Improving Human Pose Estimation: A Comparative Analysis of SimpleBaseline and Stacked Hourglass Networks Using the MPII Dataset and PCKh-0.5 Metric

DongJoo Hwang
AI Researcher
Aiffel Research 10th
hdjcool@gmail.com

Abstract

Human pose estimation is a critical technology designed to predict human posture by detecting keypoints, such as joints and essential body parts. This study evaluates and compares the performance of two prominent models: the Stacked Hourglass Network and the SimpleBaseline model, using the MPII Human Pose Dataset. The evaluation utilizes the PCKh@0.5 metric (Percentage of Correct Keypoints with head-normalized threshold) to measure prediction accuracy.

Our findings demonstrate that the SimpleBaseline model consistently outperforms the Stacked Hourglass Network in both quantitative and qualitative evaluations. The SimpleBaseline model demonstrates lower training and validation loss and improves accuracy across all keypoints.

These results highlight SimpleBaseline's efficiency and accuracy, establishing it as an ideal choice for real-world human pose estimation tasks. Future work will explore optimization strategies and additional architectural enhancements to further improve performance.

Keywords: Human Pose Estimation, SimpleBaseline, Stacked Hourglass Network, MPII Dataset, PCKh@0.5, Keypoint Detection

1 Introduction

Detecting human posture through keypoint-based human pose estimation has become a fundamental task in computer vision due to its relevance in applications like healthcare and human-computer interaction. It serves as a critical technology in various applications, including healthcare monitoring, sports performance analysis, and human-computer interaction (HCI). Despite significant advancements, pose estimation continues to face challenges such as occlusions, complex poses, and variations in image conditions. This study compares two widely-used models for human pose estimation: Stacked Hourglass Network and SimpleBaseline. By leveraging the MPII Human Pose Dataset and evaluating with the PCKh@0.5 metric, we aim to identify the strengths and weaknesses of each approach and demonstrate the practical advantages of SimpleBaseline over Stacked Hourglass Network.

2 Background

Human pose estimation involves detecting keypoints such as shoulders, elbows, and knees from images to estimate the human body's structure. Keypoint detection accuracy directly influences

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

downstream tasks, including activity recognition, motion tracking, and augmented reality. Challenges in Pose Estimation:

- Occlusions: Keypoints hidden by objects or other body parts.
- Complex Poses: Unusual body positions causing ambiguity.
- Scale Variations: Differences in person size due to camera distance.

To address these challenges, deep learning models such as Stacked Hourglass Network and SimpleBaseline have been introduced.

3 Related Works

3.1 Stacked Hourglass Network:

The Stacked Hourglass Network, operating as a multi-scale feature extraction model, utilizes stacked hourglass modules to capture fine-grained details across various resolutions. While effective, it faces challenges of high computational costs and training instability.

3.2 SimpleBaseline Model:

The SimpleBaseline model simplifies the architecture by employing a ResNet backbone for feature extraction and a deconvolutional decoder to predict keypoint heatmaps. This design enhances computational efficiency and prediction accuracy.

This study compares these two models on the MPII Human Pose Dataset using the PCKh@0.5 metric to provide clarity on their respective strengths and weaknesses.

4 Method

4.1 Dataset: MPII Human Pose Dataset

In this study, we use the MPII Human Pose Dataset to train and evaluate our human pose estimation models. The dataset comprises over 25,000 images and more than 40,000 human instances, with each human body represented by 16 keypoints, including the head, shoulders, elbows, wrists, hips, knees, and ankles. Additionally, the dataset captures a variety of poses, angles, background complexities, and occlusions, making it suitable for real-world evaluation.

4.2 Baseline Model: Stacked Hourglass Network

The Stacked Hourglass Network is designed to predict human poses through multi-resolution feature extraction and a symmetrical structure. This network stacks multiple Hourglass modules to capture multi-scale features, enabling the prediction of complex poses and fine details. However, the network suffers from high computational costs and training difficulties due to its deep architecture. To address these limitations, we propose the SimpleBaseline model.

