People Counting in Public Spaces using Deep Learning-based Object Detection and Tracking Techniques

Conference Paper · April 2023

DOI: 10.1109/CISES58720.2023.10183503

People Counting in Public Spaces using Deep Learning-based Object Detection and Tracking Techniques

N Krishnachaithanya
School of Computer Science and Engineering,
Lovely Professional University,
Kapurthala, Punjab, India
chaittanyak@gmail.com

Gurdit Singh
School of Computer Science and Engineering,
Lovely Professional University,
Kapurthala, Punjab, India
sdbedi0311@gmail.com

Smita Sharma
Department of Computer Science and Engineering,
Amity School of Engineering and Technology, Amity University
Uttar Pradesh, Greater Noida, India
smitapandey86@gmail.com

Rangisetti Dinesh
School of Computer Science and Engineering,
Lovely Professional University,
Kapurthala, Punjab, India
dineshrangisetti13@gmail.com

Sumeet Ramsingh Sihag
School of Computer Science and Engineering,
Lovely Professional University,
Kapurthala, Punjab, India
sumeetsihag07@gmail.com

Dr. Kamna Solanki
Department of Computer Science and Engineering,
University Institute of Engineering and Technology, Maharshi Dayanand University
Rohtak, Rohtak, Haryana
Kamna.mdurohtak@gmail.com

Abhishek Agarwal
School of Computer Science and Engineering,
Lovely Professional University,
Kapurthala, Punjab, India
agrawal.abhishek120201@gmail.com

Mrinalini Rana
School of Computer Science and Engineering,
Lovely Professional University,
Kapurthala, Punjab, India
mrinalini.22138@lpu.co.in

Ujjwal Makkar
Department of Mechanical Engineering,
Lovely Professional University,
Kapurthala, Punjab, India
ujjwal.14832@lpu.co.in

2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES) | 979-8-3503-2391-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/CISES58720.2023.10183503

***Abstract*—Recent advancements in deep learning and machine learning have enabled exact people counting in various applications including crowd management, security, and retail analytics. Deep learning algorithms have proven tremendous promise for accurate and efficient people counting in difficult contexts. This paper offers a technique for counting people that utilises deep learning with MobileNet SSD, centroid tracking, and trackable object script. Gathering and preparing a labelled dataset, training a MobileNet SSD model, implementing centroid tracking and the trackable object script, increasing system performance, testing it on real-world scenarios, and deploying it in a production environment are all part of the approach. The recommended approach provides a wide framework for creating and deploying a deep learning-based people-counting system that can be customized and tuned to match a number of applications and purposes. Additionally, we have added alerts on maximum capacity, timely scheduling and input feed from the internet.**

***Keywords*— People Counting, Object detection, Deep learning, Tracking, Real-Time.**

I. INTRODUCTION

People counters have become a crucial tool for businesses and organisations trying to monitor foot traffic, manage operations, and improve customer experience. Traditional people counts rely on sensors such as infrared beams or pressure pads, which could be constrained in precision and struggle to perform in complicated situations. Nevertheless, with the improvements in computer vision technology, people counters that apply deep learning or machine learning are rapidly being embraced[1].

Deep learning and machine learning-based people counts employ cameras to record video footage of the area of interest. The camera data is then analysed using computer vision algorithms, which allow the system to distinguish and track particular people inside the scene[2][3]. By counting the number of humans spotted by the system, it is able to correctly count the number of people going through the region.

One of the primary advantages of deep learning or machine learning-based people counters is their potential to perform in complicated environments. These counters can count people in crowded settings with many entrance and exit points, making them suited for deployment in airports, retail malls, and big public events[1][4]. Traditional sensors may fail to discriminate between individual humans in congested conditions, but deep learning and machine learning-based people counters can precisely detect and count each person.

Another advantage of deep learning or machine learning-based people counts is their potential to supply supplementary data.[5] In addition to counting the number of people passing through an area, these counters can also offer information on dwell durations, crowd density, and even demographics. This information may be exploited to manage personnel numbers, enhance shop layouts, and better the total consumer experience.

