Stacked Hourglass Networks for Human Pose Estimation (ECCV 2016)

2021 UGRP Pose Estimation Seminar

DGIST 기초학부 한현영

hyhan@dgist.ac.kr

Introduction

About Human Pose Estimation

Good pose estimation : **Be robust to occlusion and severe deformation(due to factors like clothing and lighting)**

Introduction

Existing : Using **ConvNets(main building block)** with feature(hand-crafted), graphical model

Stacked Hourglass(Proposed method) : **반복적인 pooling, upsampling 과정**과 **intermediate supervision**을 이용하여 performance(Accuracy, robust to occlusion) 향상

Character

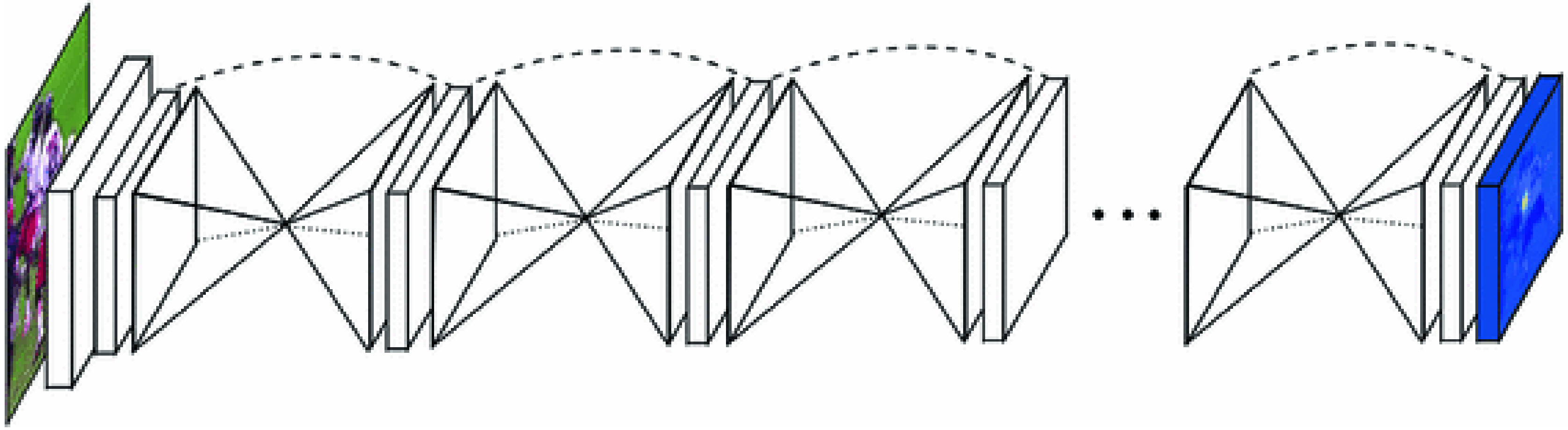
i. Network pools down to a very low resolution, then upsamples and combines features across multiple resolutions

