Multi-camera crowd density estimation for large-scale events

Mrs. Atif Faridi[1], Dr. Farheen Siddiqui[2], Dr. Md. Tabrez Nafis [3] Dr. Mohd Abdul Ahad [4].
[1] Research Scholar, [2] HoD & Associate Prof [,3,4,] Asst. Prof.
[1]cusb.atif@gmail.com, [2]fsiddiqui@jamiahamdard.ac.in, [3]tabrez.nafis@gmail.com, [4]itsmeahad@gmail.com
Department of Computer Science and Engineering,
Jamia Hamdard, New Delhi, India

Abstract—Multi-camera crowd density estimation is a challenging task, particularly for large-scale events where crowds can be massive and dynamic. This problem is important in various fields, such as public safety, event management, and urban planning. In order to perform a single processing task on these multiple views from different cameras we need to first combine the multi-view overlapped images into a single view image. In this paper, we present a deep learning-based approach for multi-camera crowd density estimation in large-scale events. The proposed method uses multiple cameras to capture the scene from different viewpoints and combines the images to estimate the crowd density. We have used cyclic GAN for panoramic image generation/stitching from multiple image views. We use a deep convolutional neural network to learn features from the input images and estimate the crowd density. The proposed method can handle the challenges of large-scale events such as dynamic crowds, occlusions, and lighting variations.

Keywords— video surveillance, crowd density estimation, large-event person density estimation, deep learning

I. INTRODUCTION

Density estimation for mass gatherings is a critical task that has important implications for public safety and event management. Accurate density estimation is necessary to ensure the safety and security of attendees, and it can also be used for planning and optimizing event logistics. Multiple cameras can be used to capture a scene from different viewpoints and provide a more complete picture of the crowd, enabling more accurate density estimation. However, density estimation for mass gatherings using multiple cameras is a challenging task. The number of people in the scene can change rapidly over time, and the cameras may be subject to occlusions, lighting variations, and viewpoint changes. Traditional methods for density estimation rely on handcrafted features and algorithms, which may not be able to handle the complexity of mass gatherings. Recent advancements in deep learning-based methods have shown promising results in various computer vision tasks, including image stitching, density estimation for mass gatherings. Image stitching [1] [2] is the process of combining multiple images with overlapping fields of view to create a single panoramic image. Traditional methods for image stitching rely on hand-crafted features and algorithms, which can be time-consuming and may not produce high-quality results in challenging scenarios. Deep learning-based image stitching [3] [4] is an approach that uses neural networks to learn features and patterns from the input images, enabling them to produce better results in a variety of scenarios. Crowd density estimation using object detection [5] [6] [8] is a popular approach for estimating the number of people in a crowd or an event. Object detection is a computer vision task that involves detecting and localizing objects of interest within an image or video stream. In the context of crowd density estimation, object detection is used to detect individual people within an image or video stream. Once the people are detected using object detection, their locations can be used to estimate the crowd density. This can be achieved by counting the number of people within a predefined area, such as a square meter, and dividing that by the area to obtain the crowd density in people per square meter.

Object detection can be performed using a variety of deep learning-based models, including Faster R-CNN [9], YOLO (You Only Look Once) [10], and SSD (Single Shot Detector) [11]. These models are trained on large datasets of labeled images, allowing them to learn to detect people with high accuracy.

One of the challenges of using object detection for crowd density estimation is the presence of occlusions, where people may be partially or completely hidden from view. This can result in inaccurate density estimates, as some people may not be detected by the object detection model. However, recent research has shown that incorporating additional contextual information, such as the locations of nearby people, can improve the accuracy of object detection in crowded scenes.

Object detection-based crowd density estimation has a wide range of applications, including event management, public safety, and transportation planning. By accurately estimating the crowd density, authorities can better manage the flow of people and resources, ensuring the safety and security of attendees. Overall, object detection-based crowd density estimation is a promising approach for estimating crowd density and has the potential to improve public safety and event management.components, incorporating the applicable criteria that follow.

