Efficient High-Resolution High-Level-Semantic Representation Learning for Human Pose Estimation

Hong Liu

Lisi Guan (Presenter)

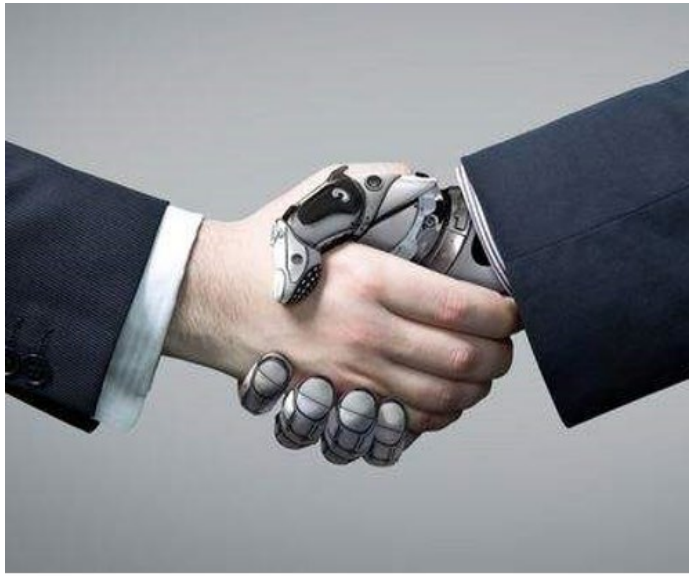
Peking University

Introduction

■ Human pose estimation & Its applications

- Human pose estimation aims at locating keypoints defined on the human body, such as the head, ankle, shoulder, etc.
-)