ii. Symmetric topology

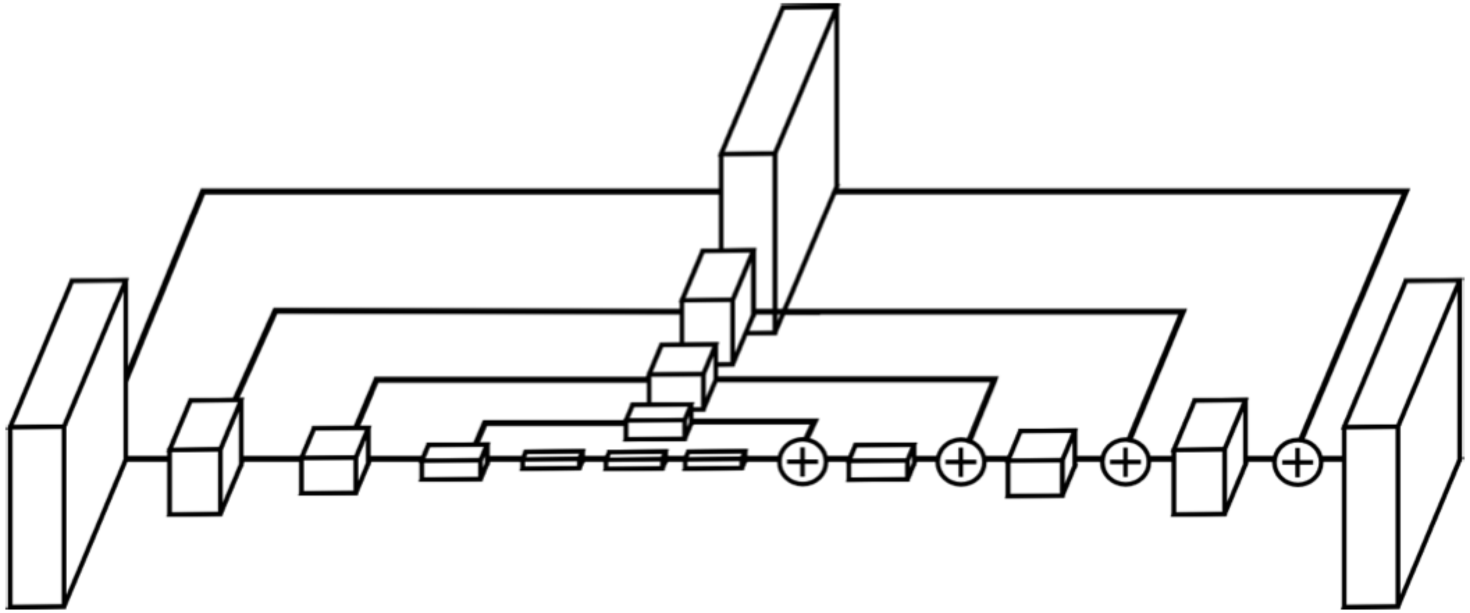
iii. Expand on a single hourglass by consecutively placing multiple hourglass modules together end-to-end

→ Allows for repeated bottom-up(pooling), top-down(upsamling) inference

iv. Conjunction : intermediate supervision; bidirectional inference



Stacked Hourglass Network



Nearest neighbor upsampling

Skip connection for top-down processing : *resolution 보존 차원*

Stacked Hourglass Network

Hourglass : simple, minimal design that has the capacity to capture all of these features and bring them together to output pixel-wise prediction

Must have some mechanism to effectively process and consolidate features across scale

→ Use single pipeline with skip layers to preserve spatial information at each resolution

==

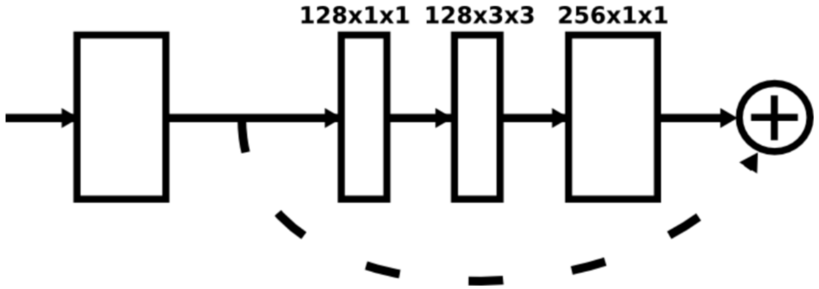
Max pooling → nearest neighbor (with skip layer)

