IJAMEC
e-ISSN: 2147-8228

INTERNATIONAL JOURNAL OF APPLIED MATHEMATICS ELECTRONICS AND COMPUTERS

www.dergipark.org.tr/ijamec

International Open Access

Volume 09
Issue 01

March, 2021

Research Article

Stacked Hourglass Network with Additional Skip Connection for Human Pose Estimation

Seung-Taek Kim [a], ***Hyo Jong Lee*** [a,*]

[a] *Division of Computer Science and Engineering, Jeonbuk National University,567 Baekje-daero, Deokjin-gu, Jeonju-si, Jeollabuk-do, South Korea*

ARTICLE INFO

Article history:
Received 1 October 2020
Accepted 11 February 2021

Keywords:
human pose estimation,
hourglass network,
deep learning

ABSTRACT

The human pose estimation is a problem of localizing human joints in a single image, and that is still a challenge in the field of computer vision. The hourglass network has been used in many researches to achieve good performance in human pose estimation problems. For human pose estimation problem, not only high-level features but also low-level features are important for understanding the whole human body. However, the vanilla hourglass network has the problem of passing only high-level features to the next stack. Therefore, we propose a network structure that can solve the problems of the vanilla hourglass by using an additional skip connection. The proposed skip connection improves network performance by passing relative low-level features to the next stack. In addition, the skip connection is a simple element-wise Sum operation, so there is no increase in the number of parameters. In this work, we use the well-known human pose estimation data set, MPII, to evaluate the proposed method. We conducted experiments to evaluate the objective performance of the proposed method, and it was confirmed through this evaluation that the proposed method improves the performance of human pose estimation of the vanilla hourglass network.

1. Introduction

The human pose estimation problem is one of challenges in the field of computer vision. The human pose is one of the key information for extracting human behavior, that used in artificial intelligence CCTV, autonomous vehicles and security system. The goal of human pose estimation problem is to localize joints from single 2D images. The traditional method estimates the pose using additional equipment (e.g., stereo camera, depth sensor, etc.). Recently, the human pose estimation problems performance has been greatly improved by development of Convolutional Neural Network (CNN) [1-3]. Nevertheless, the problem of estimating human posture is still difficult to solve due to the diversity of joints, camera angles, lighting condition, clothing and partial occlusion. Fig.1 Shows difficulty of human pose estimation.

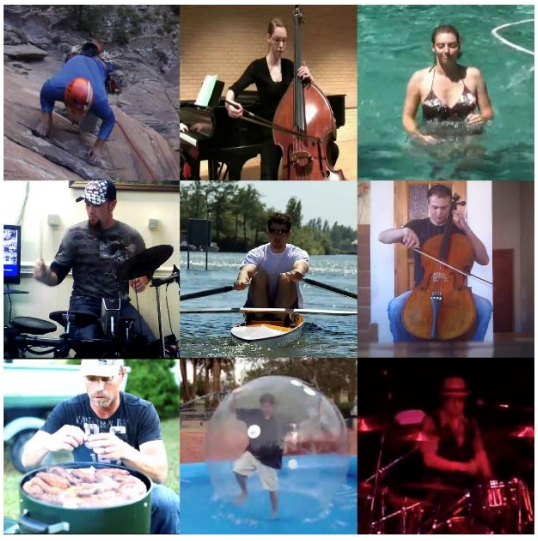

Figure 1. Images from the MPII dataset, illustrating the difficulty of human pose estimation.