4.3 Proposed Model: SimpleBaseline

SimpleBaseline adopts an encoder-decoder architecture to provide an efficient and intuitive approach to pose estimation. The model employs ResNet as the backbone to effectively extract image features and then predicts keypoint heatmaps using a decoder network. The key improvements of SimpleBaseline include:

- Simplified Architecture: Removing unnecessary complexity to improve computational efficiency.
- Accurate Feature Extraction: Leveraging ResNet's strong representational power for precise feature extraction.

These improvements contribute to faster training speed and higher accuracy in pose prediction.

4.4 Training Details

The training process utilizes a customized Mean Squared Error (MSE) loss function to minimize pixel-wise errors between the ground truth heatmaps and the predicted heatmaps. Notably, a weight of 81 is assigned to keypoint regions in the ground truth to emphasize their importance during training. The loss function is defined as:

$$\text{Loss} = \frac{1}{\text{global\_batch\_size}} \sum (\text{labels} - \text{outputs})^2 \cdot \text{weights}$$

This weighting mechanism ensures the prioritization of keypoints during the training process, maintaining a balanced emphasis across the heatmap.

Data augmentation techniques such as random rotation, horizontal flipping, and scaling adjustments were applied to enhance the model's generalization performance.

4.5 Evaluation Metric: PCKh@0.5

The PCKh@0.5 (Percentage of Correct Keypoints with head-normalized threshold) metric is used to evaluate the model's performance. This metric determines whether a predicted keypoint is correct by measuring the Euclidean distance between the predicted and ground truth keypoints. A prediction is considered correct if the distance is within 50% of the head length.

Keypoint-specific PCKh@0.5 scores are calculated to provide a detailed performance assessment, ensuring robustness against variations in human poses and scaling factors.

4.6 Preprocessing Pipeline

To efficiently preprocess input images and keypoint data, we designed a Preprocessor class. The pipeline consists of the following steps:

1. Image Resizing and Normalization:
   (a) Original images are resized to (256, 256, 3).
   (b) Pixel values are normalized to the range [-1, 1].
2. Region of Interest (ROI) Cropping:
   (a) The ROI is determined based on the minimum/maximum keypoint coordinates and body height.
   (b) A random margin is added during training to enable data augmentation.
3. Keypoint Normalization:
   (a) After ROI cropping, keypoint coordinates are normalized to a range of (0, 1).
4. Gaussian Heatmap Generation:
   (a) Each keypoint is converted into a 64x64 Gaussian Heatmap.
   (b) Invisible keypoints (visibility == 0) are excluded from the heatmap.
   (c) The heatmap consists of 16 channels, one for each keypoint.
5. Dataset Parsing

This preprocessing pipeline ensures that the data is clean, normalized, and properly formatted for effective model training and evaluation.

5 Results – Model Comparison (Stacked Hourglass Network vs. SimpleBaseline)

5.1 Experimental Setup

In this study, we compared the Stacked Hourglass Network and SimpleBaseline models to evaluate their performance on the MPII Human Pose Dataset. Both models were evaluated using the

PCKh@0.5 metric, and training loss, validation loss, and keypoint-specific PCKh values were analyzed. Here are the main settings for the experiment:

- **Number of Epochs:** 20
- **Optimizer:** Adam Optimizer
- **Learning Rate:** 0.0007
- **Evaluation Metric:** PCKh-0.5

5.2 Quantitative Comparison

Table 1: Comparison of Training and Validation Performance Between Stacked Hourglass and SimpleBaseline Models

Model	Train Loss	Val Loss	Average PCKh (Train)	Average PCKh (Val)
Stacked Hourglass	0.9959	1.1075	0.0002457	0.0002457
SimpleBaseline	0.2172	0.2917	0.12825	0.12825

Table 1 provides a comparative analysis of training loss, validation loss, and average PCKh scores for the Stacked Hourglass Network and SimpleBaseline models, evaluated on the MPII Human Pose Dataset.

