Image-Based Crowd Estimation Using Deep Convolutional Neural Networks

[1]Poonam kukana, [2]Ameer Ali

[1]Deptt. Of Computer Science, Rayat Bahra University, Mohali Punjab

[2]Deptt. Of Computer Science, Rayat Bahra University, Mohali Punjab

Abstract: Image-based crowd estimation has gained significant attention due to its applications in crowd management, urban planning, and event security. This abstract provides a comprehensive overview of the state-of-the-art techniques in image-based crowd estimation using deep convolutional neural networks (CNNs). The complexity of crowd scenes, with varying densities, occlusions, and lighting conditions, poses a formidable challenge for accurate crowd estimation. Deep CNNs have emerged as powerful tools for addressing these challenges, leveraging their ability to automatically learn hierarchical features from data. The architecture selection process involves choosing suitable CNN architectures, such as VGGNet, ResNet, or specialized crowd counting networks, based on the specific requirements of the application. Data plays a crucial role in training these models. A diverse dataset comprising images of crowded scenes is collected and annotated with ground truth crowd counts. Preprocessing steps, including resizing, normalization, and data augmentation, enhance the network's ability to generalize across different scenarios. The training process involves optimizing the network parameters through backpropagation, minimizing the discrepancy between predicted and ground truth crowd counts. Post-processing techniques are then applied to refine the crowd count predictions. These may include filtering mechanisms to address issues like double counting or false positives. The evaluation of the model's performance is conducted on a separate test dataset, employing metrics such as Mean Absolute Error (MAE) and Mean Squared Error (MSE) to quantify the accuracy of crowd estimations. Deployment of the trained model in real-world scenarios involves integrating it into larger systems for crowd management or utilizing it as a standalone tool. Continuous monitoring and potential retraining are essential for adapting the model to evolving environments and requirements. Image-based crowd estimation using deep CNNs represents a promising avenue for addressing the challenges associated with diverse crowd scenes. The advancements in this field contribute to enhancing the capabilities of crowd management systems, urban planning tools, and security applications, fostering a safer and more efficient interaction with crowded environments.

Keywords: *Crowd Estimation, Image Analysis, Deep Learning, Convolutional Neural Networks (CNNs), Computer Vision.*

Introduction:

The exponential growth of urban populations and the increasing prevalence of large-scale events have underscored the critical importance of crowd estimation in various domains, including public safety, event management, and urban planning [1]. Traditional methods for crowd estimation often relied on manual counting or simplistic computer vision techniques, which were limited in their accuracy and scalability. However, with the advent of deep learning, specifically Deep Convolutional Neural Networks (DCNN) [2], a revolutionary transformation has taken place in the field of crowd estimation, enabling more precise and efficient analysis of complex crowd scenes from images. Crowd estimation involves the quantification of the number of individuals present in a given area, making it an integral component of crowd management and urban surveillance systems

[3]. Traditional methodologies struggled to cope with the challenges posed by diverse crowd dynamics, occlusions, and varying lighting conditions. Deep learning, particularly DCNNs, has emerged as a game-changer by leveraging hierarchical feature learning, enabling the automated extraction of intricate patterns and representations from images [4]. Deep Convolutional Neural Networks (DCNNs) have gained prominence owing to their ability to automatically learn spatial hierarchies of features directly from pixel-level data [5]. This hierarchical learning approach enables DCNNs to discern complex patterns and relationships within images, which is crucial for accurate crowd estimation [6]. By processing images through multiple convolutional layers, DCNNs can effectively capture both local and global features, allowing them to understand the intricate details of crowded scenes. One of the key advantages of employing DCNNs for crowd estimation lies in their adaptability to diverse datasets [7]. Unlike traditional methods that often struggled with variations in scene complexity, DCNNs can be trained on large and diverse datasets, allowing them to generalize well to different crowd scenarios. This adaptability is particularly advantageous in real-world applications, where crowd dynamics can vary significantly across different environments, such as stadiums, urban squares, or transportation hubs [9]. The evolution of DCNNs for crowd estimation can be traced back to pioneering works that sought to address the limitations of conventional approaches. Researchers have explored different architectures and training strategies to enhance the performance of DCNNs in crowd analysis tasks [9]. Techniques such as transfer learning and fine-tuning [10] have proven effective in leveraging pre-trained models on large datasets, leading to improved generalization and robustness in crowd estimation applications. The field continues to advance, researchers are not only focusing on counting individuals but also on extracting more detailed information, such as crowd density maps and trajectory predictions [11]. The ability of DCNNs to capture spatial dependencies enables the generation of density maps, providing a richer understanding of crowd distribution within a given space. This information is invaluable for urban planners, security personnel, and event organizers in optimizing crowd flow and ensuring public safety [12]. In this era of smart cities and intelligent surveillance systems, the integration of image-based crowd estimation using DCNNs holds immense potential for transforming the way we manage and understand crowd dynamics. The following sections of this exploration will delve deeper into the underlying principles of Deep Convolutional Neural Networks, the methodologies employed in training and fine-tuning models for crowd estimation, and the practical applications and challenges associated with implementing such systems in real-world scenarios [13]. Through this comprehensive investigation, we aim to shed light on the transformative impact of DCNNs on image-based crowd estimation and envision the future possibilities that lie ahead in this dynamic field.