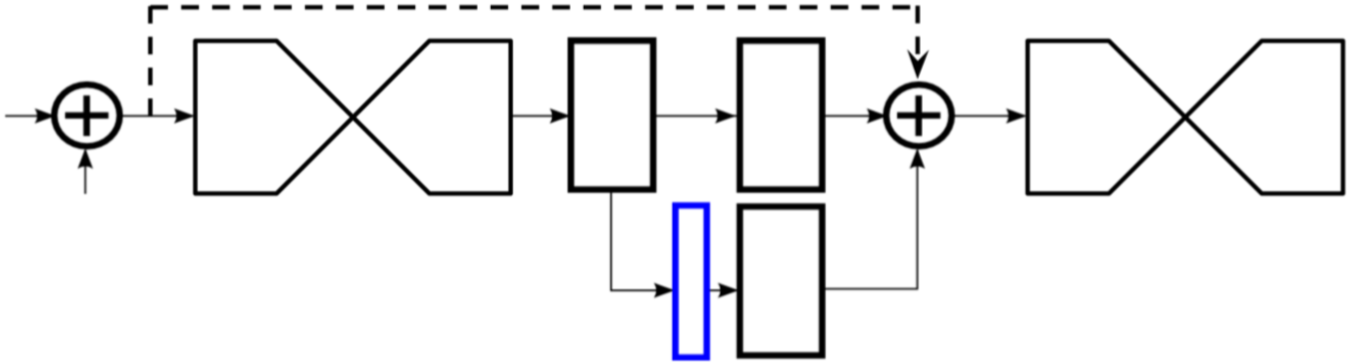
(After reaching the output resolution) 2 consecutive rounds of 1x1 convs are applied to produce the final network prediction

→ HEATMAP (the network predicts the prob of a joint’s presence)

Layer Implementation

Residual learning & “Inception” based design





Layer Implementation

반복적인 bottom-up, top-down inference를 통해 initial estimates와 이미지 전반에 대한 feature를 다시금 추정(reevaluation)할 수 있게 함

중간중간에서 얻어지는 예측값 (the prediction of intermideate heatmaps)들에 대해서도 ground truth와의 loss를 적용할 수 있음 (Intermediate Supervision)

반복적인 예측값의 조정으로 좀 더 세밀한 결과를 도출할 수 있으며, 중간중간 적용되는 loss로 인해 좀 더 깊고 안정적인 학습이 가능하리라 예상할 수 있음

Experiments

MPII, FLIC dataset

-standard Percentage of Correct Keypoints (PCK)

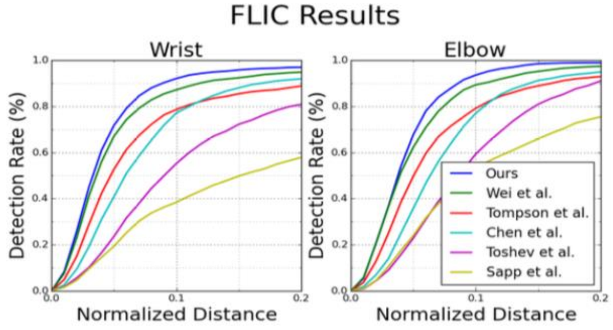
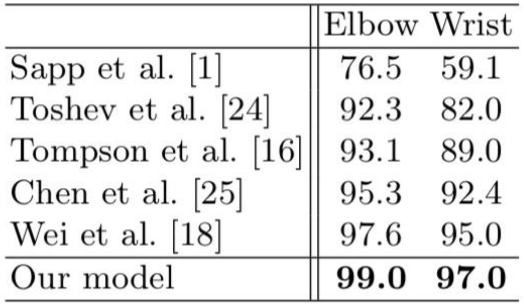


Fig. 6. PCK comparison on FLIC



	Elbow	Wrist
Sapp et al. [1]	76.5	59.1
Toshev et al. [24]	92.3	82.0
Tompson et al. [16]	93.1	89.0
Chen et al. [25]	95.3	92.4
Wei et al. [18]	97.6	95.0
Our model	**99.0**	**97.0**

Table 1. FLIC results (PCK@0.2)

