DecideNet: Counting Varying Density Crowds Through Attention Guided Detection and Density Estimation.

PAMI Meeting

BUGINGO EMMANUEL[1]

[1] School of information Science and Engineering
Department of Computer Science
Lab 301 Cloud Computing and Big Data

3^{rd} August 2018

What problem being solved ?

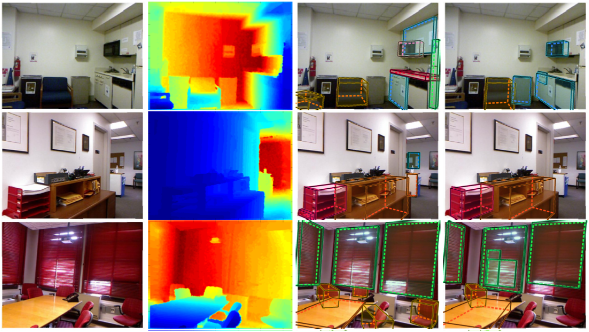
what problem being solved

Models

what problem being solved

- **Given an image**, Crowd counting provide a number of people in that image.
- Different approach have been proposed to solve this problem. However all of them can be classified into two categories : **Detection based crowd counting and Regression based crowd counting.**
- **Detection based crowd count approaches :** Use Object detectors to localize the position of each person. **Better for uncrowded patches**
- **Regression based crowd count approaches :** density map of image patches. **Better for crowded patches**
- *Can crowd counting exclusively based on either regression or detection be enough to simultaneously handle high and low density scene ?*

Why this is an issue ?



Why do we care ?, what impact ?

- **Crowd count :** is important for high level crowd analysis like : crowd monitoring, Scene understanding,

Why this is an issue ?

Why do we care ?, what impact ?

- **Crowd count :** public safety management.

Big Challenge

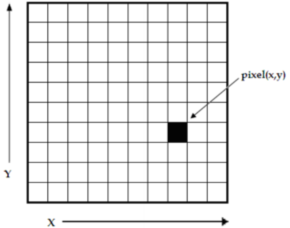
- **Main issue :** the density varies spatially and temporally, each category is better for a certain density.

Solutions

- **Regression based crowd count approach** Can totally find the number of people in a given area.
- **Detection based crowd count approach** Can also do the same job.

Some Definitions

Pixel :



Density at a specific pixel on a given image

$$\forall p \in I_i, D_i^{gt}(p|I_i) = \sum_{P \in \mathbf{P}_i^{gt}} \mathcal{N}^{gt}(p|\mu = P, \sigma^2).$$

Total person count $\sum_{p \in I_i} D_i^{gt}(p|I_i) = c_i.$

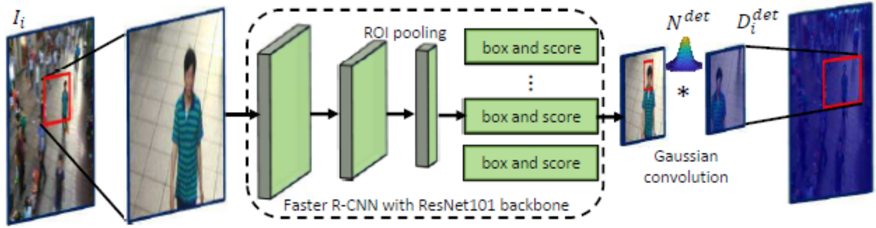
Ground trough : x and y that defines the position of the detected object on the picture

Density at a specific pixel on a given image considering the effects from all the Gaussian functions centered by annotation points.

Total person count : Summing over the density values of all pixels over the entire image.

Ω : Is a parameter that is used to minimize the difference between the prediction density map $D_i^{out}(p|I_i)$ and the ground-truth $D_i^{gt}(p|I_i)$ by learning a non-linear mapping for I_i.