Related Work: The proposed work under consideration is an interdisciplinary field that involves aspects of computer vision, image processing, deep learning, and agriculture. A brief description of the existing literature is given as under: In [14] the author offers a multi-task, patch-based deep convolution neural network for crowd tally. Counting people and estimating density are other training approaches for this network. In [15] the author suggested a multi-panel CNN that is altered to take perspective distortion related changes in head size into consideration. A filter of receptive fields in the big, medium, and tiny sizes is contained in each of the three columns of the multi-column design. As a result, the errors associated with perspective distortion can be minimized. In [16] a Cascaded CNN architecture is introduced that enables end-to-end multitask crowd enumeration and density valuation. The resulting density map demonstrates high accuracy with minimal counting inaccuracies. [17] introduces a contextual pyramid of CNNs to create accurate crowd density maps. Learning to categorize the input pictures and their patches into different densities allows one to gain both the global and local contexts. A ADCrowdNet is proposed in [18], an architecture based on convolutional neural networks, to address crowd comprehension in densely populated and noisy environments. ADCrowdNet effectively tackles challenges such as noise, obstructions, and varying crowd disseminations usually encountered in vastly congested and noisy conditions, [19] proposes an Inception Dense Estimator Network, a method based on inception that uses deep learning to calculate population density from a crowd picture, (IDEnet). Due to the inception-based network's high degree of modularity, new image-processing jobs may be readily added or changed. A framework designed for counting individuals in crowd recordings characterized by low to medium density is put forwarded in [20]. The framework utilizes an advanced detector called Faster-RCNN to identify pedestrians within crowd videos. To address false positives, Motion Guided Filters are employed, resulting in an improved mean average accuracy across all detections. [21] introduces an approach for near real-time crowd enumeration based on DCNN (Deep Convolutional Neural Networks). The method offers several advantages, including the utilization of NVIDIA GPU parallel framework for High-performance Computing. It provides a rapid and adaptable solution for processing video feeds from cameras, offering innovative solutions that can be deployed for disaster management and emergency evacuation without the need for explicit system configuration.

Model description and Methodology: Image-based crowd estimation tackles the challenge of accurately quantifying crowd sizes and densities using visual data. Traditional methods are time-consuming and subjective, prompting theneed for automated techniques. This research aims to develop computer vision and machine learning algorithms to overcome challenges like occlusions, perspective distortions, lighting variations, and complex crowd dynamics. Improved crowd estimation techniques have widespread applications in urban planning, event organization, and public safety. The problem of image-based crowd estimation can be formulated as follows: Given an input image (or video frame) of a crowded scene, estimate the number of individuals present in the image. This can be done through various techniques such as object detection, image segmentation, and density estimation. The goal is to develop a model that can accurately estimate the crowd density in an image and count the number of individuals present in the image. The problem of image-based crowd estimation involves using images, such as those captured by cameras or drones, to estimate the location. The goal is to develop algorithms that can automatically analyze the images and properly calculate the population of the crowd. Some ofthe key challenges in this problem include dealing with variations in crowd density, camera perspective, and lighting conditions. Additionally, it is important to consider the privacy and ethical implications of using images of individuals in this way. The approach to conducting research for this project is presented in Figure 1.