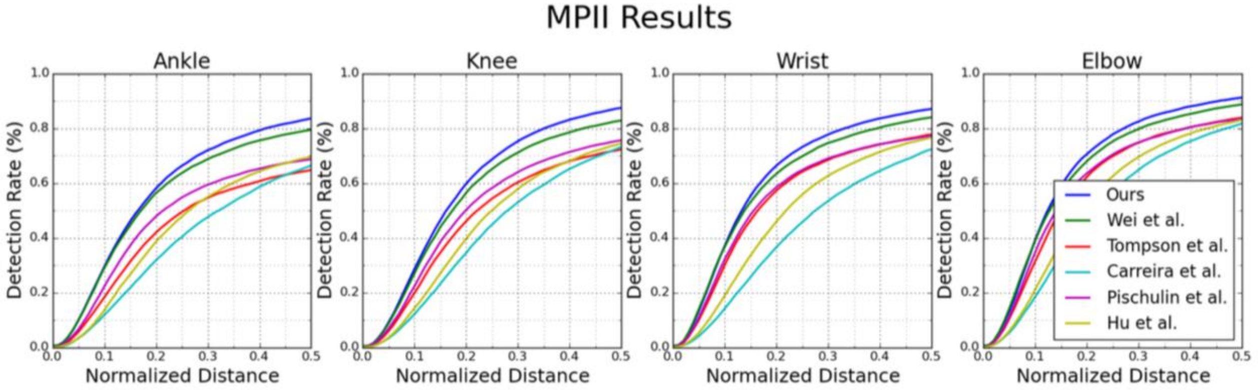
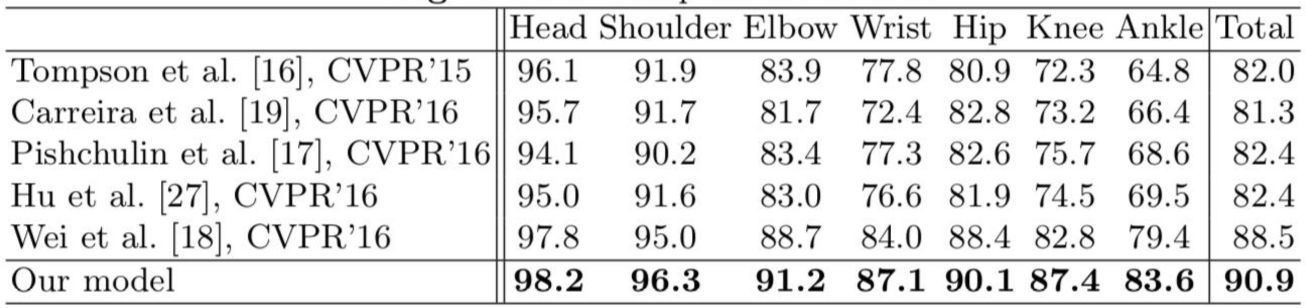


Fig. 7. PCKh comparison on MPII



	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Tompson et al. [16], CVPR'15	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Carreira et al. [19], CVPR'16	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Pishchulin et al. [17], CVPR'16	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Hu et al. [27], CVPR'16	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Wei et al. [18], CVPR'16	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Our model	**98.2**	**96.3**	**91.2**	**87.1**	**90.1**	**87.4**	**83.6**	**90.9**

Table 2. Results on MPII Human Pose (PCKh@0.5)