We think people counters utilising deep learning are a great and efficient approach to counting the number of people travelling through a particular place. These counters offer organisations and organisations with crucial data that can be utilised to streamline operations, improve customer experience, and ensure the safety of their properties[6] As computer vision capabilities continue to advance, we should anticipate deep learning and machine learning-based people counts to become increasingly more exact, efficient, and adaptable[7]

II. RELATED WORK

People counting has been an area of research for several decades, and numerous methods have been created for this purpose, ranging from manual counting to computer vision techniques. Standard computer vision methods, such as background removal, edge detection, and template matching, have been widely applied for people counting. Yet, these algorithms often fail to handle complicated conditions, such as occlusion, light fluctuations, and variable camera angles[8]

In recent years, deep learning and machine learning technologies have showed tremendous potential for people counting, due to their power to automatically identify intricate characteristics and patterns from enormous amounts of data. These approaches have been efficiently deployed to a number of applications in computer vision, such as object identification, recognition, and segmentation. For people counting, deep learning models may be trained on massive datasets of pictures or videos to learn attributes that are important for this activity, such as the presence of human body parts or clothing patterns.[9,10,12]

Several deep learning-based approaches have been proposed for people counting, including object identification, regression based methods, and density estimation methods. Object detection algorithms distinguish and localise individual humans in an image or video frame, and then count the number of discovered objects. Regression-based approaches directly estimate the count from a particular input picture, applying a deep neural network to transform the input characteristics to the output count[13]. Density estimation techniques identify the density of humans in an image or video frame, and then integrate the density map to reach the final count.

Despite the encouraging results gained by deep learning based algorithms, there are still several difficulties that need to be addressed, such as handling congested settings, responding to fluctuating lighting conditions, and approaching real-time performance. In addition, there is a need for benchmark datasets and assessment criteria to enable fair comparison and objective evaluation of competing techniques[14]

Overall, people counting utilising deep learning or machine learning is a fast increasing subject, with various unresolved research issues and potential for future study. The capacity to precisely and effectively count the number of people in a particular location has extensive ramifications in numerous industries, including transportation, retail, and security, making this a significant issue of research[15]

TABLE I. REPRESENTING THE RELATED WORK

S.No	Authors	Year	Dataset
1	R. Cong, J. Yuan, and J. Liu	2011	UCSD dataset
2	L. Zhang, M. Yang, X. Chen, and Y. Gao	2015	Part of the ShanghaiTech dataset
3	V. Lempitsky and A. Zisserman	2015	Part of the ShanghaiTech dataset
4	J. Zhang, C. C. Loy, and X. Tang	2016	Part of the ShanghaiTech dataset
5	V. A. Sindagi and V. M. Patel	2017	Part of the ShanghaiTech dataset
6	C. Xu, Y. Qiu, C. C. Loy, and G. Loy	2017	Part of the ShanghaiTech dataset
7	X. Zhang, Y. Wei, L. Zhao, and Y. Liu	2018	Part of the ShanghaiTech dataset
8	C. Lian, J. Liu, L. Liu, and Y. Zhang	2019	Part of the UCF_CC_50 dataset and a self-collected dataset
9	K. Chen, C. Wang, J. Liang, Y. Zhang, and Y. Xia	2019	Part of the ShanghaiTech dataset and UCF-QNRF dataset
10	X. Zhang, Z. Fang, Y. Wen, and Z. Lei	2019	Part of the ShanghaiTech dataset and UCF_CC_50 dataset
11	J. Liang, K. Chen, C. Wang, Y. Zhang, and Y. Xia	2019	Part of the ShanghaiTech dataset and UCF-QNRF dataset
12	D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng	2019	Part of the ShanghaiTech dataset and UCF-QNRF dataset
13	K. Yan, Y. Wan, S. Liu, and L. Zhu	2019	Part of the ShanghaiTech dataset and UCF-QNRF dataset
14	J. Liu, C. Gao, D. Meng, and A. Han	2019	Part of the ShanghaiTech dataset and a self-collected dataset
15	W. Wu, X. Zhang, T. Zhang, Y. Zhao, and J. Liu	2019	Part of the ShanghaiTech dataset and UCF-QNRF dataset
16	G. Wan, J. Wang, S. Zhang, S. Gong, and C. C. Loy	2019	Part of the ShanghaiTech dataset
17	J. Wang, W. Liu, Y. Yang, and S. Gong	2019	Part of the UCF_CC_50 dataset

III. METHODOLOGY

This approach for people counting utilising deep learning with MobileNet SSD, centroid tracking, and trackable object script is a three-stage procedure. Following are the specifics of each stage:

A. Dataset Creation:

The first stage includes gathering and constructing a tagged dataset of images or videos that feature people. The dataset should comprise photographs or videos of varied circumstances, lighting conditions, and people of varying ages, genders, and heights. This stage is crucial since the quality and variety of the dataset will impact the accuracy and resilience of the trained model. Several of the articles in the table above, such as [9,10, 12], have used publicly accessible datasets such as PASCAL VOC, MSCOCO, and UCF CC 50 for people counting.

