Deep High-Resolution Representation Learning for Human Pose Estimation (2019)

Seho Kim

https://arxiv.org/pdf/1902.09212.pdf

Introduction

- 2D human pose estimation; to localize human anatomical keypoints
  - Applications(human action recognition, human-computer interaction, animation, etc)
- High-Resolution Net(HRNet); maintain high-resolution representations
  - High-to-low resolution subnetworks(ex. Hourglass, SimpleBaseline)
  - Connect the multi-resolution subnetworks in parallel(exchanging the information); multi-scale fusions
- Benefits
  - Parallel rather than Series → more precise heatmap
  - Repeated multi-scale fusions;same depth, similar level
    → more accurate heatmap

Introduction

- Over three benchmark datasets
  - Keypoint detection: COCO keypoint detection dataset and the MPII Human Pose dataset
  - Video pose tracking: PoseTrack dataset

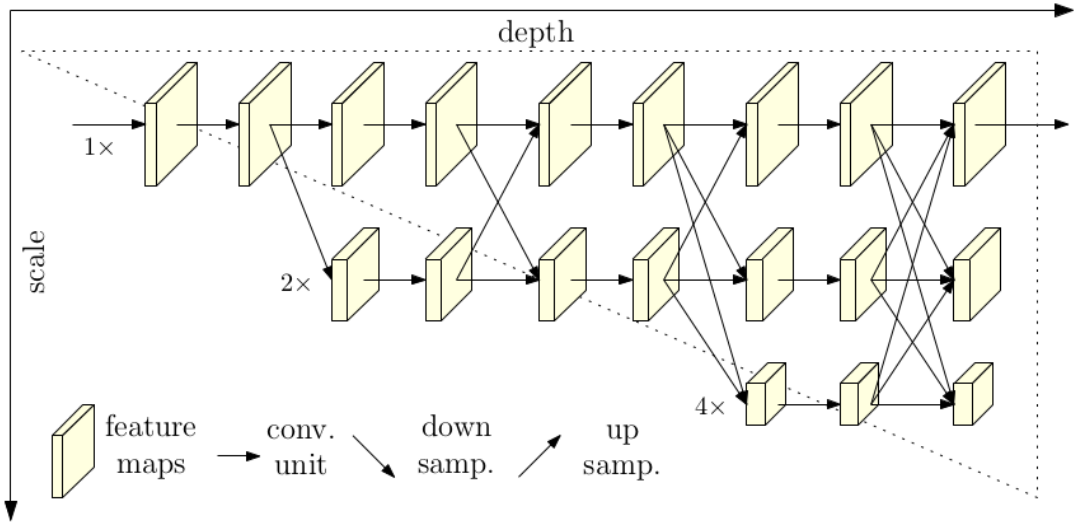

Figure 1. Illustrating the architecture of the proposed HRNet. It consists of parallel high-to-low resolution subnetworks with repeated information exchange across multi-resolution subnetworks (multi-scale fusion). The horizontal and vertical directions correspond to the depth of the network and the scale of the feature maps, respectively.

Approach

- Follow the widely-adopted pipeline
- Sequential multi-resolution subnetworks
  - Existing networks for pose estimation: connect high-to-low resolution subnetworks in series
- Parallel multi-resolution subnetworks
  - Start from a high-resolution subnetwork as the first stage
  - Gradually add high-to-low resolution subnetworks one by one
  - Connect the multi-resolution subnetworks in parallel

$$\mathcal{N}_{11} \rightarrow \mathcal{N}_{22} \rightarrow \mathcal{N}_{33} \rightarrow \mathcal{N}_{44}.$$

$$\begin{array}{ccccccc} \mathcal{N}_{11} & \rightarrow & \mathcal{N}_{21} & \rightarrow & \mathcal{N}_{31} & \rightarrow & \mathcal{N}_{41} \\ & \searrow & \mathcal{N}_{22} & \rightarrow & \mathcal{N}_{32} & \rightarrow & \mathcal{N}_{42} \\ & & & \searrow & \mathcal{N}_{33} & \rightarrow & \mathcal{N}_{43} \\ & & & & & \searrow & \mathcal{N}_{44}. \end{array}$$

Approach

- Network instantiation
  - Follow the design rule of ResNet to distribute the depth to each stage and the number of channels to each resolution
  - Four stages with four parallel subnetworks
    - Resolution is gradually decreased to a half
    - Width(the number of channels) is increased to the double
  - The first stage contains 4 residual units(the same to the ResNet-50)
  - The 2^{nd}, 3^{rd}, 4^{th} stages contain 1, 4, 3 exchange blocks, respectively; one exchange block contains 4 residual units(two 3x3 convolutions in each resolution + exchange unit)