Experiments

네트워크가 깊어질수록 intermediate prediction accuracy 증가

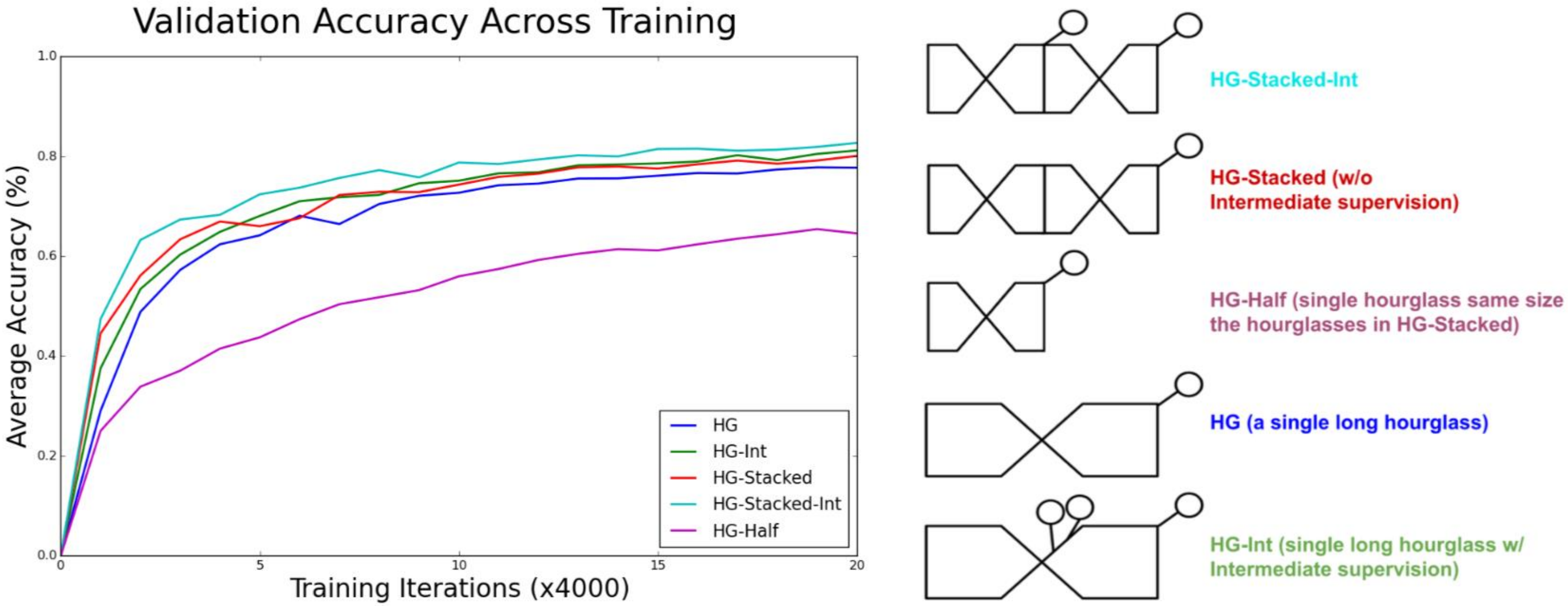
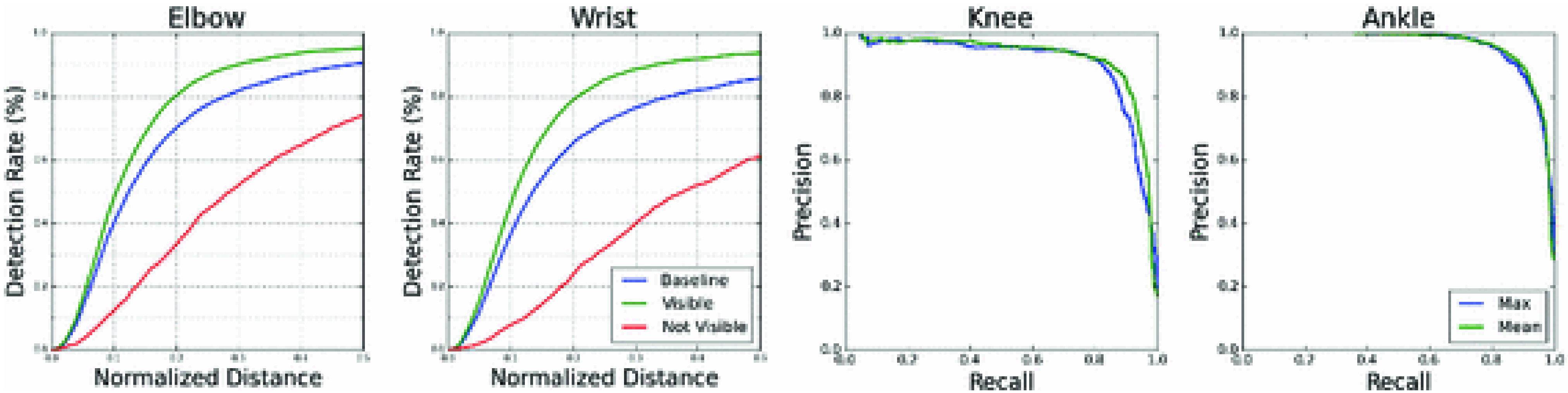


Fig. 8. Comparison of validation accuracy as training progresses. The accuracy is averaged across the wrists, elbows, knees, and ankles. The different network designs are illustrated on the right, the circle is used to indicate where a loss is applied

Occlusion

Red : Not Visible


Elbow
Detection Rate (%)
Normalized Distance
Wrist
Detection Rate (%)
Normalized Distance
Baseline
Visible
Not Visible
Knee
Precision
Recall
Ankle
Precision
Recall
Max
Mean