PHNet: Parasite-Host Network for Video Crowd Counting

Shiqiao Meng, Jiajie Li, Weiwei Guo, Lai Ye and Jinfeng Jiang
{*mengshiqiao*, *Jiajie_Li*, *jjf*}@alumni.tongji.edu.cn, {*weiweiguo*, *yelai*}@tongji.edu.cn

Problem

Many crowd counting methods for a single image have been proposed, but few studies have focused on using temporal information from image sequences of videos to improve prediction performanc. We propose a Parasite-Host Network (PHNet) which is composed of Parasite branch and Host branch to extract temporal features and spatial features respectively. To specifically extract the transform features in the time domain, we propose a novel architecture termed as "Relational Extractor"(RE) which models the multiplicative interaction features of adjacent frames.

Theoretical Basis

Let the pixels on two adjacent frames be $F_i^{t-1}$and F_j^t. Our goal is to learn the transformation relationship R between them. The way to model the transform relationship between two images is to use multiplicative interactions. The equation is given as follows:

$$R_k = \sum_{ij} \omega_{ijk} F_i^{t-1} F_j^t \tag{1}$$

where R_k denotes the multiplicative interactions between images. F_j^{t-1} and F_j^t represent the pixel at location i on the $t-1^{th}$ frame and the pixel at location j on the t^{th} frame. ω_{ijk} are parameters that can fit the multiplicative relationship. We factorized this equation and got the factorized relation model, which can be expressed as:

$$R_k = \sum_f \omega_{kf}^R \sum_i \omega_{ik}^{t-1} F_i^{t-1} \sum_j \omega_{jk}^t F_j^t \tag{2}$$

where the meaning of the symbols here is the same as that of Equation Since the Equation is difficult to implement simply with neural networks, we need to transform it into other forms that are easy to implement. So we model multiplicative interactions by implementing energy model, which can be expressed as

$$\begin{aligned} R_k &= \sum_f \omega_{kf}^R (\omega_{\mathbf{ik}}^{\mathbf{t-1}^{\mathrm{T}}} F_i^{t-1} + \omega_{\mathbf{jk}}^{\mathbf{t}^{\mathrm{T}}} F_j^t)^2 \\ &= \sum_f \omega_{kf}^R \Big[2(\omega_{\mathbf{ik}}^{\mathbf{t-1}^{\mathrm{T}}} F_i^{t-1})(\omega_{\mathbf{jk}}^{\mathbf{t}^{\mathrm{T}}} F_j^t) + \\ &\quad (\omega_{\mathbf{ik}}^{\mathbf{t-1}^{\mathrm{T}}} F_i^{t-1})^2 + (\omega_{\mathbf{jk}}^{\mathbf{t}^{\mathrm{T}}} F_j^t)^2 \Big] \end{aligned} \tag{3}$$

Relation Extractor

According to the theories in the previous subsection, we designed a Relation Extractor(RE) to model temporal relation and extract high-dimensional features. In the first layer of the neural network, using 3D convolution can establish a connection between adjacent frames in channel level, and then using a square operation followed by a multi-layer convolution operation enables the network to learn the parameters in the Equation (3).