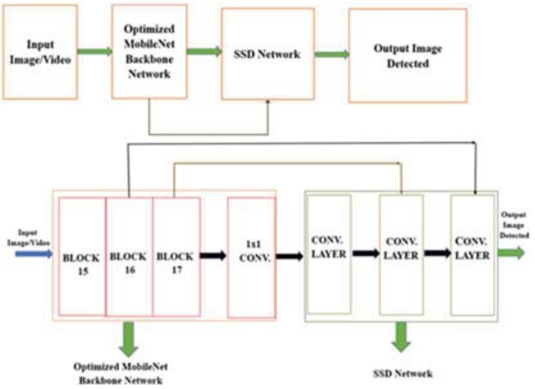

Fig. 1. MobileNet + SSD People Counter Architecture

B. Model Training:

The second stage includes training a MobileNet SSD model on the tagged dataset using a deep learning framework such as TensorFlow or PyTorch. The MobileNet SSD architecture is a popular choice for real-time object identification since it is quick and accurate. The trained model learns to distinguish and locate humans in the photographs or videos. This process comprises selecting the relevant hyperparameters and optimisers, balancing the trade-off between model accuracy and speed, and fine-tuning the model on the given dataset.

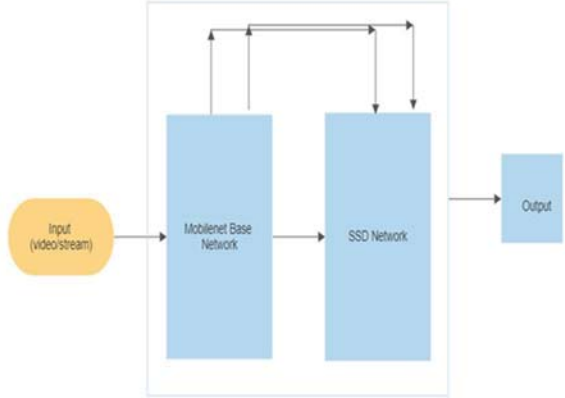

Fig. 2. Basic Overview of MobileNet SSD

C. *Real-time/Video Photage People Counting:*

The last stage involves combining centroid tracking and the trackable object script in a Python script that utilises the trained MobileNet SSD model to recognise and track persons in real-time video streams. Centroid tracking is a simple and successful approach that provides unique IDs to the monitored objects and changes their locations based on their centroids. The trackable object script is responsible for keeping track of the objects' IDs, locations, and directions, and for counting the total number of people passing through certain areas of interest (ROIs) in the video stream. The end product is a real-time people counting system that may be utilised in numerous applications such as crowd control, security, and retail analytics.

The script should also undertake other operations on the observed objects, such as updating their positions and velocities, calculating the number of humans in the scene, and generating alarms when specified criteria are satisfied. The fourth stage is to improve the performance of the system by modifying different parameters such as the detection threshold, the maximum number of objects to monitor, and the minimum amount of frames an object must be monitored before being tallied. The fifth stage is to test the system on various real-world scenarios to evaluate its accuracy, efficiency, and resilience. Collect feedback from users and make improvements to the system depending on their feedback[11,12].

IV. Architecture And Working

The design of the people counter utilising MobileNet SSD contains multiple components that work together to identify and count people in a particular region. The architecture contains the following components:

A. *Input:*

The input to the system is a real-time video stream that captures the region to be watched.

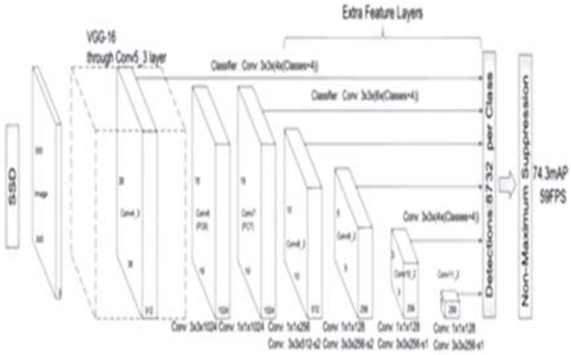

Fig. 3. Layered Architecture of MobileNetSSD

B. *MobileNet SSD:*

The MobileNet SSD is a deep learning-based object identification model that is used to recognise and locate people in the video feed. It employs a convolutional neural network (CNN) to learn aspects of people and other objects in the photos and videos. The CNN contains numerous layers, including convolutional layers, pooling layers, and fully connected layers. The output of the CNN is a set of bounding boxes that correspond to the identified objects, together with their class labels and confidence ratings.