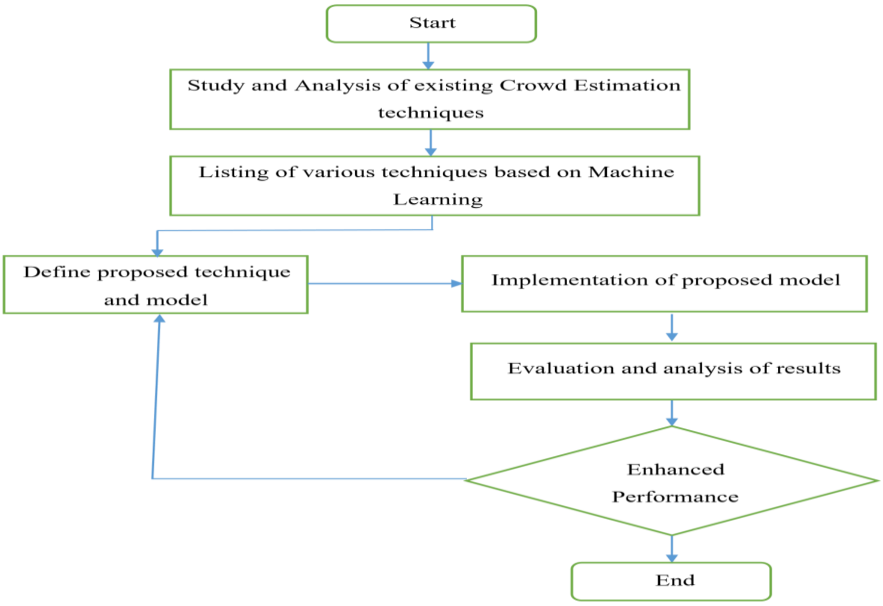
Start
Study and Analysis of existing Crowd Estimation techniques
Listing of various techniques based on Machine Learning
Define proposed technique and model
Implementation of proposed model
Evaluation and analysis of results
Enhanced Performance
End

Figure 1: Proposed Methodology

The prime objective of this training is to utilize deep learning techniques in developing a model that can learn and analyze crowd characteristics, enabling the prediction of crowd density levels. This research aims to facilitate efficient crowd supervision and incident inhibition in crowded environments. The aim is to automate crowd counting without human involvement to effectively handle crowds in a specialized and timely manner. To achieve this, a dataset consisting ofphotographs depicting various crowd sizes was utilized. The focus was on determining the number of heads within densely populated areas. Given the need to avoid crowds to prevent the spread of Covid-19, maintaining distance from crowds and providing support to security personnel in crowd dispersal can help mitigate accidents as shown in figure 2.