* Corresponding author. E-mail address: *hlee@jbnu.ac.kr*
DOI: 10.18100/ijamec.803330

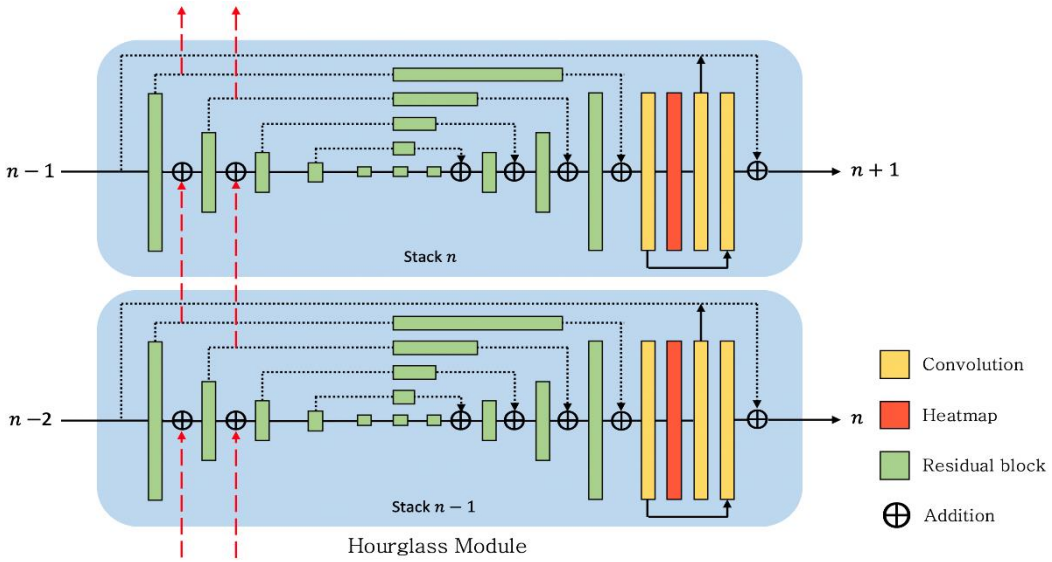

Figure 3. Proposed hourglass structure, that uses an additional skip connection (red dashed arrow in figure).

The stacked hourglass network [1] is one of well-known method for human pose estimation. It has a stacked structure of hourglass modules composed of residual blocks [10]. Since the hourglass network has performed in the human pose estimation problem, a number of studies have used it as a backbone [4-9].

In stacked hourglass network, the output of the current stack is added to the input of the current stack and used as the input of the next stack. Because of this structure, only relatively high-level features are passed to the next stack. This can be a factor that degrades network performance.

Therefore, we propose a new stacked hourglass network structure that solves the problem that only high-level features are delivered to the next stack. The proposed structure can maintain relative low-level features using additional skip connections. And that structure can improve performance without increasing the number of network parameters. We used the well-known human pose estimation dataset, MPII, for the objective evaluation of the proposed method. We confirmed through experiments that the proposed method improved the performance of the stacked hourglass network.

2. APPROACH

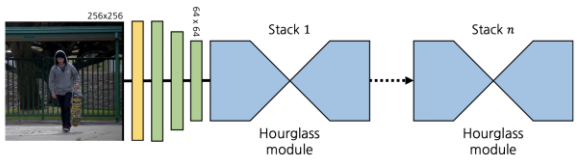

Figure 2. Vanilla Stacked hourglass network structure.

An input of the hourglass network is given an image of size 256x256. The input is reduced to 64x64 through the residual block and given as the input of the hourglass module. The hourglass module consists of a residual block with a bottleneck structure. In the encoder part of the hourglass module, the size of the feature is reduced using max-pooling, and the size is restored again using the nearest neighbour in the decoder. This structure is repeated by the stack and more accurate features are extracted. The hourglass network used in this paper is shown in Figure 2.

Each stack in the hourglass network is stacked using a skip connection. Therefore, input and output of the previous stack $(n-1)$ are added and passed to the input of the next stack (n). In this structure, only the high-level features continue to be passed to the next stack. However, low-level features are also important for the network to understand a human whole body. So, we propose a new structure of hourglass network that can maintain information of low-level features. Fig.3 is detail of the proposed new hourglass network structure. The proposed additional skip connection (Red dashed arrow in Fig.3) is located in front of the hourglass module encoder that can deliver the previous stacks low-level feature directly to next stack. This structure helps the network understand the entire human body by maintains low-level features as well as high-level features throughout the network.

