Crowd Clustering with Perspective Misrepresentation Correction using Supervised Learning

Mr.M.Chennakesavarao[1], Sayyed Karimulla[2], Shaik Jasmin[3], Y. Bhargavi[4], Syed Mastan Sharif[5]

[1]*Asst. Prof, Dept of CSE, Tirumala Engineering College, Narasaraopet, Guntur, A.P., India*

[2,3,4,5]*B. Tech Students, Dept of CSE, Tirumala Engineering College, Narasaraopet, Guntur, A.P., India*

ABSTRACT - Crowd counting is used to estimate the number of people in a picture. Regression is now a popular technique for counting people's numbers. It is worth noting that, with the advent of convolutional neural networks (CNN), CNN-based approaches have become a research hotspot. It is a more fascinating subject to predict the number of people in the picture than to predict the location of the person in the image. Since perspective distortion causes variations in the size of the crowd in the picture, the perspective transformation present is still a challenge. To reduce viewpoint distortion and more accurately locate a person's location, we created a novel system called the Adaptive Learning Network (CAL).

Keywords: crowd counting; localization; adaptive learning; convolutional neural network

I. INTRODUCTION

The crowd counting task requires you to count the number of people in a picture. The crowd counting task is critical in manufacturing, life, disaster management, security control, and public space design [1–3]. Crowd counting has received increased publicity as people's safety perception has improved. Recently, the crowd counting task has used a convolutional neural network (CNN) to resolve the scale variance problem, resulting in significant improvements in crowd density estimation [4,5].

However, the image's perspective distortion remains a significant challenge for crowd counting; specifically, the model is not especially accurate in predicting avatars with large differences in size in the same image. As a result, understanding how to properly manage items of varying sizes is critical to improving the crowd counting model. Recently, the demand for crowd counting has shifted from simply counting the total number of people in a picture to also locating a precise personal location, allowing for more accurate counting. The Visual Geometry Group (VGG) [6] serves as the foundation for the majority of the current work. Following that, after each max pooling process, extract and decode different sizes of features separately. After each max pooling, we can obtain features with sizes of 1/2, 1/4, 1/8, and 1/16 of the original image scale. The current approach simply superimposes features of varying sizes without taking into account the mixture of varying sizes brought on by different image inputs and different scene inputs. The level The pattern of perspective transformation in each picture is not the same, which means that if our branch information is merged in the same way, the acquired knowledge would not cover all samples. On this basis, we propose using a dynamic framework to combine branch knowledge based on various image features in order to achieve the aim of dynamic evolution. We also suggest a dynamic learning branch combination approach based on the adaptive scenario discovery system (ASD) [7]model. Unlike ASD, our model is concerned with more than just basic counting tasks; we also include the location of complex items in the model.

II. RELATED WORK

The early detection-based methods [8–11] are based on the detection style system, which employs the slide window to detect people in photographs. These methods estimate the number of pedestrians in low-density crowd scenes by detecting their entire body. However, due to occlusions, heads are usually the only identifiable component in high-density conditions. The detectors of some body parts detection methods were proposed as a further development. Regression-based methods, on the other hand, regressed the crowd density plot, the integration of which is the crowd counting result. Earlier methods counted by mapping global image features or combining local patch features, yielding roughly counts. When these two strategies are compared, regression-based approaches outperform in high-density situations. Furthermore, detection-based methods can typically address both counting and localization problems at the same time.

Recently, CNN-based methods have demonstrated their benefits in crowd image feature mapping and people/head recognition for crowd counting and localization. The Multi-column Convolutional Neural Network (MCNN) approach is tested in which three columns of different filters are used to extract features of heads at various scales.

Sam et al. [21] proposed the Switching-CNN and trained each of the three columns with a subset of the patches, while a density selector is used to extract structural and functional differences. Li et al. propose the CSRNet as a method for concentrating on encoding deeper features in congested scenes. In addition, Idrees presented a deep CNN with composition loss method for counting, density map estimation, and localization.

Since regression-based crowd counting approaches are commonly used in counting scenes, the most straightforward approach is to tackle the localization task by sharpening crowd density maps. However, the poor accuracy of the density plot, as argued in most previous studies, remains an unavoidable disadvantage. An early anomaly detection and localization method implemented normalcy models, which jointly demonstrate the appearance and dynamics of complex congested scenes in which MDTs are trained at multiple scales to deal with the problems of empirical evaluation of anomaly detectors on crowded scenes.

For pixel-level image classification, the Fully Convolutional Network (FCN) is suggested. Matan et al. enhanced the classic LeNet to identify digit strings. Ning et al. have used entirely convolutional inference to plan a convent in the segmentation of C. elegans tissues scene. Multi-layered nets have also used entirely convolutional computation in recent years.

III. PROPOSED ARCHITECTURE

We suggest a new model with three components: the backbone, the pathways, and the adaptive branch. Figure 1 depicts the architecture of a novel system called Adaptive Learning Network (CAL). We will go over the framework and implementation in depth in the following parts.