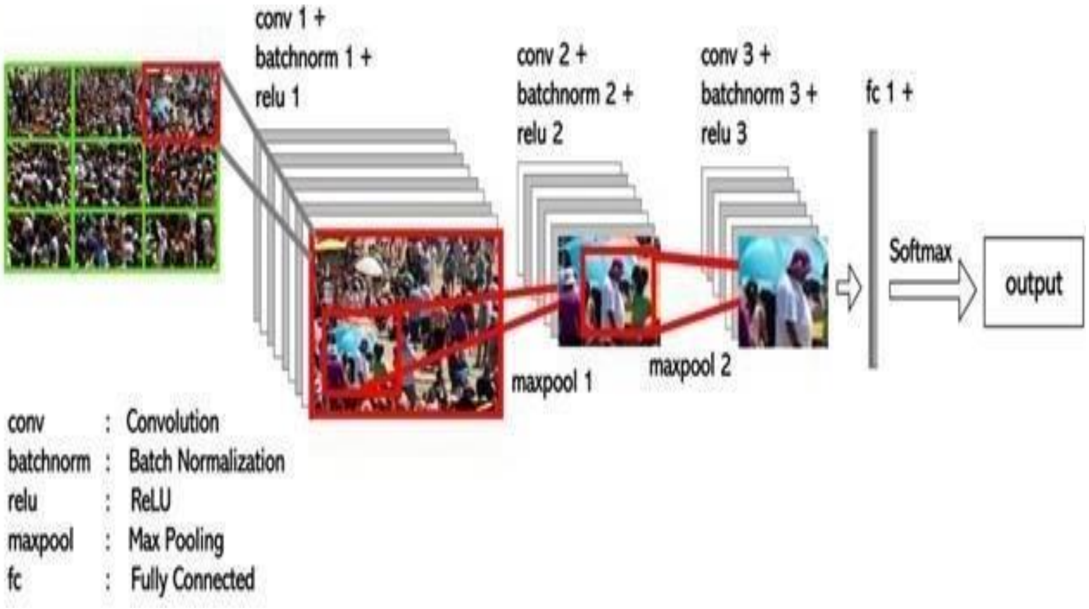

Figure 2: Proposed CNN Classifier

The most substantial part of the enumeration system is data acquisition, and choosing a suitable sample for machine learning trials is imperative. The dataset of crowd estimation wasgathered from the Internet for this study. The dataset contains 1198 images of the ShanghaiTech dataset. The preparation phase involves augmenting the existing dataset by incorporating additional photographs through the utilization of a segmentation technique. The image is divided into nine separate, non-overlapping segments, and density maps are generated for each segment. The total count of heads within each map indicates the number of heads present. Furthermore, understanding the distribution of the total count across the segmented images is crucial. To facilitate this, a dot map is initially formed for the whole image before being divided into segments. The count and positions of heads within a particular image segment are represented by the sum and places of "1" in the corresponding section of the dot map. The complete dot map is also necessary for labeling purposes. During this stage, the primary feature of the image, namely the headcount, is extracted. Subsequently, the headcount of each image is utilized to classify them collectively. The images are then separated into 20 and 33 classes for two separate trials. To shift the focus from counting to density estimation, the approach involves creatingmultiple classes that encompass different assortments of crowd counts. These classes are carefully chosen to represent precise density levels, with higher levels indicating a higher riskof congestion situations. In this phase, a Deep Convolutional Neural Network with three convolution layers (DCNN3) is proposed for crowd enumeration in a given scene.

Results: Deep learning models have been used to estimate crowd density maps from an input image as shown in figure 3, where each pixel in the density map represents the crowd density at the corresponding location in the input image as shown in figure 4.. Some models are designed to directly predict the count of individuals in a crowd from an input image. Besides counting, there is also emphasis on accurately localizing individuals in the crowd. Object detection techniques within CNNs can be applied for this purpose. Robustness to changes in scale, viewpoint, and occlusion has been a focus, allowing models to perform well under different conditions. Techniques like data augmentation have been utilized to improve model generalization by creating variations in the training data. Pre-trained models on large datasets or related tasks have been used as a starting point, followed by fine-tuning on specific crowd estimation datasets.

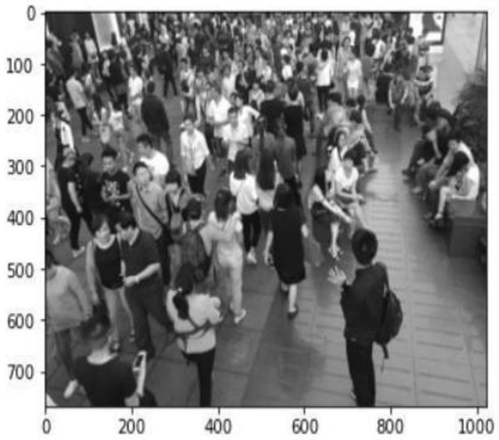

Figure 3: Test Image for density detection in Crowd estimation.