Intelligent Surveillance System



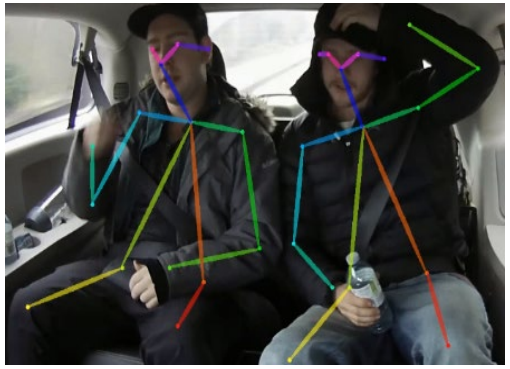
Human-Robot Interaction

Sports Video Analysis

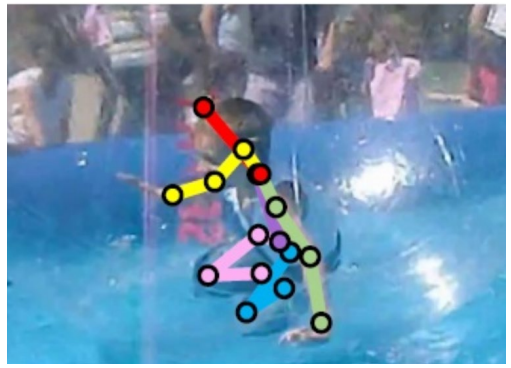
Introduction

Challenges in human pose estimation

Various Human Poses



Various Clothes Types

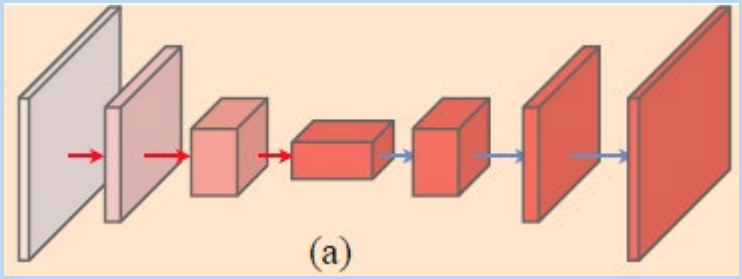


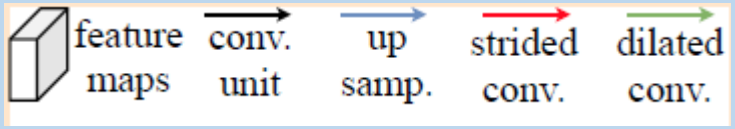
Illumination Variations

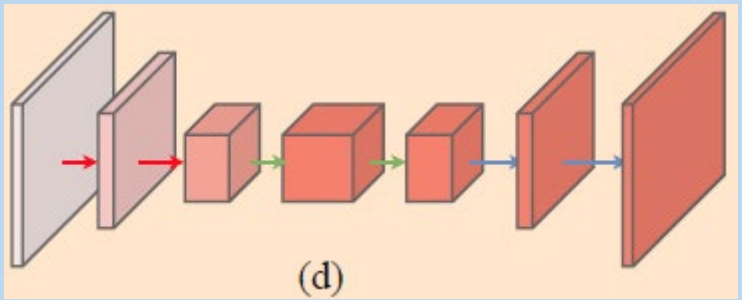
- Quantization error caused by heatmap
- Complex spatial interference caused by interactions between people

Introduction

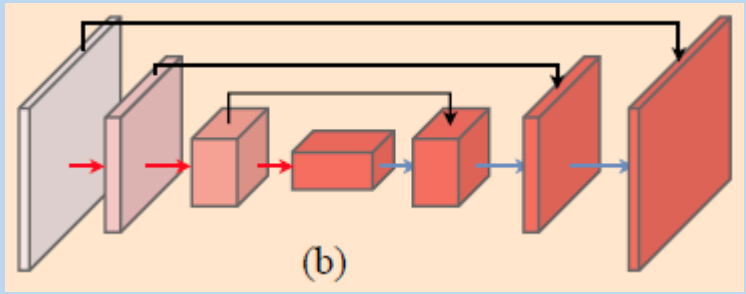
■ **High-Resolution High-Level-Sematic Representation Extraction**

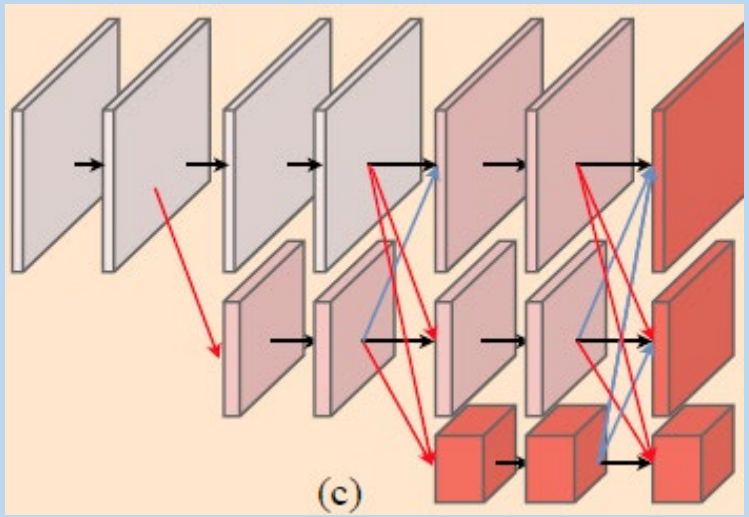






Spatial Resolution Loss





Sematic information Mismatch

Method

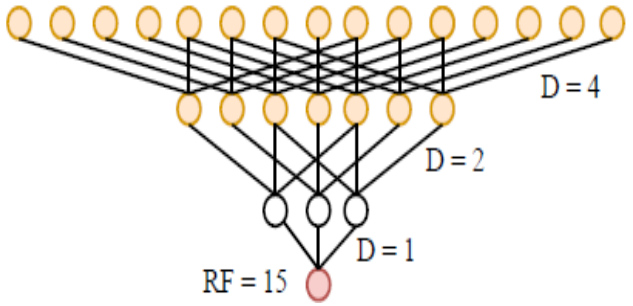
- **Dilation Pyramid Module** (DPM) can enlarge receptive fields multiplicatively without spatial information loss and semantic information mismatch. DPM is composed of N consecutive dilated convolution layers, of which dilation radius is specially designed. DPM is defined as follow:

$$X_{out} = f_N^{d_N}(f_{N-1}^{d_{N-1}}(...(f_2^{d_2}(f_1^{d_1}(X_{in})))))$$

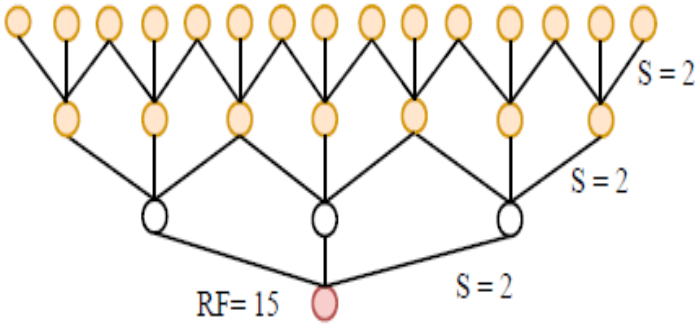
Where $f_i^{d_i}$ is the i_{th} dilated convolution layer and d_i is the dilation radius of it, d_i is determined by 2^{N-i}, X_{in} and X_{out} are input features and output features, respectively. The kernel size of all the dilated convolution of DPM is set to k x k. The receptive fields of DPM can be formulated as follows:

$$\begin{aligned} RF_{total} &= k + (k-1) * 2^1 + ... + (k-1) * 2^{N-1} \\ &= (k-1)(1 + 2^1 + ... + 2^{N-1}) + 1 \\ &= (k-1)(2^N - 1) + 1, \end{aligned}$$

where RF_{total} is the total receptive fields of DPM. DPM can enlarge receptive fields multiplicatively as subsampling and keep spatial resolution unchanged. By default, the kernel size of the dilated convolution used in DPM is set to 3 x 3 for the consideration of parameter consumption and computation cost.



DPM of three convolution neural layers



Three convolution neural layers with stride 2

Method

Dilation Pyramid Net

- **DPN** can extract high-resolution and high-level-semantic features efficiently without spatial information loss and does not need to recovery resolution by fusing features of different resolution.

- **DPN** achieves efficiency in both parameter and computation. Compared to previous advanced methods, it contains considerable fewer parameters and lower computation cost.

TABLE I: Architectures of proposed DPN.

layer name	output size	DPN4_32	DPN4_64
conv0	128x96	7x7,64,stride 2	7x7,64,stride 2
pooling	64x48	3x3 max pool,stride 2	3x3 max pool,stride 2
conv1	64x48	[1x1, 64 3x3, 64 1x1,256] ×3	[1x1, 64 3x3, 64 1x1,256] ×3
conv2	32x24	1x1,128,stride 2	[1x1,128 3x3,128 1x1,512] ×4 1x1,256,stride 2
DPM x4	32x24	[1x1,32 3x3,32 1x1,128] ×6	[1x1,64 3x3,64 1x1,256] ×6
heatmap	64x48	upsample x2 1x1,17	
#Params	-	0.7 M	3.3 M

Ablation Study

- **Dataset: MS COCO dataset**

TABLE II: Comparison between normal convolution and DPM. DPN1 means that the model contains one DPM module. Normal means that the dilation radiuses of all the dilated convolution layers are set to one. So all the compared methods consume the same parameters and computation cost. AP is average precision.

		DPN1_64	DPN2_64	DPN3_64	DPN4_64
AP	DPM	**64.5**$_{\uparrow 4.8}$	**68.0**$_{\uparrow 4.0}$	**69.2**$_{\uparrow 4.9}$	**69.9**$_{\uparrow 5.3}$
	Normal	59.7	64.0	64.3	64.6

TABLE III: Comparison with Stacked Dilated Convolution. AP is average precision. D means that the dilation radius of all the dilated convolution layers in DPM is set to D.