Approach

- Repeated multi-scale fusion
  - Exchange units
    - Each output is an aggregation of the input maps (resolutions and widths of outputs are the same to the input)
    - Strided 3x3 convolutions for downsampling
    - Simple nearest neighbor sampling following a 1x1 convolution for aligning the number of channels

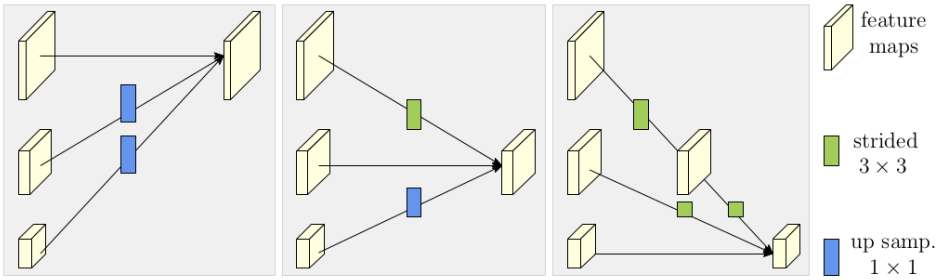

Figure 3. Illustrating how the exchange unit aggregates the information for high, medium and low resolutions from the left to the right, respectively. Right legend: strided 3×3 = strided 3×3 convolution, up samp. 1×1 = nearest neighbor up-sampling following a 1×1 convolution.

$$\begin{matrix} \mathcal{C}^1_{31} & \searrow & & \nearrow & \mathcal{C}^2_{31} & \searrow & & \nearrow & \mathcal{C}^3_{31} & \searrow & \\ \mathcal{C}^1_{32} & \rightarrow & \mathcal{E}^1_3 & \rightarrow & \mathcal{C}^2_{32} & \rightarrow & \mathcal{E}^2_3 & \rightarrow & \mathcal{C}^3_{32} & \rightarrow & \mathcal{E}^3_3, \\ \mathcal{C}^1_{33} & \nearrow & & \searrow & \mathcal{C}^2_{33} & \nearrow & & \searrow & \mathcal{C}^3_{33} & \nearrow & \end{matrix}$$

$$\mathbf{Y}_k = \sum_{i=1}^{s} a(\mathbf{X}_i, k).$$
$$\mathbf{Y}_{s+1} = a(\mathbf{Y}_s, s+1).$$

Approach

- Heatmap estimation
  - Regress the heatmaps simply from the high-resolution representations output by the last exchange unit
  - Loss function: the mean squared error; for comparing the predicted heatmaps and the groundtruth heatmaps
  - Groundtruth heatmaps: 2D Gaussian with standard deviation of 1 pixel centered on the groundtruth location of each keypoint

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Approach

- Network instantiation
  - Follow the design rule of ResNet to distribute the depth to each stage and the number of channels to each resolution
  - Four stages with four parallel subnetworks
    - Resolution is gradually decreased to a half
    - Width(the number of channels) is increased to the double
  - The first stage contains 4 residual units(the same to the ResNet-50)
  - The 2^{nd}, 3^{rd}, 4^{th} stages contain 1, 4, 3 exchange blocks, respectively; one exchange block contains 4 residual units(two 3x3 convolutions in each resolution + exchange unit)
  - Experiments: HRNet-W32(small net), HRNet-W48(big net)

Experiments

- COCO Keypoint Detection
  - Dataset
  - Evaluation metric

    Object Keypoint Similarity(OKS) = $\frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i>0)}{\sum_i \delta(v_i>0)}$.
  - Training
  - Testing
  - Results on the validation set
  - Results on the test-dev set

Experiments

- COCO Keypoint Detection

Table 1. Comparisons on the COCO validation set. Pretrain = pretrain the backbone on the ImageNet classification task. OHKM = online hard keypoints mining [11].