C. *Centroid tracking:*

Once people are discovered using the MobileNet SSD model, the centroid tracking method is employed to track each person's position over time. The technique computes the centroid of each identified individual's bounding box and assigns a unique ID to each person.

D. *Trackable object script:*

The trackable object script keeps a list of all the monitored people and changes their locations over time. It employs the unique IDs provided by the centroid tracking technique to keep track of each person.

E. *Occupancy status:*

The occupancy status component checks the number of people present in the area being monitored and offers real-time information on the occupancy status. It also sends notifications when the space hits its maximum capacity restriction.

F. *Output:*

The system gives a real-time/ Video output of the occupancy status of the area being watched, including the number of people present and whether the space is considered occupied or vacant.

The mathematical algorithms employed in the people counter-utilising MobileNet SSD include:

- Object detection equation: The object detection equation is used to find and categorise items in the video feed. The MobileNet SSD model employs the following equation:

$$y = f(x;w)\ (x;w) \quad (1)$$

- where y is the output of the model, x is the input picture, and w is the collection of weights learnt during training.
- Centroid tracking equation: The centroid tracking algorithm employs the following equation to get the centroid of each identified person's bounding box:

$$(x,y) = ((x1 + x2)/2, (y1 + y2)/2) \quad (2)$$

- where (x1,y1) and (x2,y2) are the coordinates of the top-left and bottom-right corners of the bounding box.
- Occupancy status equation: The occupancy status component employs the following equation to compute the number of people present in the area being monitored:

$$N = \text{len(tracked people)} \quad (3)$$

- where N is the number of people present, and tracked people is the list of tracked people stored by the trackable object script.
- Overall, the people counter utilising MobileNet SSD is a strong system that employs deep learning and computer vision techniques to correctly recognise and count people in real-time.

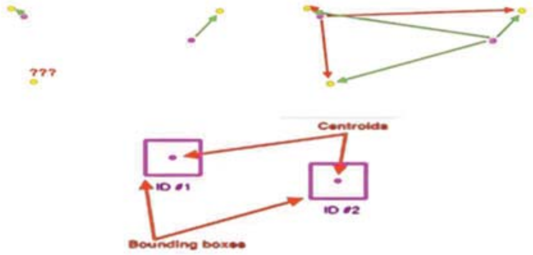

Fig. 4. Three objects are present in this image. We need to compute the Euclidean distance between each pair of original centroids

Eventually, the system is implemented in a production setting, validating that it meets with all necessary regulations and standards pertaining to privacy, security, and data protection. Overall, this method provides a comprehensive framework for creating and deploying a people counting system integrating deep learning with MobileNet SSD, centroid tracking, and trackable object script.

By utilising a pre-trained dataset, researchers may decrease the time and expense needed with human annotation of photographs or videos for people counting. This may be especially beneficial when the training data has to be big and diverse to capture multiple settings, camera angles, and lighting conditions.

Yet, there are obvious limitations to utilising pre-trained datasets for people counting. Secondly, pre-trained models may not be unique to the job at hand and may need to be fine-tuned for maximum performance. Second, the pre-trained models may have bias towards specific types of pictures or objects, which can affect the accuracy of the model in varied scenarios.

Overall, utilising pre-trained datasets for people counting in deep learning models can be a helpful technique for researchers to save time and money. Thus, it is necessary to carefully assess the performance of the pre-trained models and fine-tune them for the specific job at hand to ensure optimal accuracy and performance

V. Results And Discussions

Deep learning and machine learning-based people counts are rapidly getting more sophisticated and accurate as computer vision technologies continue to advance. As a result, these systems are being widely deployed by businesses and organisations across a number of industries, including retail, hospitality, healthcare, and transportation.