Detailed method : DetectionNet



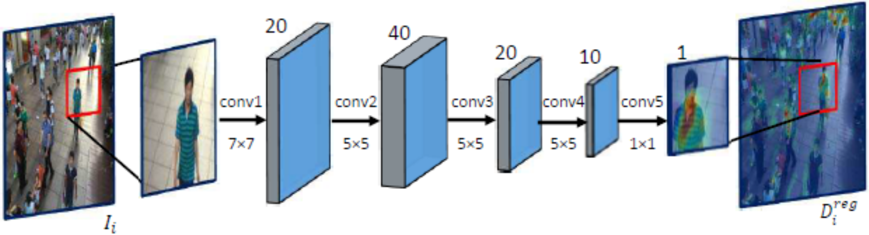
Detection based density map

$$D_i^{det}(p|\Omega_{det}, I_i) = \sum_{P \in \mathbf{P}_i^{det}} \mathcal{N}^{det}(p|\mu = P, \sigma^2).$$

Loss for Detection base

$$L_{det} = \frac{1}{N} \sum_i \left[L_{cls}(\mathbf{P}_i^{det}, \mathbf{P}_i^{gt}|\Omega_{det}) + L_{loc}(\mathbf{P}_i^{det}, \mathbf{P}_i^{gt}|\Omega_{det}) \right]$$

Detailed method continue : RegressionNet



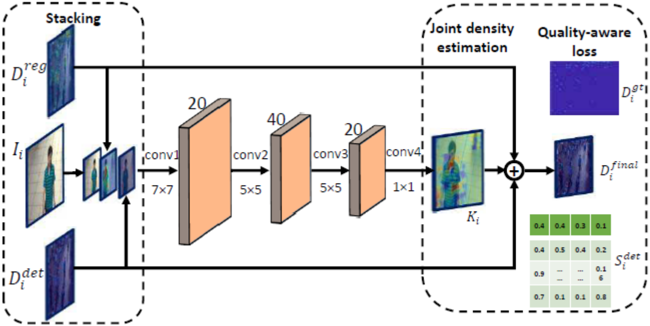
Estimated cloud density map for all pixels

$$\mathcal{F}^{reg}(I_i|\Omega_{reg}) = D_i^{reg}(p|\Omega_{reg}, I_i).$$

Loss for based base

$$L_{reg} = \frac{1}{N}\sum_{i}\sum_{p\in I_i}\left[D_i^{reg}(p|\Omega_{reg}, I_i) - D_i^{gt}(p|I_i)\right]^2.$$

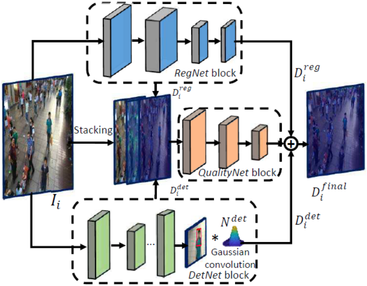
Detailed method continue : QualityNet



$$L_{qua} = \frac{1}{N} \sum_{i} \sum_{p \in I_i} \left\{ \left[D_i^{final}(p|\Omega_{qua}, I_i) - D_i^{gt}(p|I_i) \right]^2 + \lambda \| K_i(p|\Omega_{qua}, I_i) - S^{det}(p|I_i) \|^2 \right\},$$

It receive as input : image I_i, 2 density map from detection and regression and **It outputs a probabilistic attention map** $K_i(P|\Omega_{qua}, I_i)$

Detailed method continue :Model DecideNet



$$D_i^{final}(p|I_i) = K_i(p|\Omega_{qua}, I_i) \odot D_i^{det}(p|\Omega_{det}, I_i) + (\mathbf{J} - K_i(p|\Omega_{qua}, I_i)) \odot D_i^{reg}(p|\Omega_{reg}, I_i),$$

$$L_{decide} = L_{reg} + L_{det} + L_{qua},$$

Results

Evaluation settings : 40k steps of iteration, initial learning rate 0.005. each 10k cut LR by half, Images are cropped into 4x3 patches.