3. RESULT

We use the well-known MPII [11] data set to evaluate the performance of the proposed new hourglass network. The MPII dataset contains over 40,000 images of people with joint information, of which around 25,000 images were collected in real-world contexts. For human pose estimation, 16 coordinates for each joint were labeled for each person.

In order to evaluate the performance of our method, we compare the performance with the state-of-the-art lightweight method for stacked hourglass network with various experiments. As an evaluation method, we used Percentage of Correct Keypoints head (PCKh) as used in [12]. The PCKh@0.5 uses 50% of the ground-truth head segment length as a threshold. If the error rate is lower than

the threshold value when comparing predicted value with ground-truth, it is determined to be the correct answer.

We followed the same training process as used for the original stacked-hourglass network with an input-image size of 256×256. For the data augmentation required for training, rotation ($\pm 30^{\circ}$), scaling (± 0.25), and flipping were performed. The model used in all experiments was written using PyTorch [13]. We used the Adam optimizer [14] for training and with a batch size of 8. The number of training epochs was 300, and initial learning rate was 2.5×10^{-4} , which was reduced to 2.5×10^{-5}, 2.5×10^{-6} in the 150th and 220th epochs. The network was initialized by a normal distribution $\mathcal{N}(m, \sigma^2)$ with mean m = 0 and standard deviation σ = 0.001.

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} \sum_{ij} \left\| H_n(i,j) - \hat{H}_n(i,j) \right\|^2 \qquad (1)$$

The ground-truth heat map $H = \{H_k\}_{k=1}^{K}$ was generated by applying gaussian around k body joints. The loss $\mathcal{L}$ between the heat map $\hat{H} = \{\hat{H}_k\}_{k=1}^{K}$ and H predicted by network used Mean Squared Error (MSE). Loss is calculated using the predicted heatmaps from each stack and summed up by intermediate supervision.

TABLE I
COMPARISON OF VANILLA HOURGLASS NETWORK WITH MPII VALIDATION DATASET. (DOUBLE STACKED)

Network Architecture	PCKh@0.5 (Mean)
Hourglass (Vanilla)	*87.8*
Ours	***88.7***

TABLE II
COMPARISON OF OTHER METHODS WITH MPII VALIDATION DATASET.

Method	Hea	Sho	Elb	Wri	Hip	kne	Ank	Mean
Pishchulin et al. [15]	74.3	49.0	40.8	34.1	36.5	34.4	35.2	44.1
Tompson et al. [3]	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
Carreira et al. [16]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Tompson et al. [17]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Hu et al. [18]	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Pishchulin et al. [19]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Lifshitz et al. [20]	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Gkioxary et al. [21]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Rafi et al. [22]	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Belagiannis et al. [23]	97.7	95.0	88.2	83.0	87.9	82.6	78.4	88.1
Insafutdinov et al. [24]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei et al. [25]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Newell et al. [1]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Sun et al. [26]	98.1	96.2	91.2	87.2	89.8	87.4	84.1	91.0
Ours (8 stacks)	98.1	96.4	92.1	87.7	90.3	88.1	82.5	91.1

We trained and compared the double stack hourglass network to compare it with the vanilla hourglass network. We confirmed that the proposed method improved the vanilla hourglass network by this experiment. The results of this experiment are summarized in Table I.

We compared the proposed stacked hourglass network (8 stack) with other methods. The results of this experiment are summarized in Table II. Fig.4 presents a visualization of pose estimation results for the MPII data set in the proposed 8-stack network. We confirmed from these experiments that the proposed method improved the performance of stacked the hourglass network.