Method	Backbone	Pretrain	Input size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
8-stage Hourglass [40]	8-stage Hourglass	N	256 × 192	25.1M	14.3	66.9	–	–	–	–	–
CPN [11]	ResNet-50	Y	256 × 192	27.0M	6.20	68.6	–	–	–	–	–
CPN + OHKM [11]	ResNet-50	Y	256 × 192	27.0M	6.20	69.4	–	–	–	–	–
SimpleBaseline [72]	ResNet-50	Y	256 × 192	34.0M	8.90	70.4	88.6	78.3	67.1	77.2	76.3
SimpleBaseline [72]	ResNet-101	Y	256 × 192	53.0M	12.4	71.4	89.3	79.3	68.1	78.1	77.1
SimpleBaseline [72]	ResNet-152	Y	256 × 192	68.6M	15.7	72.0	89.3	79.8	68.7	78.9	77.8
HRNet-W32	HRNet-W32	N	256 × 192	28.5M	7.10	73.4	89.5	80.7	70.2	80.1	78.9
HRNet-W32	HRNet-W32	Y	256 × 192	28.5M	7.10	74.4	90.5	81.9	70.8	81.0	79.8
HRNet-W48	HRNet-W48	Y	256 × 192	63.6M	14.6	75.1	90.6	82.2	71.5	81.8	80.4
SimpleBaseline [72]	ResNet-152	Y	384 × 288	68.6M	35.6	74.3	89.6	81.1	70.5	79.7	79.7
HRNet-W32	HRNet-W32	Y	384 × 288	28.5M	16.0	75.8	90.6	82.7	71.9	82.8	81.0
HRNet-W48	HRNet-W48	Y	384 × 288	63.6M	32.9	**76.3**	**90.8**	**82.9**	**72.3**	**83.4**	**81.2**

Experiments

- COCO Keypoint Detection

Table 2. Comparisons on the COCO test-dev set. #Params and FLOPs are calculated for the pose estimation network, and those for human detection and keypoint grouping are not included.

Method	Backbone	Input size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Bottom-up: keypoint detection and grouping										
OpenPose [6]	–	–	–	–	61.8	84.9	67.5	57.1	68.2	66.5
Associative Embedding [39]	–	–	–	–	65.5	86.8	72.3	60.6	72.6	70.2
PersonLab [46]	–	–	–	–	68.7	89.0	75.4	64.1	75.5	75.4
MultiPoseNet [33]	–	–	–	–	69.6	86.3	76.6	65.0	76.3	73.5
Top-down: human detection and single-person keypoint detection										
Mask-RCNN [21]	ResNet-50-FPN	–	–	–	63.1	87.3	68.7	57.8	71.4	–
G-RMI [47]	ResNet-101	353 × 257	42.6M	57.0	64.9	85.5	71.3	62.3	70.0	69.7
Integral Pose Regression [60]	ResNet-101	256 × 256	45.0M	11.0	67.8	88.2	74.8	63.9	74.0	–
G-RMI + extra data [47]	ResNet-101	353 × 257	42.6M	57.0	68.5	87.1	75.5	65.8	73.3	73.3
CPN [11]	ResNet-Inception	384 × 288	–	–	72.1	91.4	80.0	68.7	77.2	78.5
RMPE [17]	PyraNet [77]	320 × 256	28.1M	26.7	72.3	89.2	79.1	68.0	78.6	–
CFN [25]	–	–	–	–	72.6	86.1	69.7	78.3	64.1	–
CPN (ensemble) [11]	ResNet-Inception	384 × 288	–	–	73.0	91.7	80.9	69.5	78.1	79.0
SimpleBaseline [72]	ResNet-152	384 × 288	68.6M	35.6	73.7	91.9	81.1	70.3	80.0	79.0
HRNet-W32	HRNet-W32	384 × 288	28.5M	16.0	74.9	92.5	82.8	71.3	80.9	80.1
HRNet-W48	HRNet-W48	384 × 288	63.6M	32.9	**75.5**	**92.5**	**83.3**	**71.9**	**81.5**	**80.5**
HRNet-W48 + extra data	HRNet-W48	384 × 288	63.6M	32.9	**77.0**	**92.7**	**84.5**	**73.4**	**83.1**	**82.0**

Experiments

- MPII Human Pose Estimation
  - Dataset
  - Testing
  - Evaluation metric
  - Results on the test set

Experiments

- MPII Human Pose Estimation

Table 3. Performance comparisons on the MPII test set (PCKh@0.5).