1

Method	MAE	MSE
SquareChn Detector	20.55	439.1
R-FCN	6.02	5.46
Faster R-CNN	5.91	6.60
Count Forest	4.40	2.40
Exemplary Density	1.82	2.74
Boosting CNN	2.01	N/A
MoCNN	2.75	13.40
Weighted VLAD	2.41	9.12
DecideNet	**1.52**	**1.90**

2

Method	MAE	MSE
R-FCN	52.35	70.12
Faster R-CNN	44.51	53.22
Cross-scene	32.00	49.80
M-CNN	26.40	41.30
FCN	23.76	33.12
Switching-CNN	21.60	33.40
CP-CNN	**20.1**	30.1
DecideNet	21.53	31.98
DecideNet+R3	20.75	**29.42**

3

Method	MAE					
	S1	S2	S3	S4	S5	Ave
Cross-scene	**2.00**	29.50	9.70	**9.30**	**3.10**	12.90
M-CNN	3.40	20.60	12.90	13.00	8.10	11.60
Local&Global	7.80	15.40	15.30	25.60	4.10	11.70
CNN-pixel	2.90	18.60	14.10	24.60	6.90	13.40
Switching-CNN	4.40	15.70	10.00	11.00	5.90	9.40
DecideNet	**2.00**	**13.14**	**8.90**	17.40	4.75	**9.23**

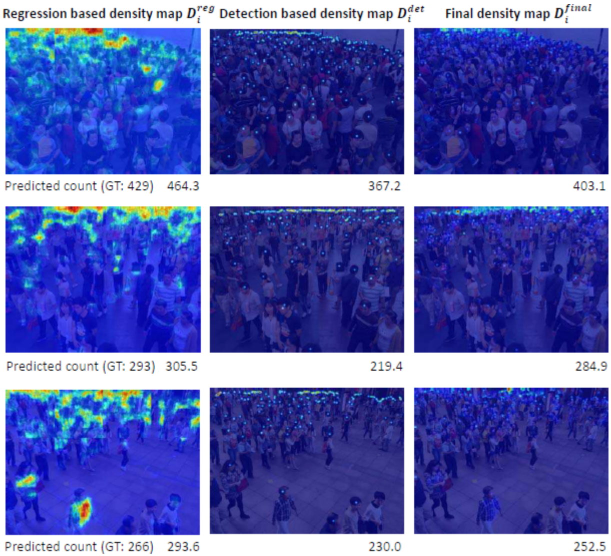
4

Method	MAE		MSE	
	Mall	SHB	Mall	SHB
RegNet only	3.37	42.85	4.22	63.63
DetNet only	4.50	44.90	5.60	73.18
RegNet+DetNet (Late Fusion)	3.93	38.63	4.96	65.27
RegNet+DetNet+QualityNet	1.83	24.93	2.27	41.86
RegNet+DetNet+QualityNet (quality-aware loss)	**1.52**	**21.53**	**1.90**	**31.98**

Keys

- Dataset(1.Mall,2. ShanghaiTech PartB, 3.WorldExpo)
- 4. Qualitative results(Dataset 1 and 2)
- **Accuracy of the algorithm in estimating** MAE : Mean Absolute Error
- **Metrics that indicates the robustness** MSE : Mean Squared Error of estimation

Results continue



Conclusion and Extension

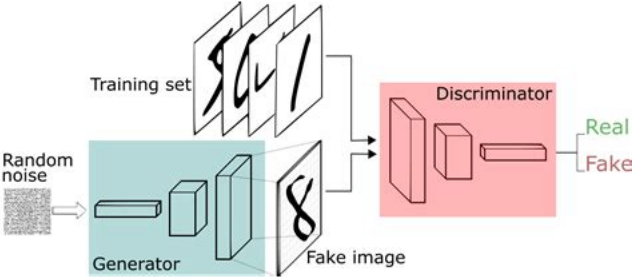
Conclusion

- The author of this paper has considered the advantage of the two category of approach in count the number of people in the crowd.
- They have proposed an architecture that combines **Detection based crowd count advantages and those of Regression based crowd count approaches.**
- They claim the proposed architecture to be the first framework that uses both regression and Detection at the same time.

Conclusion and Extension

Extension

Use **Generative Adversarial Network** and **Retrain detector** for not only detecting head but also other features that can represent a human.



Q and A