Muti-view crowd counting

Kaiyi Zhang

Abstract

Crowd counting is a challenging problem in computer vision tasks with wide applications, such as urban planning, surveillance, and monitoring. It suffers many obstacles like perspective distortion, occlusion, high density. Traditional methods are usually detection-based or extract hand-craft features to reformulate it as a regression problem. In recent years, a great progress has been made in crowd counting task with the development of deep learning techniques. It outperforms a lot than previous solutions especially in congested scene with severe occlusion. While muti-view counting is a popular way to solve the existing problems in single-view counting. This article will introduce an cnn based network published in CVPR 2019 in the field of muti-view counting.

Keywords： computer vision, crowd counting, density estimation, deep learning, muti-view.

1 Introduction

Crowd counting is a very important task in computer vision. It is used to count the number of people in an image or video. This has great applications in industry, such as urban planning, security monitoring, stampede prevention and so on. Crowd counting also has a high correlation with many other computer vision tasks, such as human behavior analysis, crowd positioning, tracking, scene understanding and so on. Therefore, crowd counting has always received a lot of attention from researchers. At the same time, because of the universality of counting task, the method of crowd counting can also be extended to other counting tasks.

As with other computer vision tasks, crowd counting also faces many unsolved problems. For example, people will block each other in a crowded scene, which seriously affects the effect of the algorithm. Another challenge is the scale variations as the perspective distortion of camera, so distant pedestrians tend to be small and nearby pedestrians usually larger. This will make it difficult for neural network to extract scale-aware features. In addition, annotation of datasets requires enormous human and financial resources.

2 Related works

In the long course of research, many classical methods have been put forward. Most of the earlier methods were detection-based[1-3] in which pedestrians in a picture were identified and framed by a rectangle, allowing a natural count of how many people there were. This works well for images containing only sparse crowds, but often fails for dense crowds. Because dense crowds often have serious occlusion, it is difficult to extract human body features. The person in the picture is probably just a head or a dot. Subsequently, to solve this problem, many regression based methods were proposed to build a mapping from image features to density maps. Some early approaches extract handcraft features[4-7] like foreground area, edge orientation or texture to build the regression model. While in recent years, deep learning based regression model bring a great boosting

performance to this task[8-12]. The most common structures can be divided into single column with variable receptive field and muti-columns with relatively fixed receptive field. Both of these two can capture muti-scale information from the input image. The most representatives are[13-14]. And also, dilated convolution and deformable convolution[15-17] can powerfully capture scale features and have good effects on images with drastic scale changes. Beside above,[9] additionally propose a cross scene crowd counting strategy that by searching the patches similar to the test image in the training set, the network will be fine-tuned and have a better generalization ability. For the question that output density map is usually not clear,[18] develop a learnable upsampling block to obtain a high quality density map. Inspired by ensemble learning,[19] propose a boosting cnn to get a better performance.[20] develop a novel feature fusion strategy to fuse key features from different layers.[21] think that we should learn a density map in low resolution first, and then fuse it with next stage features to get a high resolution result.[22] revisit the problem of crowd counting from the perspective of Bayesian theory, giving us a whole new idea. It is worth mentioning that above proposed network architectures are concerned with how to extract scale context information most, which is very significant to counting task. The attention mechanism[10,23-25] is also widely used in crowd counting tasks to mark areas of interest which can make neural network learning focused more on key areas and features. A large number of studies have shown that this mechanism can effectively improve the accuracy of crowd counting task. Shi *et al.*[26] give a segmentation mask, which is a binary image converted from annotation image as a spatial attention. Neural network will just focus on the each small blob region that contains a point, which can be extended to any object counting problem as a general strategy. Wan *et al.*[27] introduce semantic prior to the network, the pixel values of all background areas are multiplied by an attention factor, making the network pay more attention to the crowd and avoid the interference of background noise. Another spotlight is a residual regression strategy to force the network to learn the difference between the input image and the supporting image. Supporting images are obtain by K-means algorithm, which can represent different density levels.