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Total
Insafutdinov et al. [27]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei et al. [69]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Bulat et al. [4]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Newell et al. [40]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Sun et al. [58]	98.1	96.2	91.2	87.2	89.8	87.4	84.1	91.0
Tang et al. [63]	97.4	96.4	92.1	87.7	90.2	87.7	84.3	91.2
Ning et al. [44]	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Luvizon et al. [37]	98.1	96.6	92.0	87.5	90.6	88.0	82.7	91.2
Chu et al. [14]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chou et al. [12]	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Chen et al. [10]	98.1	96.5	92.5	88.5	90.2	**89.6**	86.0	91.9
Yang et al. [77]	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Ke et al. [31]	98.5	96.8	92.7	88.4	90.6	89.3	**86.3**	92.1
Tang et al. [62]	98.4	**96.9**	92.6	88.7	**91.8**	89.4	86.2	**92.3**
SimpleBaseline [72]	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
HRNet-W32	**98.6**	**96.9**	**92.8**	**89.0**	91.5	89.0	85.7	**92.3**

Experiments

- Application to Pose Tracking
  - Dataset
  - Evaluation metric
  - Training
  - Testing
  - Results on the PoseTrack2017 test set

Experiments

- Application to Pose Tracking

Table 4. #Params and GFLOPs of some top-performed methods reported in Table 3. The GFLOPs is computed with the input size 256×256.

Method	#Params	GFLOPs	PCKh@0.5
Insafutdinov et al. [27]	42.6M	41.2	88.5
Newell et al. [40]	25.1M	19.1	90.9
Yang et al. [77]	28.1M	21.3	92.0
Tang et al. [62]	15.5M	15.6	92.3
SimpleBaseline [72]	68.6M	20.9	91.5
HRNet-W32	28.5M	9.5	92.3

Table 5. Results of pose tracking on the PoseTrack2017 test set.

Entry	Additional training Data	mAP	MOTA
ML-LAB [84]	COCO+MPII-Pose	70.3	41.8
SOPT-PT [53]	COCO+MPII-Pose	58.2	42.0
BUTD2 [29]	COCO	59.2	50.6
MVIG [53]	COCO+MPII-Pose	63.2	50.7
PoseFlow [53]	COCO+MPII-Pose	63.0	51.0
ProTracker [19]	COCO	59.6	51.8
HMPT [53]	COCO+MPII-Pose	63.7	51.9
JointFlow [15]	COCO	63.6	53.1
STAF [53]	COCO+MPII-Pose	70.3	53.8
MIPAL [53]	COCO	68.8	54.5
FlowTrack [72]	COCO	74.6	57.8
HRNet-W48	COCO	**74.9**	**57.9**

Experiments

- Ablation Study
  - Repeated multi-scale fusion
  - Resolution maintenance
  - Representation resolution

Table 6. Ablation study of exchange units that are used in repeated multi-scale fusion. Int. exchange across = intermediate exchange across stages, Int. exchange within = intermediate exchange within stages.

Method	Final exchange	Int. exchange across	Int. exchange within	AP
(a)	✓			70.8
(b)	✓	✓		71.9
(c)	✓	✓	✓	73.4

Experiments

- Ablation Study

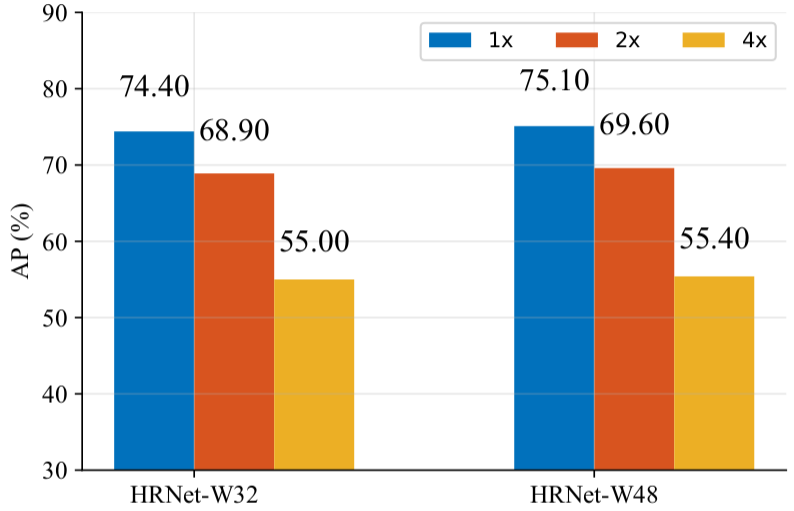

Figure 5. Ablation study of high and low representations. $1\times$, $2\times$, $4\times$ correspond to the representations of the high, medium, low resolutions, respectively.