II. Literature Review

Multi-camera crowd density estimation is a vital task in computer vision and surveillance systems. It involves analyzing multiple camera feeds to estimate the density and movement patterns of crowds in different locations. In recent years, there has been extensive research on this topic, and several methods have been proposed to improve the accuracy of crowd density estimation.

H Rahmalan et al. (2006) [12] have proposed a crowd density estimation technique considering various real-time scenarios like changing background, lighting conditions etc. The authors [12] discuss a method for crowd density estimation based on a combination of background subtraction and optical flow. The method uses a camera to capture video of a crowd, which is then processed to estimate the density of people in the scene. The proposed method aims to overcome the limitations of traditional approaches that use a fixed camera position and assume that the background is static. The method first performs background subtraction to separate the foreground objects, i.e., people, from the background. The resulting binary image is then processed using optical flow to estimate the motion of the people. The optical flow algorithm is used to track the movement of the people in the scene, which is used to estimate their density. The authors [12] evaluate the proposed method on three different datasets: (i) a real-time video captured at a train station, (ii) a video captured at a shopping center, and (iii) a synthetic dataset generated using computer graphics. The results show that the proposed method can accurately estimate the density of crowds in different scenarios. There are several limitations to their methodology i) Limited applicability: The method proposed in the paper is primarily designed for indoor surveillance scenarios with fixed cameras. It may not be suitable for outdoor scenarios, where lighting conditions, weather, and other factors can significantly affect the accuracy of the density estimates, ii) Lack of generalization: The approach is trained and tested on a specific dataset, which may limit its generalizability to other scenarios. The dataset used in the paper consists of only two scenes, one with low density and another with high density, which may not represent the full range of crowd densities encountered in real-world scenarios, iii) Assumption of static cameras: The proposed method assumes that the cameras are fixed and do not move. This assumption may not hold in all surveillance scenarios, especially those involving mobile cameras or cameras mounted on drones, iv) Limited accuracy: While the proposed method achieves good accuracy in estimating crowd density in the specific scenarios tested in the paper, the accuracy may degrade in more complex scenarios involving occlusions, overlapping crowds, or irregular shapes, and v) Computational complexity: The proposed method involves a relatively complex pipeline of image processing and machine learning techniques, which may require significant computational resources and processing time, limiting its real-time applicability.

Wu, G. Liang, KK Lee, and Y. Xu [13] proposes a method for estimating crowd density in surveillance scenarios using texture analysis and learning. The proposed method aims to overcome some of the limitations of traditional methods by using texture analysis and learning techniques to extract features that are relevant to crowd density estimation. The paper presents a comprehensive analysis of the proposed method, including an evaluation of its performance on various datasets and a comparison with state-of-the-art methods. The results show that the proposed method achieves high accuracy in crowd density estimation, even in complex scenarios with overlapping crowds and occlusions. One of the strengths of the proposed method is its adaptability to different scenarios. The texture-based analysis and learning techniques enable the method to extract relevant features from the input images, regardless of the camera angle, height, or distance. This makes it more applicable to real-world situations where the camera placement may vary. Another strength of the proposed method is its ability to operate in real-time. The method is designed to process images in a fast and efficient manner, making it suitable for surveillance applications where timely responses are required. However, there are some limitations to the proposed method. The method requires a large amount of training data to achieve high accuracy, which may be difficult to obtain in some scenarios. Additionally, the method may not generalize well to scenarios that are significantly different from the training data.

M. Fu et al. [14] presented a research paper that presents a novel approach for crowd density estimation using convolutional neural networks (CNNs). The proposed method is designed to achieve high accuracy in crowd density estimation while operating in real-time. The paper provides a detailed description of the proposed method, including the network architecture, training data, and evaluation metrics. The method is evaluated on several datasets and compared with state-of-the-art methods. The results show that the proposed method outperforms other methods in terms of both accuracy and computational efficiency.