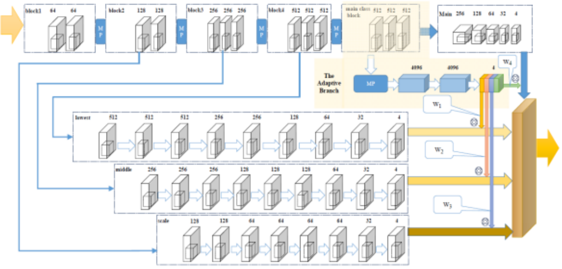
block1 64 64
MP
block2 128 128
block3 256 256 256
block4 512 512 512
main class block 512 512 512
Main 256 128 64 32 4
The Adaptive Branch
MP
4096
4096
4
W4
W1
W2
W3
lowest 512 512 512 256 256 128 64 32 4
middle 256 256 256 128 128 128 64 32 4
scale 128 128 64 64 64 64 32 4

Figure 1. Proposed Architecture

The VGG [6] network is currently used as a backbone in the mainstream approach for extracting features from crowd counting tasks. The use of a backbone network can be divided into two categories: starting from scratch to build a new network or migrating a pre-trained subnet from an existing network. Between these two options, the second has more benefits in terms of both time savings and performance. This is also the basis for our network architecture. We started by creating a feature extraction structure using VGG16 as the backbone. We did, however, replicate and fine-tune some blocks in order to adjust the feature extraction task to multiple resolutions. More precisely, as seen in Table 1, our backbone eliminates the completely connected layer of VGG16. Furthermore, our VGG model is pre-trained on the ImageNet dataset.

IV. RESULTS AND OBSERVATION

In this segment, we will look at three popular crowd counting datasets that are commonly used in crowd counting and localization tasks. Furthermore, several methods for evaluating the efficiency of the architectures are added. Following that, we compare previous experimental findings and assess our approach on these datasets.

Table 1. Results Observed

Approach	MSE	MAE
CAL	63.5	99.2
NO-CAL	70.8	119.5
MCNN	43.1	56.8
CMTL	56.9	67.3
Switching CNN	64.5	53.9

V. CONCLUSION

We present a novel architecture for counting crowds with perspective distortion correction using adaptive learning in this paper. Our approach focuses on using a dynamic learning network to learn the dynamic combination relationship under different samples and then applying this dynamic combination relationship to form different ratios for each image sample. The efficacy and efficiency of our proposed adaptive scenario discovery method for the crowd counting task were tested in relation to state-of-the-art approaches.

VI. REFERENCES

1. Gao, G.; Gao, J.; Liu, Q.; Wang, Q.; Wang, Y. CNN-based Density Estimation and Crowd Counting: A Survey. arXiv **2020**, arXiv:2003.12783.

2. Kang, D.; Ma, Z.; Chan, A.B. Beyond Counting: Comparisons of Density Maps for Crowd Analysis Tasks Counting, Detection, and Tracking. IEEE Trans. Circuits Syst. Video Technol. **2018**, 29, 1408–1422.

3. Sindagi, V.A.; Patel, V.M. A survey of recent advances in cnn-based single image crowd counting and density estimation. Pattern Recognit. Lett. **2018**, 107, 3–16.

4. Tong, M.; Fan, L.; Nan, H.; Zhao, Y. Smart Camera Aware Crowd Counting via Multiple Task Fractional Stride Deep Learning. Sensors **2019**, 19, 1346.

5. Yu, Y.; Huang, J.; Du, W.; Xiong, N. Design and analysis of a lightweight context fusion CNN scheme for crowd counting. Sensors **2019**, 19, 2013.

6. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings

of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.

7. Wu, X.; Zheng, Y.; Ye, H.; Hu, W.; Ma, T.; Yang, J.; He, L. Counting crowds with varying densities via adaptive scenario discovery framework. Neurocomputing **2020**, 397, 127–138.

8. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 1, pp. 886–893.

9. Leibe, B.; Seemann, E.; Schiele, B. Pedestrian detection in crowded scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 1, pp. 878–885.

10. Tuzel, O.; Porikli, F.; Meer, P. Pedestrian detection via classification on riemannian manifolds. IEEE Trans. Pattern Anal. Mach. Intell. **2008**, 30, 1713–1727.

11. Enzweiler, M.; Gavrila, D.M. Monocular pedestrian detection: Survey and experiments. IEEE Trans. Pattern Anal. Mach. Intell. **2008**, 31, 2179–2195.

12. Li, M.; Zhang, Z.; Huang, K.; Tan, T. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In Proceedings of the International Conference on Pattern Recognition (ICPR), Tampa, FL, USA, 8–11 December 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 1–4.

13. Chan, A.B.; Vasconcelos,N. Bayesian poisson regression for crowd counting. In Proceedings of the International Conference on Computer Vision (ICCV), Kyoto, Japan, 29 September–2 October 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 545–551.

14. Ryan, D.; Denman, S.; Fookes, C.; Sridharan, S. Crowd counting using multiple local features. In Proceedings of the Digital Image Computing: Techniques and Applications, Melbourne, Australia, 1–3 December 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 81–88.

15. Kong, D.; Gray, D.; Tao, H. A viewpoint invariant approach for crowd counting. In Proceedings of the International Conference on Pattern Recognition (ICPR), Hong Kong, China, 20–24 August 2006; IEEE: Piscataway, NJ, USA, 2006; Volume 3, pp. 1187–1190.