The structure of the PHNet for video crowd counting

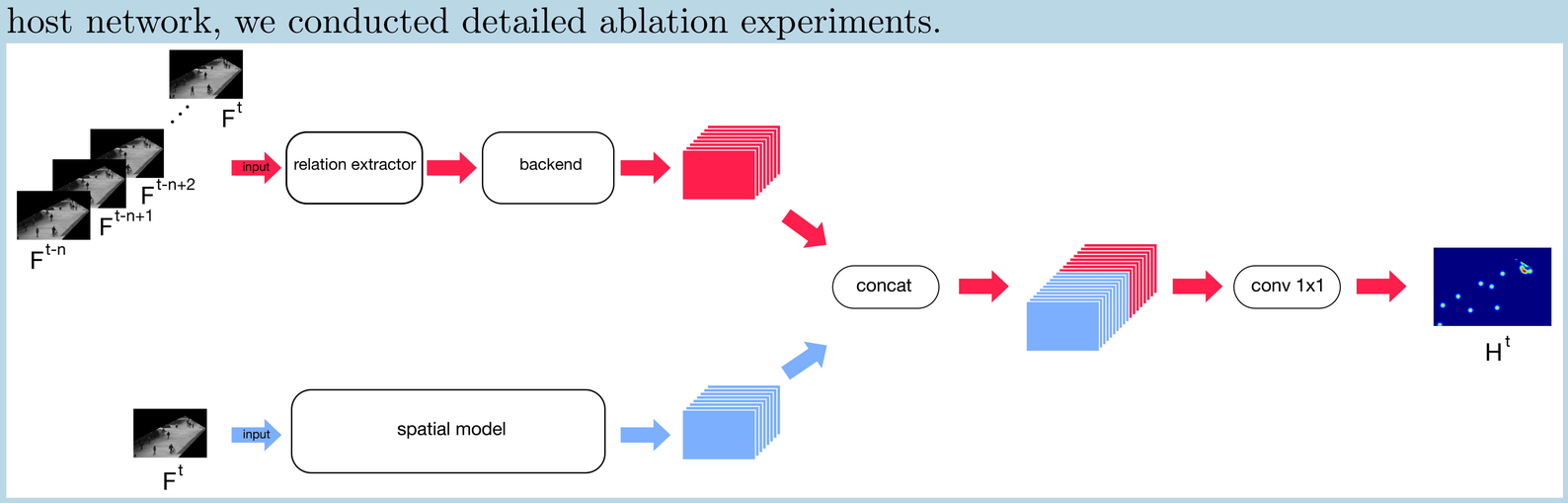
The overall structure of PHNet is shown in the figure below. This model contains two branches, which are used to extract temporal information and spatial information respectively. The spatial model part can be replaced by any crowd density estimate model which uses a single image to predict. The part of extracting time information is completed by RE module, which is also one of our main contributions.

Parasitic Branch: appearance-independent temporal branch

Parasitic temporal branch first utilize RE to learn appearance-independent relation between frames. Then, we deploy the backend using multiple dilated convolution layers, which enlarge receptive fields and extract deeper features without losing resolutions.

Host Branch: alternative spatial branch

As demonstrated earlier, the RE can effectively extract the branch focusing on temporal information. And as an easy-to-deploy fully convolutional network branch, this branch can easily upgrade any network model used for static crowd counting into a spatial-temporal model. For the selection of the host network, we conducted detailed ablation experiments.



Experiments

Since PHNet is focus on video data, we conduct comparative experiments using four annotated crowd counting datasets which include the UCSD dataset, Venice dataset, FDST dataset, and CrowdFlow dataset. We use MAE (mean absolute error) and RMSE (root mean squared error) to evaluate the performance of the model. The results are shown as follow, it can be seen that our effect has reached the state-of-the-art. And the visualization of the comparison of PHNet and CSRNet is shown as follows:

ESTIMATION ERRORS ON THE UCSD DATASET

Method	Venue	Year	MAE	RMSE
Switch-CNN	CVPR	2017	1.62	2.10
Zhang et al.	CVPR	2015	1.60	3.31
ConvLSTM	ICCV	2017	1.30	1.79
CSRNet	CVPR	2018	1.16	1.47
Bi-ConvLSTM	ICCV	2017	1.13	1.43
MCNN	CVPR	2016	1.07	1.35
SANet	ECCV	2018	1.02	1.29
ADCrowdNet	CVPR	2019	0.98	1.25
PACNN	CVPR	2019	0.89	1.18
PHNet(ours)	-	-	**0.82**	**1.05**