One of the strengths of the proposed method is its ability to operate in real-time. The network architecture is designed to minimize computational complexity, making it suitable for surveillance applications where timely responses are required. Additionally, the method achieves high accuracy in crowd density estimation, even in complex scenarios with overlapping crowds and occlusions. Another strength of the proposed method is its adaptability to different scenarios. The network architecture can be fine-tuned for specific scenarios by adjusting the number of layers and filters, and the training data can be augmented to include different types of crowds. However, there are some limitations to the proposed method. The method requires a large amount of training data to achieve high accuracy, which may be difficult to obtain in some scenarios. Additionally, the method may be sensitive to camera parameters such as camera angle, height, and distance.

S. Wang et al. [15] proposes a lightweight convolutional neural network (CNN) based approach for crowd density estimation. The proposed method is designed to operate on edge devices, enabling real-time crowd density estimation in resource-constrained environments. The method is evaluated on several datasets and compared with state-of-the-art methods. The results show that the proposed method achieves high accuracy in crowd density estimation while operating on edge devices. One of the strengths of the proposed method is its lightweight design. The network architecture is designed to minimize computational complexity, making it suitable for edge devices with limited processing power and memory. Additionally, the method achieves high accuracy in crowd density estimation, even in complex scenarios with overlapping crowds and occlusions. Another strength of the proposed method is its adaptability to different scenarios. The network architecture can be fine-tuned for specific scenarios by adjusting the number of layers and filters, and the training data can be augmented to include different types of crowds.

However, there are some limitations to the proposed method. The method may not perform as well in scenarios with significantly different lighting conditions or camera angles, as the training data may not fully represent these scenarios. Additionally, the method may require a large amount of training data to achieve high accuracy, which may be difficult to obtain in some scenarios.

III. Proposed Method

Multi-camera crowd density estimation for large-scale events using deep learning image stitching and people density estimation is a proposed technique that leverages the power of deep learning algorithms and computer vision techniques to estimate the crowd density in large events. The technique involves using multiple cameras positioned at different angles to capture images of the crowd from different viewpoints. These images are then stitched together using deep learning algorithms to create a single panoramic image of the crowd. PIINET (360-degree Panoramic Image Inpainting Network using a Cube Map) [16] is a deep learning network that can be used for inpainting missing regions in 360-degree panoramic images. The network is designed to use a cube map representation of the panoramic image to address the challenges posed by the spherical geometry of such images. 360-degree panoramic images are often represented using equirectangular projection, which maps the spherical image onto a 2D plane. However, this representation can lead to significant distortions and discontinuities at the poles of the image. Cube maps, on the other hand, provide a more uniform representation of the spherical image by dividing it into six faces of a cube. PIINET uses a three-stage approach to inpainting missing regions in a cube map representation of a 360-degree panoramic image. In the first stage, a deep learning network is used to estimate the missing regions based on the surrounding pixels in the image. In the second stage, a fusion network is used to combine the estimated regions with the known regions of the image. Finally, a refinement network is used to refine the inpainted image and remove any artifacts. The advantage of using a cube map representation is that it enables PIINET to inpaint missing regions in a way that is more consistent with the geometry of the original image. This can result in more accurate and visually appealing inpainted images. PIINET has applications in a wide range of fields, including virtual reality, video editing, and surveillance. However, it is important to carefully evaluate the performance of the network in different scenarios to ensure its reliability and accuracy.

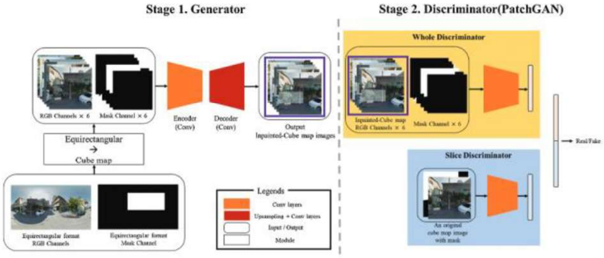
Stage 1. Generator
Stage 2. Discriminator(PatchGAN)
RGB Channels × 6
Mask Channel × 6
Encoder (Conv)
Decoder (Conv)
Output Inpainted-Cube map images
Equirectangular → Cube map
Equirectangular format RGB Channels
Equirectangular format Mask Channel
Legends
Conv layers
Upsampling + Conv layers
Input / Output
Module
Whole Discriminator
Inpainted-Cube map RGB Channels × 6
Mask Channel × 6
Real/Fake
Slice Discriminator
An original cube map image with mask

Figure 1: PIINET [16] Panoramic image creation using multiple views.