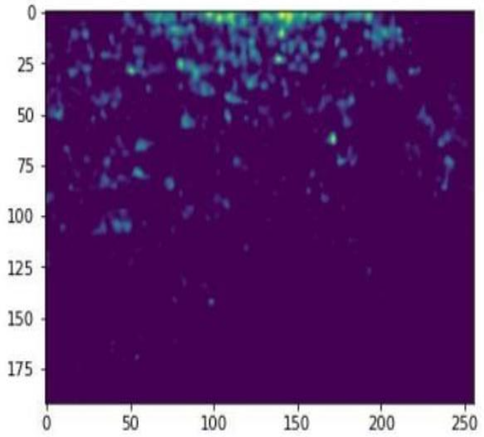

Figure 4: Density map represents the crowd density at the corresponding location

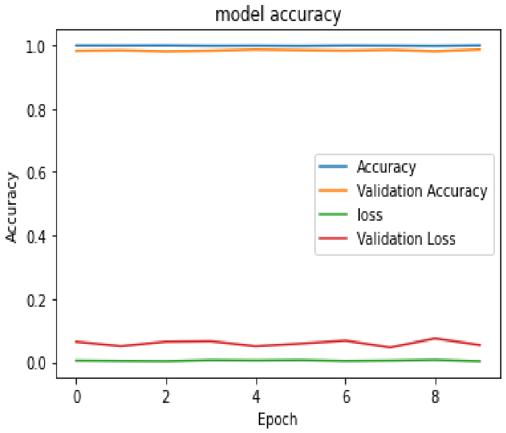

Figure 5: Training Accuracy, Validation Accuracy, Training Loss and Validation Loss.

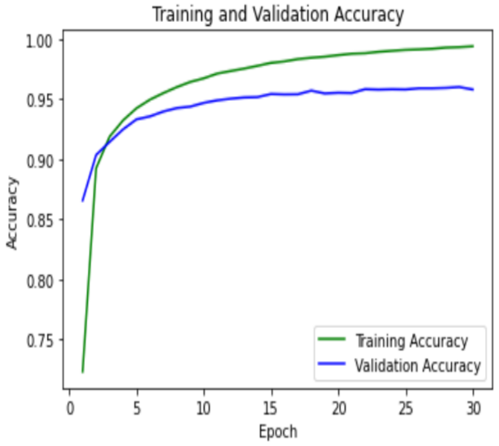

Figure 6: Training and Validation Accuracy at 30 Iterations

Training Accuracy the percentage of correctly classified samples in the training dataset. It is calculated as the number of correct predictions divided by the total number of training samples as shown in figure 5 and figure 6. While high training accuracy indicates that the model is learning well on the training data, it doesn't necessarily guarantee good performance on new, unseen data. Validation Accuracy measures the percentage of correctly classified samples in a separate validation dataset. The validation dataset is not used during the training process and serves as an independent evaluation set. Training Loss is a measure of how well the model is performing on the training data. It represents a quantitative measure of the difference between the predicted values and the actual values for the training samples. validation loss measures the difference between predicted and actual values, but it is computed on the validation dataset. It provides an indication of how well the model generalizes to new, unseen data. An increase in validation loss may suggest overfitting, especially if training loss is decreasing.

Conclusion:

As the population increases, crowd analysis is becoming increasingly important, especially in events like political rallies, sporting events, and music concerts where crowd gatherings are more common. Crowd analysis encompasses tasks such as estimating crowd density and conducting in-depth analysis. Both traditional methods that rely on manually engineered features and methods founded on Convolutional Neural Networks (CNNs) can be employed for these purposes. CNN-based techniques can be further categorized based on the training procedure, network characteristics, and image perspective. Examining crowd density poses various challenges, including occlusion, non-uniform density, changes in scale, and viewpoint variations. Traditional methods are typically limited to handling low-density crowds. However, when comparing the accuracy of different traditional and CNN-based techniques, it becomes evident that CNN-based approaches are better suited for handling dense crowds with varying object scales and scene perspectives. Different models can be used for classification, but CNN appears to be more accurate than other models. It is used as a backbone for many computer vision tasks like crowd estimation.CNN classified images and counted 131 people on 100 epochs. The model's accuracy will improve as the size of the dataset is increased.

References:

[1] Zhang, Cong and Li, Hongsheng and Wang, Xiaogang and Yang, Xiaokang,” Cross- scene crowd counting via deep convolutional neural networks,” in Proceedings of the IEEE Conference on computer vision and pattern recognition,2015, pp. 833-841.