ESTIMATION ERRORS ON THE VENICE DATASET

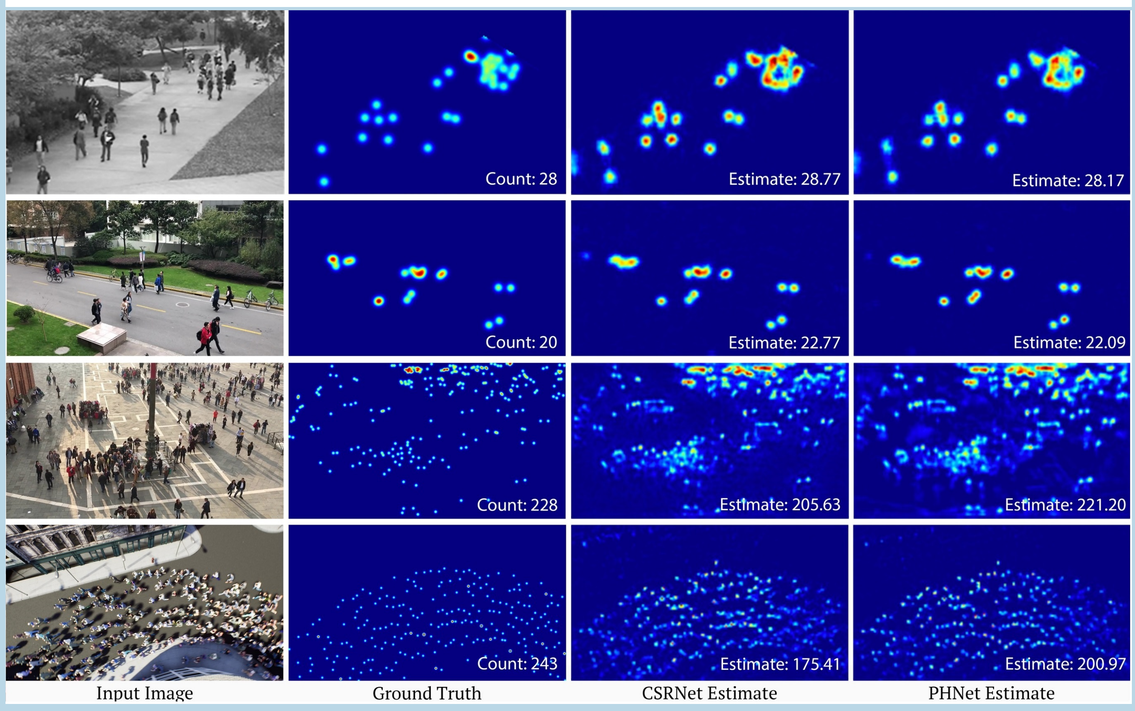
Method	Venue	Year	MAE	RMSE
MCNN	CVPR	2016	145.4	147.3
Switch-CNN	CVPR	2017	52.8	59.5
CSRNet	CVPR	2018	35.8	50.0
CAN	CVPR	2019	23.5	38.9
ECAN	CVPR	2019	20.5	29.9
GPC	IROS	2019	18.2	26.6
PHNet(ours)	-	-	**18.1**	**25.1**

ESTIMATION ERRORS ON THE FDST DATASET

Method	Venue	Year	MAE	RMSE
ConvLSTM	ICCV	2017	4.48	5.82
WithoutLST	ICME	2019	3.87	5.16
MCNN	CVPR	2016	3.77	4.88
LST	ICME	2019	3.35	4.45
PHNet(ours)	-	-	**1.65**	**2.16**

ESTIMATION ERRORS ON THE CROWDFLOW DATASET

Method	Venue	Year	MAE	RMSE
MCNN	CVPR	2016	172.8	216.0
CSRNet	CVPR	2018	137.8	181.0
CAN	CVPR	2019	124.3	160.2
PHNet(ours)	-	-	**107.9**	**127.6**



Reference

[1] Antoni B Chan, Nuno Vasconcelos: *Counting peo-ple with low-level features and Bayesian regression*, IEEE Transactions on image processing (2011)

[2] Roland Memisevic: *Learning to relate images*, IEEE transactions on pattern analysis and machine intelligence (2013)

[3] Yingying Zhang et al.: *Single-Image Crowd Counting via Multi-Column Convolutional Neural Network*. In:2016 IEEE Conference on Computer Vision and PatternRecognition (CVPR) (2016)