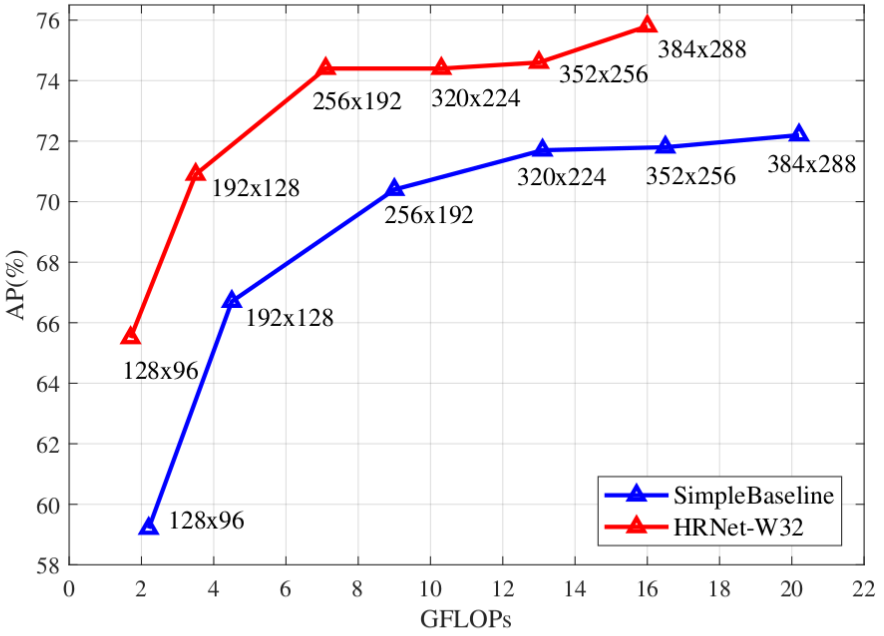

Figure 6. Illustrating how the performances of our HRNet and SimpleBaseline [72] are affected by the input size.

Conclusion and Future Works

- https://jingdongwang2017.github.io/Projects/HRNet/index.html.

Appendix

Table 7. Performance comparisons on the MPII validation set (PCKh@0.5).

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Total
Single-scale testing								
Newell et al. [40]	96.5	96.0	90.3	85.4	88.8	85.0	81.9	89.2
Yang et al. [77]	96.8	96.0	90.4	86.0	89.5	85.2	82.3	89.6
Tang et al. [62]	95.6	95.9	90.7	86.5	89.9	86.6	82.5	89.8
SimpleBaseline [72]	97.0	95.9	90.3	85.0	89.2	85.3	81.3	89.6
HRNet-W32	97.1	95.9	90.3	86.4	89.1	87.1	83.3	**90.3**
Multi-scale testing								
Newell et al. [40]	97.1	96.1	90.8	86.2	89.9	85.9	83.5	90.0
Yang et al. [77]	97.4	96.2	91.1	86.9	90.1	86.0	83.9	90.3
Tang et al. [62]	97.4	96.2	91.0	86.9	90.6	86.8	84.5	90.5
SimpleBaseline [72]	97.5	96.1	90.5	85.4	90.1	85.7	82.3	90.1
HRNet-W32	97.7	96.3	90.9	86.7	89.7	87.4	84.1	**90.8**

Table 8. Multi-person pose estimation performance (MAP) on the PoseTrack2017 dataset. "*" means models trained on thr train+valid set.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Total
PoseTrack validation set								
Girdhar et al. [19]	67.5	70.2	62.0	51.7	60.7	58.7	49.8	60.6
Xiu et al. [75]	66.7	73.3	68.3	61.1	67.5	67.0	61.3	66.5
Bin et al. [72]	81.7	83.4	80.0	72.4	75.3	74.8	67.1	76.7
HRNet-W48	82.1	83.6	80.4	73.3	75.5	75.3	68.5	**77.3**
PoseTrack test set								
Girdhar et al.* [19]	–	–	–	–	–	–	–	59.6
Xiu et al. [75]	64.9	67.5	65.0	59.0	62.5	62.8	57.9	63.0
Bin et al.* [72]	80.1	80.2	76.9	71.5	72.5	72.4	65.7	74.6
HRNet-W48*	80.1	80.2	76.9	72.0	73.4	72.5	67.0	**74.9**

Table 9. Multi-person pose tracking performance (MOTA) on the PoseTrack2017 test set."*" means models trained on the train+validation set.

Method	Head	Sho.	Elb.	Wri	Hip	Knee	Ank.	Total
Girdhar et al.* [19]	–	–	–	–	–	–	–	51.8
Xiu et al. [75]	52.0	57.4	52.8	46.6	51.0	51.2	45.3	51.0
Xiao et al.* [72]	67.3	68.5	52.3	49.3	56.8	57.2	48.6	57.8
HRNet-W48*	67.1	68.9	52.2	49.6	57.7	57.0	48.5	**57.9**