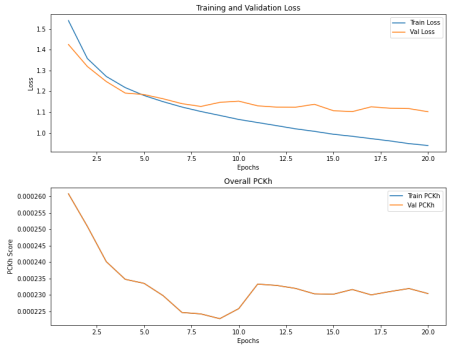

Figure 1: Training and Validation Loss and Overall PCKh for Stacked Hourglass Network

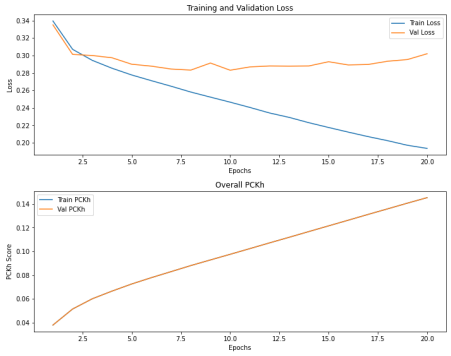

Figure 2: Training and Validation Loss and Overall PCKh for SimpleBaseline

Figure 1 illustrates the training and validation loss curves (top) along with the overall PCKh scores (bottom) over 20 epochs for the Stacked Hourglass Network model.

Figure 2 illustrates the training and validation loss curves (top) along with the overall PCKh scores (bottom) over 20 epochs for the SimpleBaseline model.

Table 2 demonstrates the superiority of SimpleBaseline in terms of both training efficiency and loss minimization. With 49% less training time per epoch and significantly lower training loss, SimpleBaseline outperforms the Stacked Hourglass Network in practical scenarios requiring both efficiency and accuracy.

Analysis:

- Average PCKh: SimpleBaseline demonstrated a significant improvement in the average PCKh score across both training and validation phases.
- Loss: The SimpleBaseline model achieved lower training and validation losses compared to the Stacked Hourglass Network, indicating better convergence.

The training times and performance of the Stacked Hourglass Network and SimpleBaseline model were compared over 5 epochs on the same dataset. The results, summarized in Table 2, indicate that SimpleBaseline requires 825.03 seconds per epoch on average, compared to 1687.45 seconds for the

Table 2: Model Training Time per Epoch and Loss

Model	Epoch	Training Time (seconds)	Training Loss
Stacked Hourglass	1	1783.27	1.5394
	2	1659.95	1.3574
	3	1648.77	1.2713
	4	1641.31	1.2179
	5	1703.94	1.1797
Average		**1687.45**	**1.3131**
SimpleBaseline	1	828.63	0.3396
	2	816.67	0.3071
	3	816.83	0.2944
	4	839.62	0.2854
	5	823.38	0.2777
Average		**825.03**	**0.3008**

Stacked Hourglass Network. This 49% reduction in training time highlights SimpleBaseline's computational efficiency, which is attributed to its simpler architecture leveraging a ResNet backbone and deconvolution-based decoder. In addition to faster training, SimpleBaseline achieves a consistently lower training loss, averaging 0.3008, compared to 1.3131 for the Stacked Hourglass Network. This suggests that SimpleBaseline not only trains faster but also optimizes better, making it a preferable choice for scenarios requiring efficient and accurate human pose estimation.

5.3 Keypoint-wise PCKh Comparison

Table 3: Keypoint-wise PCKh-0.5 Performance Comparison Between Stacked Hourglass Network and SimpleBaseline