3 Method

3.1 Overview

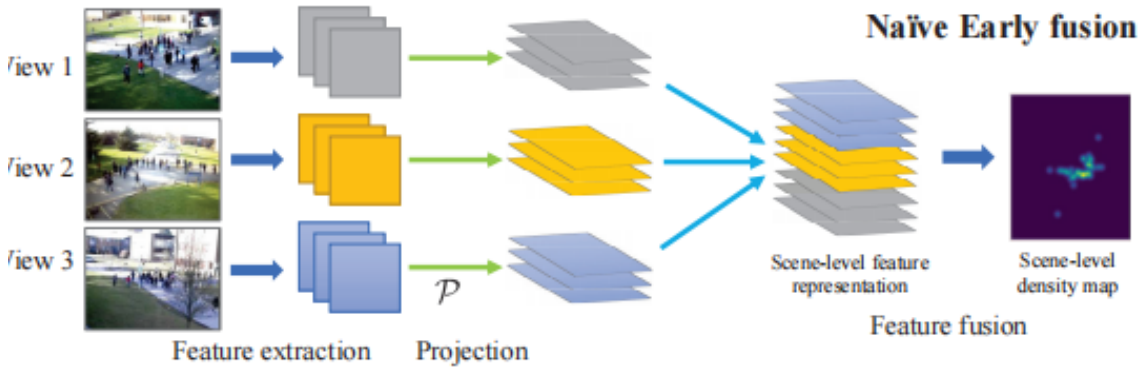

Figure 1: Overview of the method

3.2 Feature extraction

FCN-7	
Layer	Filter
conv 1	$16 \times 1 \times 5 \times 5$
conv 2	$16 \times 16 \times 5 \times 5$
pooling	2×2
conv 3	$32 \times 16 \times 5 \times 5$
conv 4	$32 \times 32 \times 5 \times 5$
pooling	2×2
conv 5	$64 \times 32 \times 5 \times 5$
conv 6	$32 \times 64 \times 5 \times 5$
conv 7	$1 \times 32 \times 5 \times 5$

Fusion	
Layer	Filter
concat	-
conv 1	$64 \times n \times 5 \times 5$
conv 2	$32 \times 64 \times 5 \times 5$
conv 3	$1 \times 32 \times 5 \times 5$

Figure 2: Feature extraction

3.3 Loss

Loss function used in this method is MSE loss, which is commmonly used in deep learning based computer vision task. The pixel-wise distance between output density map and ground truth density map is expected to be small.

4 Implementation details

4.1 Comparing with released source codes

Only tensorflow version codes are available. A new pytorch version is written by myself. No other codes are referenced.

