9] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems*, pages 1799–1807, 2014.

[10] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1653–1660. IEEE, 2014.

[11] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. *arXiv preprint arXiv:1602.00134*, 2016.