Keypoint	Stacked Hourglass (Val)	SimpleBaseline (Val)	Description
Keypoint 1	0.00017223	0.06383678	Top of the head
Keypoint 2	0.00018018	0.07470323	Neck
Keypoint 3	0.00029677	0.0686248	Right shoulder
Keypoint 4	0.00022258	0.0677345	Left shoulder
Keypoint 5	0.00018283	0.07642024	Right elbow
Keypoint 6	0.00013249	0.0663752	Left elbow
Keypoint 7	0.00025172	0.11232644	Right wrist
Keypoint 8	0.00024907	0.21215156	Left wrist
Keypoint 9	0.00024377	0.2322761	Right hip
Keypoint 10	0.00025702	0.18190514	Left hip
Keypoint 11	0.00025172	0.1150106	Right knee
Keypoint 12	0.00026232	0.13591945	Left knee
Keypoint 13	0.00021993	0.13631955	Right ankle
Keypoint 14	0.00028352	0.1368283	Left ankle
Keypoint 15	0.00031532	0.13261527	Right eye
Keypoint 16	0.00023317	0.11382618	Left eye

Table 3 presents a detailed comparison of PCKh-0.5 scores for each keypoint between the Stacked Hourglass Network and the SimpleBaseline model. The PCKh-0.5 metric measures the accuracy of each keypoint prediction, normalized by head length. Higher values indicate better prediction accuracy.

Analysis:

- The SimpleBaseline model outperformed the Stacked Hourglass Network across all keypoints.
- Significant improvements were observed in keypoints like left wrist (Keypoint 8), right hip (Keypoint 9), and left hip (Keypoint 10).

6 Discussion

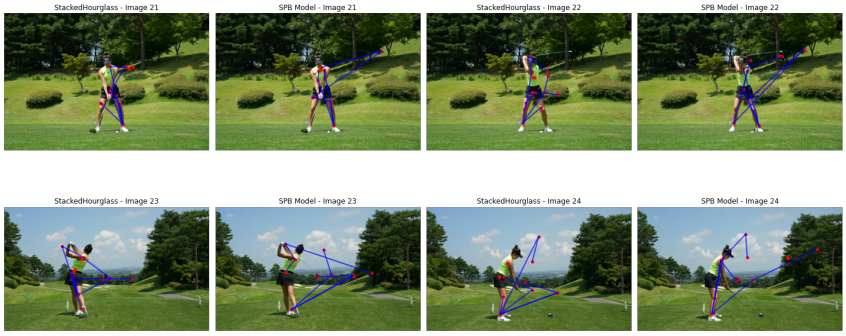

Figure 3: Qualitative Comparison of Keypoint Predictions Between Stacked Hourglass Network and SimpleBaseline

The qualitative evaluation (as shown in Figure 3) illustrates the challenges of both the Stacked Hourglass Network and the SimpleBaseline model in accurately predicting human pose keypoints during complex and dynamic scenarios. Despite achieving reasonable results for certain keypoints, significant issues persist:

1. **Prediction Errors**: Both models show significant inaccuracies, especially in complex poses or scenarios involving occlusion, which are critical for practical applications like sports analysis.
2. **Incomplete Skeleton Structures**: The connections between keypoints tend to be incomplete or misaligned, leading to unnatural skeletal representations. These errors limit the applicability of these models in real-world scenarios where reliability is crucial.

7 Conclusion and Future Work

The SimpleBaseline model showed higher accuracy and robustness compared to the Stacked Hourglass Network and also proved to be more efficient. However, the above results suggest that further advances in human pose estimation models are needed to meet the requirements of real-world applications. Future research should focus on bridging the gap between research performance and practical utility. By simultaneously achieving accurate predictions and real-time performance, models should be deployable in a variety of practical applications, such as sports analytics.

References

[1] Tompson, J., Goroshin, R., Jain, A., LeCun, Y., & Bregler, C. (2015). Efficient Object Localization Using Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 648–656. New York University. Retrieved from `https://arxiv.org/abs/1411.4280`.

[2] Newell, A., Yang, K., & Deng, J. (2016). Stacked Hourglass Networks for Human Pose Estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 483–499. University of Michigan, Ann Arbor. Retrieved from `https://arxiv.org/abs/1603.06937`.

[3] Xiao, B., Wu, H., & Wei, Y. (2018). Simple Baselines for Human Pose Estimation and Tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 466–481. Microsoft Research Asia and University of Electronic Science and Technology of China. Retrieved from `https://arxiv.org/abs/1804.06208`.