5 Results and analysis

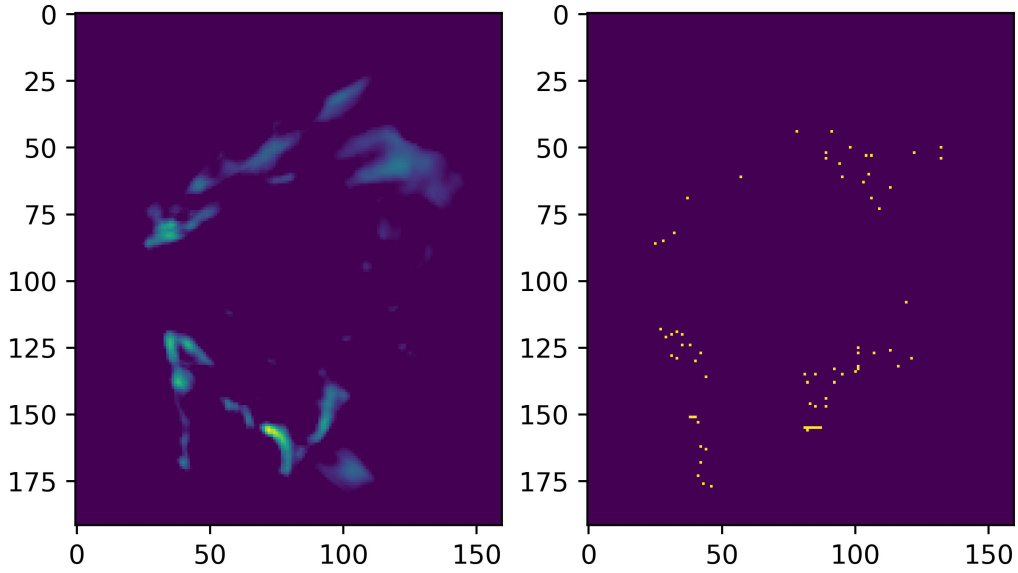

Figure 3: Experimental results

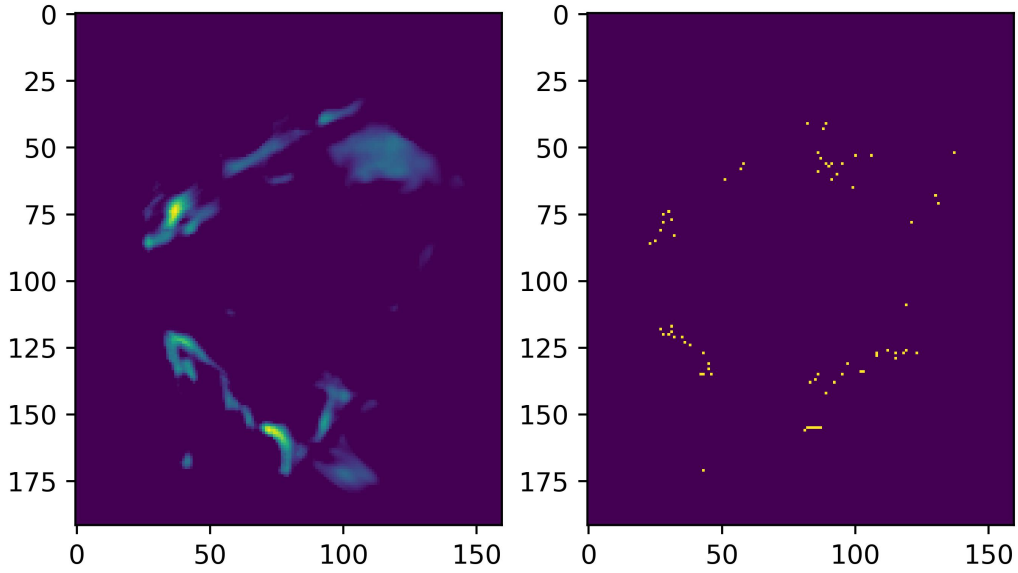

Figure 4: Experimental results

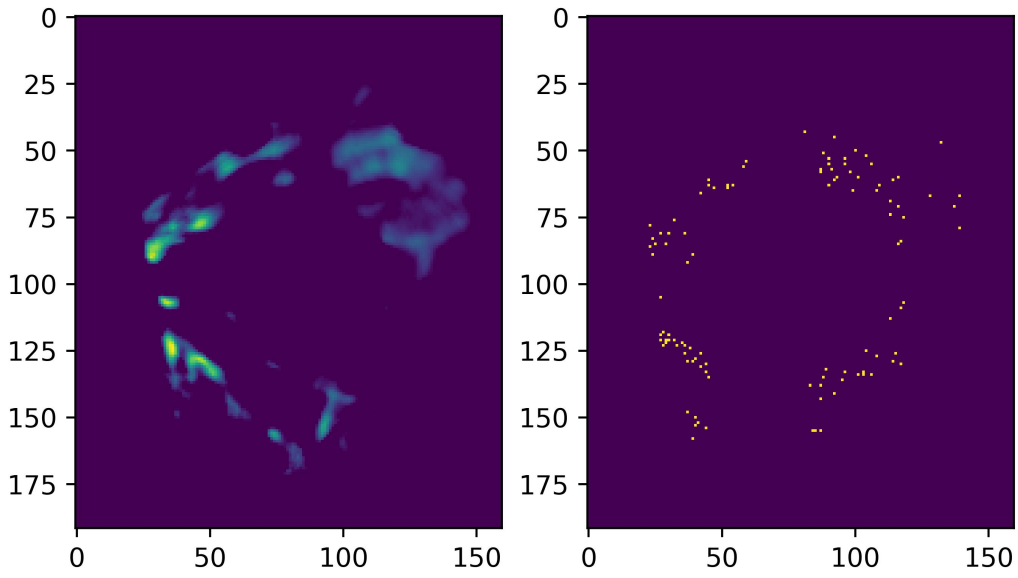

Figure 5: Experimental results

6 Conclusion and future work

Muti-view crowd counting: Many existing crowd counting methods are based on single view. Due to the limitation of lens angle, when the scene size is very large, the single view picture may not capture the whole scene. So multi view fusion counting is a promising direction.