Figure 1 shows the architecture of cyclic GAN for training the model to generate panoramic image inpainting using multiple image views. Once the panoramic image is created, density estimation techniques are applied to estimate the density of people in different regions of the image.

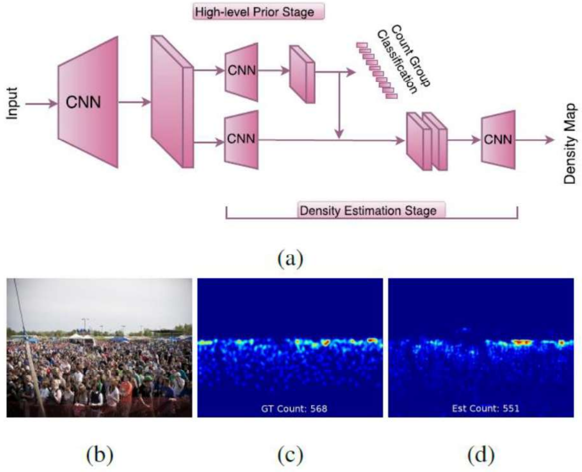

Figure 2: a) CNN based model architecture for crowd counting b) An input image c) Ground truth, d) Density map generated

Figure 2, shows the architecture of the proposed person density estimator, the input image and comparison between ground truth and predicted values. The method uses a cascaded multi-task learning framework to jointly estimate the high-level semantic information and the crowd density in an image. The proposed method consists of two stages: the first stage estimates the high-level semantic information using a deep convolutional neural network (CNN), and the second stage estimates the crowd density using a multi-column CNN. The output of the first stage is used as a prior for the second stage to improve the accuracy of crowd density estimation. The first stage network is trained to predict a heat map of high-level semantic information, such as the locations of buildings and roads, which are closely related to the crowd distribution. The second stage network is composed of multiple columns, each predicting a crowd density map. The final crowd density map is obtained by aggregating the outputs of all columns. This is done using deep learning algorithms that are trained on large datasets of images to recognize and count the number of people in an image. The proposed method is trained using a large-scale crowd counting dataset, such as the ShanghaiTech dataset, which contains over 1000 images with ground truth crowd density maps. The performance of the method is evaluated using the mean absolute error (MAE) and the mean squared error (MSE) between the predicted and ground truth crowd counts. Experimental results show that the proposed method outperforms state-of-the-art methods in terms of both MAE and MSE. The authors also demonstrate the effectiveness of the high-level semantic prior in improving the accuracy of crowd density estimation.

By combining these two techniques, the proposed technique can provide accurate and real-time estimates of crowd density in large-scale events such as concerts, sports events, and festivals. This information can be used by event organizers and authorities to manage crowd control and ensure public safety. However, it is important to note that the effectiveness of this technique may depend on various factors such as the quality of cameras used, the lighting conditions, and the complexity of the crowd movements. Therefore, it is necessary to carefully evaluate and test the technique in different scenarios to ensure its reliability and accuracy.

IV. Results

Image inpainting lacks a good quantitative evaluation scale. However, structural similarity (SSIM) and peak signal to noise ratio (PSNR) are used by Yu et. al. [17]. Table I, shows the performance comparison of generative image inpainting with contextual attention (GI) and our adopted technique.

TABLE I. PERFORMANCE COMPARISON OF IMAGE INPAINTING TECHNIQUES

	Scenery Dataset		**Building Dataset**	
	Ours	**GI [17]**	**Ours**	**GI [17]**
SSIM	**0.9020**	0.790	**0.897**	0.805

PSNR	**37.7**	32.4	**37.4**	32.7

In order to measure the performance of crowd density estimation techniques mean absolute error (MAE) and root mean square error are used. Table II, shows performance comparison between M-CNN [18] and our adopted technique.