[2] Zhang, Yingying and Zhou, Desen and Chen, Siqin and Gao, Shenghua and Ma, Yi,” Singleimage crowd counting via a multi-column convolutional neural network,” inProceedings of the IEEE Conference on computer vision and pattern recognition,2016, pp.589-597.

[3] Sindagi, Vishwanath A and Patel, Vishal M,” Generating high-quality crowd density maps using contextual pyramid CNNs,” in Proceedings of the IEEE International Conference on Computer Vision,2017, pp.1861-1870.

[4] Sindagi, Vishwanath A and Patel, Vishal M,” Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting,” in 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS),2017, pp. 1-6.

[5] Liu, Ning and Long, Yongchao and Zou, Changqing and Niu, Qun and Pan, Li and Wu, Hefeng,” Adcrowdnet: An attention-injective deformable convolutional network forcrowd understanding,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2019,3225-3234.

[6] 2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI). (n.d.). IEEE.

[7] Saqib, M., Khan, S. D., Sharma, N., & Blumenstein, M. (2019). Crowd Counting in LowResolution Crowded Scenes Using Region-Based Deep Convolutional Neural Networks. IEEE Access, 7, 35317–35329.

[8] Bhangale, U., Patil, S., Vishwanath, V., Thakker, P., Bansode, A., & Navandhar, D. (2020).Near Real-time Crowd Counting using Deep Learning Approach. Procedia Computer Science, 171, 770–779.

[9] Ryan, David, Simon Denman, Clinton Fookes, and Sridha Sridharan. "Crowd counting using multiple local features." In 2009 Digital Image Computing: Techniques and Applications, pp. 81-88. IEEE.

[10] Chan, Antoni B., and Nuno Vasconcelos. "Counting people with low-level features and Bayesian regression." IEEE Transactions on Image Processing 21, no. 4 (2012): 2160- 2177.

[11] Idrees, Haroon, Imran Saleemi, Cody Seibert, and Mubarak Shah. "Multi-source multiscale counting in extremely dense crowd images." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2547-2554.

[12] Kong, Dan, Douglas Gray, and Hai Tao. "A viewpoint invariant approach for crowd counting." In 18th International Conference on Pattern Recognition (ICPR'06), vol. 3, pp. 1187-1190. IEEE.

[13] Cho, Siu-Yeung, Tommy WS Chow, and Chi-Tat Leung. "A neural-based crowd estimation by the hybrid global learning algorithm." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 29, no. 4: 535-541.

[14] Leibe, Bastian, Edgar Seemann, and Bernt Schiele. "Pedestrian detection in crowded scenes." In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 878-885. IEEE.

[15] Wang, Lu, and Nelson HC Yung. "Crowd counting and segmentation in visual surveillance." In 2009 16th IEEE International Conference on Image Processing (ICIP), pp.
2573-2576. IEEE.

[16] Gavrila, Dariu M., Jan Giebel, and Stefan Munder. "Vision-based pedestrian detection: The protector system." In Proc. of the IEEE Intelligent Vehicles Symposium, Parma, Italy.

[17] Mahadevan, Vijay, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. "Anomaly detection in crowded scenes." In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1975-1981. IEEE.

[18] Tuzel, Oncel, Fatih Porikli, and Peter Meer. "Pedestrian detection via classification on Riemannian manifolds." IEEE Transactions on pattern analysis and machine intelligence 30, no. 10 (2008): 1713-1727. Viola, Paul, and Michael J. Jones. "Robust real-time face detection." International journal of computer vision 57, no. 2 (2004): 137-154.

[19] Viola, Paul, and Michael J. Jones. "Robust real-time face detection." International journal of computer vision 57, no. 2 (2004): 137-154.

[20] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." In International Conference on computer vision & Pattern Recognition (CVPR'05), vol. 1, pp. 886-893. IEEE Computer Society, 2005.

[21] Wu, Bo, and Ramakant Nevatia. "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors." In Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, vol. 1, pp. 90-97. IEEE, 2005.