References

[1] GE W, COLLINS R T. Marked point processes for crowd counting[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009: 2913-2920.

[2] LEIBE B, SEEMANN E, SCHIELE B. Pedestrian detection in crowded scenes[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05): vol. 1. 2005: 878-885.

[3] LI M, ZHANG Z, HUANG K, et al. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection[C]//2008 19th international conference on pattern recognition. 2008: 1-4.

[4] CHEN K, LOY C C, GONG S, et al. Feature mining for localised crowd counting.[C]//Bmvc: vol. 1: 2. 2012: 3.

[5] IDREES H, SALEEMI I, SEIBERT C, et al. Multi-source multi-scale counting in extremely dense crowd images[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013:

2547-2554.

[6] PHAM V Q, KOZAKAYA T, YAMAGUCHI O, et al. Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation[C]//Proceedings of the IEEE international conference on computer vision. 2015: 3253-3261.

[7] CHAN A B, VASCONCELOS N. Bayesian poisson regression for crowd counting[C]//2009 IEEE 12th international conference on computer vision. 2009: 545-551.

[8] SHI Z, ZHANG L, LIU Y, et al. Crowd counting with deep negative correlation learning[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5382-5390.

[9] ZHANG C, LI H, WANG X, et al. Cross-scene crowd counting via deep convolutional neural networks [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 833-841.

[10] ZHANG A, SHEN J, XIAO Z, et al. Relational attention network for crowd counting[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6788-6797.

[11] ZOU Z, SHAO H, QU X, et al. Enhanced 3D convolutional networks for crowd counting[J]. arXiv preprint arXiv:1908.04121, 2019.

[12] SINDAGI V A, PATEL V M. Ha-ccn: Hierarchical attention-based crowd counting network[J]. IEEE Transactions on Image Processing, 2019, 29: 323-335.

[13] BOOMINATHAN L, KRUTHIVENTI S S, BABU R V. Crowdnet: A deep convolutional network for dense crowd counting[C]//Proceedings of the 24th ACM international conference on Multimedia. 2016: 640-644.

[14] BABU SAM D, SURYA S, VENKATESH BABU R. Switching convolutional neural network for crowd counting[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5744-5752.

[15] YAN Z, ZHANG R, ZHANG H, et al. Crowd counting via perspective-guided fractional-dilation convolution[J]. IEEE Transactions on Multimedia, 2021, 24: 2633-2647.

[16] DAI F, LIU H, MA Y, et al. Dense scale network for crowd counting[C]//Proceedings of the 2021 International Conference on Multimedia Retrieval. 2021: 64-72.

[17] DEB D, VENTURA J. An aggregated multicolumn dilated convolution network for perspective-free counting[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2018: 195-204.

[18] LI Y, ZHANG X, CHEN D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1091-1100.

[19] WALACH E, WOLF L. Learning to count with cnn boosting[C]//European conference on computer vision. 2016: 660-676.

[20] SINDAGI V A, PATEL V M. Multi-level bottom-top and top-bottom feature fusion for crowd counting [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1002-1012.

[21] RANJAN V, LE H, HOAI M. Iterative crowd counting[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 270-285.

[22] MA Z, WEI X, HONG X, et al. Bayesian loss for crowd count estimation with point supervision[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6142-6151.

[23] JIANG X, ZHANG L, XU M, et al. Attention scaling for crowd counting[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 4706-4715.

[24] HOSSAIN M, HOSSEINZADEH M, CHANDA O, et al. Crowd counting using scale-aware attention networks[C]//2019 IEEE Winter Conference on Applications of Computer Vision (WACV). 2019: 1280-1288.

[25] PAN X, MO H, ZHOU Z, et al. Attention guided region division for crowd counting[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020: 2568-2572.

[26] SHI Z, METTES P, SNOEK C G. Counting with focus for free[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4200-4209.

[27] WAN J, LUO W, WU B, et al. Residual regression with semantic prior for crowd counting[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4036-4045.