TABLE II. PERFORMANCE COMPARISON OF CROWD DENSITY ESTIMATION

	Ours	**M-CNN [18]**
MAE	**69.4**	110.2
RMSE	**96.4**	173.2

Results show the proposed approach is good in comparison to other techniques.

V. Conclusion and Future Scope of Work

The proposed method can be useful in automatic crowd management in urban planning and large event organization. Combining multi-camera views provides a centralized 360-degree view. Processing one combined image will consume less compute power than processing views from individual cameras. In the future we can incorporate more data sources (like data from real large events) to the model to make the model more robust. Future work could focus on developing multi-scale analysis techniques that can capture both local and global variations in crowd density. Future work could focus on developing standardized evaluation metrics and datasets to enable fair comparison of different methods. Future work could focus on deploying crowd density estimation methods at large-scale events and evaluating their performance in real-time.

REFERENCES

[1] Brown, M., & Lowe, D. G. (2007). Automatic panoramic image stitching using invariant features. *International journal of computer vision*, *74*, 59-73.

[2] Lin, C. C., Pankanti, S. U., Natesan Ramamurthy, K., & Aravkin, A. Y. (2015). Adaptive as-natural-as-possible image stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1155-1163).

[3] Hoang, V. D., Tran, D. P., Nhu, N. G., Pham, T. A., & Pham, V. H. (2020). Deep feature extraction for panoramic image stitching. In *Intelligent Information and Database Systems: 12th Asian Conference, ACIIDS 2020, Phuket, Thailand, March 23–26, 2020, Proceedings, Part II 12* (pp. 141-151). Springer International Publishing.

[4] Nie, L., Lin, C., Liao, K., Liu, S., & Zhao, Y. (2022). Deep Rectangling for Image Stitching: A Learning Baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5740-5748).

[5] Nie, L., Lin, C., Liao, K., Liu, S., & Zhao, Y. (2022). Deep Rectangling for Image Stitching: A Learning Baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5740-5748).

[6] Rodriguez, M., Laptev, I., Sivic, J., & Audibert, J. Y. (2011, November). Density-aware person detection and tracking in crowds. In *2011 International Conference on Computer Vision* (pp. 2423-2430). IEEE.

[7] Saleh, S. A. M., Suandi, S. A., & Ibrahim, H. (2015). Recent survey on crowd density estimation and counting for visual surveillance. *Engineering Applications of Artificial Intelligence*, *41*, 103-114.

[8] Liu, J., Gao, C., Meng, D., & Hauptmann, A. G. (2018). Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5197-5206).

[9] Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448).

[10] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

[11] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* (pp. 21-37). Springer International Publishing.

[12] Rahmalan, H., Nixon, M. S., & Carter, J. N. (2006). On crowd density estimation for surveillance.

[13] Wu, X., Liang, G., Lee, K. K., & Xu, Y. (2006, December). Crowd density estimation using texture analysis and learning. In *2006 IEEE international conference on robotics and biomimetics* (pp. 214-219). IEEE.

[14] Fu, M., Xu, P., Li, X., Liu, Q., Ye, M., & Zhu, C. (2015). Fast crowd density estimation with convolutional neural networks. *Engineering Applications of Artificial Intelligence*, *43*, 81-88.

[15] Wang, S., Pu, Z., Li, Q., & Wang, Y. (2022). Estimating crowd density with edge intelligence based on lightweight convolutional neural networks. Expert Systems with Applications, 206, 117823.

[16] Han, S. W., & Suh, D. Y. (2020). PIINET: A 360-degree panoramic image inpainting network using a cube map. arXiv preprint arXiv:2010.16003, 4.

[17] Yu, Z. Lin, J. Yang, X. Shen and X. Lu et al. ,“Generative image inpainting with contextual attention,” in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition , Salt Lake City, Utah, USA, pp. 5505 –5514, 2018.

[18] Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y., 2016b. Singleimage crowd counting via multi-column convolutional neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 589–597.