Figure 3. Prediction results of proposed method for MPII dataset

4. Conclusion

In this paper, we have proposed a stacked-hourglass network with additional skip connection for human pose estimation. The vanilla stacked hourglass network delivers only the relatively high-level features, which are the outputs of each hourglass module, to the next stack. To solve this problem, we added an additional skip-connection to the hourglass module, which reflects low-level features to the next stack to improve network performance. In addition, since the added skip-connection is an elementwise-sum operation, so there is no significant effect on the computational cost and the flow of the gradient can be improved. We conducted various experiments to evaluate the experiment, and through this, we confirmed that the proposed method improved the existing hourglass network architecture.

Acknowledgement

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (GR2019R1D1A3A03103736).

Author's Note

Abstract version of this paper was presented at 9th International Conference on Advanced Technologies (ICAT'20), 10-12 August 2020, Istanbul, Turkey with the title of "Stacked Hourglass Network with Additional Skip Connection for Human Pose Estimation".

References

[1] Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." European conference on computer vision. Springer, Cham, 2016.

[2] Bulat, Adrian, and Georgios Tzimiropoulos. "Human pose estimation via convolutional part heatmap regression." European Conference on Computer Vision. Springer, Cham, 2016.

[3] Tompson, Jonathan J., et al. "Joint training of a convolutional network and a graphical model for human pose estimation." Advances in neural information processing systems. 2014.

[4] Wang, Rui, et al. "Human pose estimation with deeply learned multi-scale compositional models." IEEE Access 7 (2019): 71158-71166.

[5] Chu, Xiao, et al. "Multi-context attention for human pose estimation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[6] Peng, Xi, et al. "Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[7] Bulat, Adrian, and Yorgos Tzimiropoulos. "Hierarchical binary CNNs for landmark localization with limited resources." IEEE transactions on pattern analysis and machine intelligence (2020).

[8] Yang, Wei, et al. "Learning feature pyramids for human pose estimation." proceedings of the IEEE international conference on computer vision. 2017.

[9] Tang, Wei, and Ying Wu. "Does Learning Specific Features for Related Parts Help Human Pose Estimation?." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[10] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[11] Andriluka, Mykhaylo, et al. "2d human pose estimation: New benchmark and state of the art analysis." Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. 2014.

[12] Xiao, Bin, Haiping Wu, and Yichen Wei. "Simple baselines for human pose estimation and tracking." Proceedings of the European conference on computer vision (ECCV). 2018.

[13] Paszke, Adam, et al. "Automatic differentiation in pytorch." (2017).

[14] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[15] Pishchulin, Leonid, et al. "Strong appearance and expressive spatial models for human pose estimation." Proceedings of the IEEE international conference on Computer Vision. 2013.

[16] Carreira, Joao, et al. "Human pose estimation with iterative error feedback." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[17] Tompson, Jonathan, et al. "Efficient object localization using convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[18] Hu, Peiyun, and Deva Ramanan. "Bottom-up and top-down reasoning with hierarchical rectified gaussians." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[19] Pishchulin, Leonid, et al. "Deepcut: Joint subset partition and labeling for multi person pose estimation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[20] Lifshitz, Ita, Ethan Fetaya, and Shimon Ullman. "Human pose estimation using deep consensus voting." European Conference on Computer Vision. Springer, Cham, 2016.

[21] Gkioxari, Georgia, Alexander Toshev, and Navdeep Jaitly. "Chained predictions using convolutional neural networks." European Conference on Computer Vision. Springer, Cham, 2016.

[22] Rafi, Umer, et al. "An Efficient Convolutional Network for Human Pose Estimation." BMVC. Vol. 1. 2016.

[23] Belagiannis, Vasileios, and Andrew Zisserman. "Recurrent human pose estimation." 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017.

[24] Insafutdinov, Eldar, et al. "Deepercut: A deeper, stronger, and faster multi-person pose estimation model." European Conference on Computer Vision. Springer, Cham, 2016.

[25] Wei, Shih-En, et al. "Convolutional pose machines." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016.

[26] Sun, Ke, et al. "Human pose estimation using global and local normalization." Proceedings of the IEEE International Conference on Computer Vision. 2017.