CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes

Yuhong Li[1,2], Xiaofan Zhang[1], Deming Chen[1]

[1]ECE, University of Illinois at Urbana-Champaign, [2]Beijing University of Posts and Telecommunications

center for cognitive computing systems research

For full text ↑

CVPR 2018 SALT LAKE CITY • JUNE 18-22

Motivations

Major tasks for understanding the highly congested scenes

1. Count the crowd → get the number

2. Generate density map → get the crowd distribution

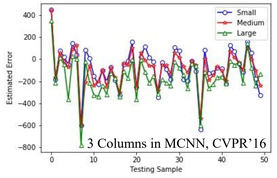
3. Understand the potential risks and make decisions

- Total crowd number, Density patterns, Change rate, etc.

Drawbacks of the state-of-the-art

E.g., MCNN, CVPR'16
Crowdnet, ACMMM'16
SwitchingCNN, CVPR'17

- Use multi-column (MCNN) structures (branches/columns for different levels of density)
- Different columns are likely to learn similar features



Method	Parameters	MAE	MSE
Col. 1 of MCNN	57.75k	141.2	206.8
Col. 2 of MCNN	45.99k	160.5	239.0
Col. 3 of MCNN	25.14k	153.7	230.2
MCNN Total	127.68k	110.2	185.9
A deeper CNN	83.84k	**93.0**	**142.2**

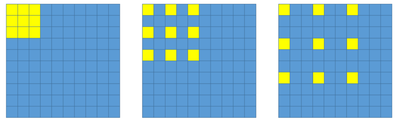
A deeper CNN (w/o branches) performs even better with less parameters

- MCNN-like structures require more efforts to train
- Hard to deploy on IoT devices (larger size with more parameters)

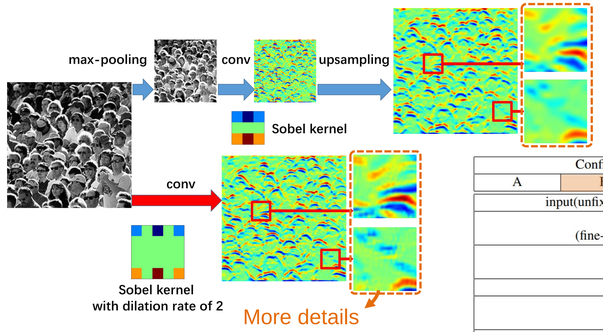
Solution: CSRNet

An end-to-end single-column structure delivering the best performance

- Front-end: CNN → Extract 2D features
- Back-end: Dilated CNN → Larger reception fields without losing accuracy (compared to pooling)



Advantage of using dilated CONV instead of pooling



For the same output size, dilated CONV keeps more detailed information

CSRNet Network configuration

conv(kernel size)-(# of filters)-(dilation rate)

max-pooling layers are conducted over a 2×2 pixel window with stride 2.

CSRNet-B is the best configuration

Architecture	MAE	MSE
CSRNet A	69.7	116.0
CSRNet B	**68.2**	**115.0**
CSRNet C	71.91	120.58
CSRNet D	75.81	120.82

Evaluation in ShanghaiTech Part A

Configurations of CSRNet			
A	B	C	D
input(unfixed-resolution color image)			
front-end (fine-tuned from VGG-16)			
conv3-64-1 conv3-64-1			
max-pooling			
conv3-128-1 conv3-128-1			
max-pooling			
conv3-256-1 conv3-256-1 conv3-256-1			
max-pooling			
conv3-512-1 conv3-512-1 conv3-512-1			
back-end (four different configurations)			
conv3-512-1 conv3-512-1 conv3-512-1 conv3-256-1 conv3-128-1 conv3-64-1	conv3-512-2 conv3-512-2 conv3-512-2 conv3-256-2 conv3-128-2 conv3-64-2	conv3-512-2 conv3-512-2 conv3-512-2 conv3-256-4 conv3-128-4 conv3-64-4	conv3-512-4 conv3-512-4 conv3-512-4 conv3-256-4 conv3-128-4 conv3-64-4
conv1-1-1			

Results

We demonstrate CSRNet on **ShanghaiTech**, **UCF CC 50**, **WorldExpo'10**, **UCSD** datasets for crowd counting, and **TRANCOS** dataset for car counting.

Results in ShanghaiTech dataset

	Part_A		Part_B	
Method	MAE	MSE	MAE	MSE
Zhang *et al.* [3]	181.8	277.7	32.0	49.8
Marsden *et al.* [38]	126.5	173.5	23.8	33.1
MCNN [18]	110.2	173.2	26.4	41.3
Cascaded-MTL [39]	101.3	152.4	20.0	31.1
Switching-CNN [4]	90.4	135.0	21.6	33.4
CP-CNN [5]	73.6	**106.4**	20.1	30.1
CSRNet (ours)	**68.2**	115.0	**10.6**	**16.0**

Results in UCF CC 50 dataset

Method	MAE	MSE
Idrees *et al.* [22]	419.5	541.6
Zhang *et al.* [3]	467.0	498.5
MCNN [18]	377.6	509.1
Marsden *et al.* [38]	338.6	424.5
Cascaded-MTL [39]	322.8	397.9
Switching-CNN [4]	318.1	439.2
CP-CNN [5]	295.8	**320.9**
CSRNet (ours)	**266.1**	397.5

Results in WorldExpo'10 dataset

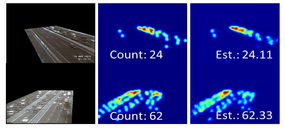
Method	Sce.1	Sce.2	Sce.3	Sce.4	Sce.5	Avg.
Chen *et al.* [46]	2.1	55.9	9.6	11.3	3.4	16.5
Zhang *et al.* [3]	9.8	14.1	14.3	22.2	3.7	12.9
MCNN [18]	3.4	20.6	12.9	13.0	8.1	11.6
Shang *et al.* [37]	7.8	15.4	14.9	11.8	5.8	11.7
Switching-CNN [4]	4.4	15.7	10.0	11.0	5.9	9.4
CP-CNN [5]	2.9	14.7	10.5	**10.4**	5.8	8.86
CSRNet (ours)	**2.9**	**11.5**	**8.6**	16.6	**3.4**	**8.6**

Results in UCSD dataset

Method	MAE	MSE
Zhang *et al.* [3]	1.60	3.31
CCNN [20] CCNN	1.51	-
Switching-CNN [4]	1.62	2.10
FCN-rLSTM [24]	1.54	3.02
CSRNet (ours)	1.16	1.47
MCNN [18]	**1.07**	**1.35**

Results in TRANCOS dataset

Method	GAME 0	GAME 1	GAME 2	GAME 3
Fiaschi *et al.* [47]	17.77	20.14	23.65	25.99
Lempitsky *et al.* [32]	13.76	16.72	20.72	24.36
Hydra-3s [20]	10.99	13.75	16.69	19.32
FCN-HA [24]	4.21	-	-	-
CSRNet (Ours)	**3.56**	**5.49**	**8.57**	**15.04**



Source code: https://github.com/leeyeehoo/CSRNet

Conclusions

- **The single-column structure is more efficient than MCNNs**
  Less complexity & parameters, easier to train & reproduction
- **We adapt dilated CONV to crowd counting applications**
  Get larger receptive field without losing details
- **We deliver the state-of-the-art performance in five datasets**

Acknowledgement

This work was supported by the IBM-Illinois Center for Cognitive Computing System Research (C[3]SR) - a research collaboration as part of the IBM AI Horizons Network.