	DPM	D=1	D=2	D=4	D=6	D=8
DPN1_32	**50.6**	40.1	**49.3**	48.7	43.9	38.5
DPN4_32	**62.0**	58.2	**60.6**	56.6	50.0	48.7

lation Study

- **Dataset: MS COCO**

TABLE IV: Comparison between high-resolution representation and medium-resolution representation. AP is average precision. DPN1 means that model contains one DPM module. Resolution is spatial resolution of input features to DPM module. The input image size is 256×192. Here the high-resolution model is trained for 210 epochs.

	Epochs	Resolution	DPN1_64	DPN2_64	DPN3_64	DPN4_64
AP	210	64×48	65.5	68.7	70.3	71.1
	140	32×24	**65.2**	**68.0**	**69.2**	**69.9**
FLOPs	210	64×48	6.0 G	7.2 G	8.4 G	9.6 G
	140	32×24	**2.1** G	**2.4** G	**2.7** G	**3.0** G

TABLE V: Comparison between reversed DPM (RDPM) and DPM. DPN1 means that the model contains one DPM module. The dilation radius of RDPM is defined by 2^{i-1}, so the dilation radius scheme of RDPM is increasing. AP is of average precision.

		DPN1_32	DPN2_32	DPN3_32	DPN4_32
AP	DPM	**50.6**$_{\uparrow 1.6}$	56.9$_{\uparrow 0.4}$	60.0$_{\uparrow 0.4}$	62.0$_{\uparrow 0.4}$
	RDPM	**49.0**	56.6	59.6	61.6

Compare with state-of-the-arts

TABLE VI: Compared with the state-of-the-art methods on COCO val2017 dataset. AP is average precision. Resolution of input image is 256×192.

Method	Backbone	#Params	GFLOPs	AP
OpenPose [25]	-	-	-	61.8
Mask-RCNN [26]	ResNet-50	-	-	63.1
Hourglass [5]	8-stage Hourglass	25.1 M	14.3	66.9
CPN [6]	ResNet-50	27.0 M	6.2	69.4
SBL [8]	ResNet-50	34.0 M	8.9	70.4
HRNet-w32 [7]	HRNet-w32	28.5 M	7.1	74.4
HRNet-w48 [7]	HRNet-w48	**63.6** M	**14.6**	**75.1**
DPN4_32 (ours)	DPN	**0.7** M	**1.1**	**62.0**
DPN4_64 (ours)	DPN	**3.3** M	**3.0**	**69.9**
DPN2_hrnet (ours)	HRNet-w32	29.5 M	7.8	**75.0**

Compare with state-of-the-arts

TABLE VII: Compared with the state-of-the-art methods on MPII test set (PCKh@0:5). Resolution of input image is 256 × 256.

Method	#Patams	GFLOPs	PCKh@0.5
Deepercut [9]	42.6 M	41.2	88.5
SBL [8]	68.6 M	20.9	91.5
MCA [27]	58 M	128	91.5
JADA [28]	26.0 M	55.0	91.5
LFP [29]	28.0 M	46.0	92.0
Deeply [30]	15.5 M	**15.6**	92.3
PIL [31]	**26.0** M	63.0	92.4
DPN4_32 (ours)	**0.7** M	**1.4**	**86.0**
DPN4_64 (ours)	**3.3** M	**4.0**	**88.6**

Contributions & Conclusions

- To extract high-resolution high-level-semantic features directly and effectively, we propose a novel dilation pyramid module (DPM), which can directly extract high-resolution high-level-semantic features without spatial information loss and semantic information mismatch.
- Based on DPM, we further propose an efficient and effective dilation pyramid network (DPN), which can achieve comparable performance to state-of-the-art methods. As a result, with DPN, computation cost and parameter consumption can be reduced considerably.
- Extensive experiments demonstrate the superior keypoint detection performance over two benchmark datasets: the COCO keypoint detection dataset and the MPII Human Pose dataset.

Thanks for listening!