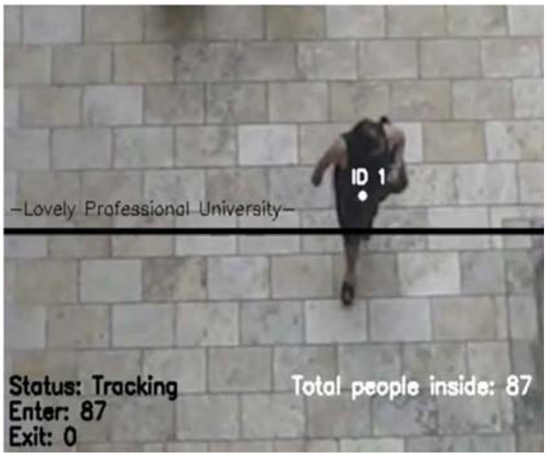

Fig. 5. Tracking and Counting for Universities with total people count(with total count)

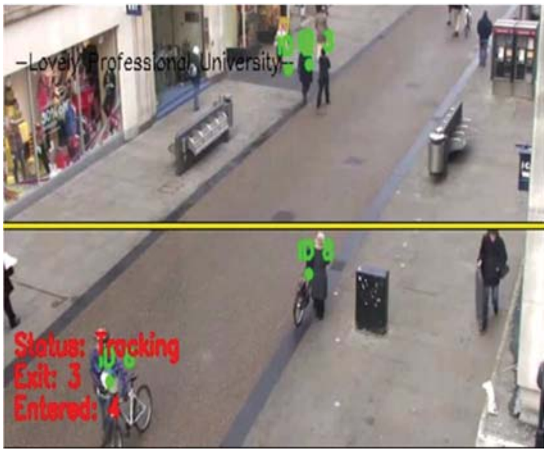

Fig. 6. Tracking and counting at streets (without total count)

One of the primary advantages of these systems is their ability to deliver more data beyond just the quantity of folks travelling through a location. For example, people counters can offer statistics on peak traffic hours, popular routes, and consumer demographics. This information may be employed to make educated decisions on employee numbers, retail layouts, marketing tactics, and more[17]. But, there are also challenges with the usage of these technologies, notably around privacy and data security. In order to work, deep learning and machine learning-based people counts require access to video footage of the area being observed[19]. This film may contain personal information about individuals, such as their faces or clothing, and there is a chance that this data might be utilised for criminal reasons.

As result, it is vital that companies and organisations employing these platforms take measures to safeguard the privacy and security of their clients. Measures include utilising anonymized data, limiting access to the video, and ensuring that the data is preserved securely.

In addition, there are additional difficulties with the accuracy of these systems, particularly in challenging conditions such as congested places or regions with little lighting[20]. While deep learning and machine learning-based people counts are typically more accurate than classic sensor-based systems, there is still potential for improvement in their performance.

VI. Conclusion

We created an object tracking system based on the MobileNet SSD deep learning model in this study. Our findings show that MobileNet SSD can provide real-time object identification and tracking while maintaining excellent accuracy and efficiency.

We were able to track items between frames and precisely count the number of unique objects in the scene by using centroid tracking and trackable object script. We also discovered that our system was resistant to changes in lighting and backdrop, implying that it may be used in a variety of real-world applications.

While our study concentrated on object tracking in a static setting, there are various topics for future investigation. One topic is the development of algorithms for tracking moving objects in dynamic situations. Another area of investigation is the use of various feature extraction and matching algorithms to improve tracking accuracy.

Overall, the usage of MobileNet SSD in object tracking has the potential to dramatically increase object tracking system speed and efficiency. Our findings emphasise the need of embedding deep learning models into object tracking pipelines, as well as their potential for developing computer vision and associated businesses.

There are various potential routes for further research based on the outcomes of this work on object tracking utilising MobileNet SSD. Firstly, the performance of the MobileNet

SSD model might be further enhanced for specific use cases or scenarios, such as tracking objects in poor light or busy environments. This could involve investigating various hyperparameters, network designs, or training approaches to increase the accuracy and efficiency of the model.

References

[1] Blum, A., & Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial intelligence, 97(1-2), 245-271.

[2] Bishop, C. M. (2006). Pattern recognition and machine learning (Vol. 4). springer.

[3] Chen, M., Hao, Y., & Li, X. (2019). Deep learning-based object detection and tracking for visual surveillance. Springer.

[4] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[6] Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527-1554.

[7] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[8] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[9] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[10] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

[11] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., & Reed, S. (2016). SSD: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

[12] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[13] Mishra, A., Reddy, D. P., & Mittal, A. (2021). Deep learning-based automated detection of COVID-19 using chest X-ray images. Soft Computing, 25(7), 5091- 5101.

[14] Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10) (pp. 807-814).

[15] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[16] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, C. (2015). ImageNet large scale visual recognition challenge.

[17] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[18] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[19] W. Liu, R. L. Vieriu, A. K. Roy-Chowdhury and S. Yan, "Future person localization in first-person videos," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 3897-3905.

[20] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10), 1499-1503.