A Deep Learning Based Person Detection and Heatmap Generation Technique with a Multi-Camera System

Conference Paper · December 2022

DOI: 10.1109/ICECE57408.2022.10089044

2 authors, including:

Shakhrul Iman Siam
The Ohio State University
7 PUBLICATIONS **48** CITATIONS

SEE PROFILE

A Deep Learning Based Person Detection and Heatmap Generation Technique with a Multi-Camera System

Md Shakhrul Iman Siam[1], Subrata Biswas[2]
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology, Dhaka 1205, Bangladesh
Email: [1]shakhrul.buet.eee@gmail.com, [2]sbiswas@wpi.edu

***Abstract*—This paper outlines a technical method for video analysis that may be used to identify persons in footage from several CCTV cameras and provide a heatmap of that information for a certain floor layout. The analysis of customer and employee behavior in retail and office settings, as well as motion tracking and advertising effectiveness research, can all be aided by the automatic creation of people density maps. With the use of video recordings made by common video surveillance cameras, density maps were created. We made advantage of CCTV cameras, which are dispersed across a retail establishment. Because the Yolov5 object detection algorithm may produce findings more quickly, we have chosen to employ it for human detection. Additionally, due to the short inference time, it is appropriate for real-time applications.**

***Index Terms*—Object Detection, YOLO, Heatmap, KDE, Homography Transform.**

I. Introduction

Real-time human detection is one of the most fundamental tasks in computer vision and it has become one of the most popular research topics in different fields over the last few years since it has numerous commercial applications [1]. Over the past few decades, human identification, tracking, and segmentation have been the subject of substantial research. Although several algorithms have been put out, there are still issues in the discipline. There are additional obstacles in detecting and tracking for the object class of humans. First, because the human body can move freely at numerous joints, the way people look can change depending on the angle from which they are viewed as well as how their body parts are positioned. People also wear a range of clothing and accessories, which when combined can create hundreds of different combinations of hues, textures, materials, and fashions.

Detection of humans from surveillance cameras can be done by various techniques [2] including Motion Based detection and Deep Learning based detection. Deep learning person detection algorithms have advanced quickly in recent years, considerably enhancing both detection speed and accuracy. Deep learning-based computer vision technology outperforms conventional image processing and recognition techniques in terms of detection speed, algorithm robustness, and feature extraction without manual design [3]. Due to the fact that deep learning techniques are data-hungry, a specific image dataset for the construction industry is needed in order to apply object detection on building sites. Because of the complexity and dynamic nature of construction activities, it is difficult to gather and annotate images, which is why there aren't many well-annotated picture datasets created specifically for the construction industry in widely used open databases.

A Heatmap is a graphical representation of data where values are depicted by color. The more congested data is at a particular location, the hotter will be the color used to represent this data. From the heatmap, we can easily find the areas that are more attractive to customers or visitors. we can get info about which places of a retail store are crowded and which are less crowded. Also, we can analyze this over time. Like, At which time of the day or which day of the week do people come the most. This info will help businesses with further analysis [4]. In this work, we represent a scalable solution for real-time human detection and heatmap generation on a floor layout in order to generate useful business insights such as customer behavior, shopping pattern, etc.

II. Related Works

In the past, person detection in surveillance videos was done manually. The task of identifying people in images has gained significant attention due to the increasing importance of biometrics and surveillance. Deepak et al. [5] developed an algorithm based on the background subtraction method for real-time object detection using Faster-R-CNN. Kajabad et al. [6] describe a people detection approach using a deep learning method. They also proposed an algorithm to find the hot zones of people's movement in the image. Parzych et al. [7] explained how to create a density map of people's movement from video footage analysis in a salesroom. However, their detection method is based on people's movement activity which can be implemented only if there are continuous movements. Punn et al. [8] used Yolov3 with deepsort tracking technique to detect people in order to monitor social distancing. Khan et al. [9] used Yolo, Faster-R-CNN, and SSD for identifying hotspots of people to mitigate the transmission of the coronavirus.

However, none of these works, address the question of how to locate the congested region on a 2-Dimensional floor layout. Existing literature only describes how to create a heatmap of persons across single-camera photos, making it impossible to combine data obtained from multiple cameras. To the best of our knowledge, this is the first paper that describes a full pipeline of detecting people from camera images and mapping the information on a floorplan taken

from multiple cameras to find the areas that are frequently visited by people.

III. Methodology

A. Cameras and Floor Layout

The proposed system is based on a multi-camera system that covers the entire floor region. A Floor layout and position of various surveillance cameras of a Retail store are shown in Fig.1. The position and coverage areas of each camera are known.

B. Person Detection

Images are captured from the video footage of the CCTV cameras. For Detecting Humans in the camera images, we use the Yolov5 object Detection algorithm [10].

1) Dataset: To get the best result from the object detection model, we train it on Crowdhuman Dataset [11]. This dataset contains 15000, 4370, and 5000 images for training, validation, and testing, respectively. There are a total of 340k person instances in the training set, and each human instance is annotated with a head bounding box and human full-body bounding box. Some of the sample images from the CrowdHuman dataset are shown in Fig.2.

2) Model Architecture: The proposed system uses Yolov5-m architecture for object detection task. Yolov5-m is a medium-sized model with 21.2 million parameters which consists of:

- **Backbone (CSP-Darknet53):** CSP-Darknet53 [12] with a Spatial Pyramid Pooling (SPP) layer is used as the backbone for Yolov5 which acts as a feature extractor. It is a convolutional neural network (CNN) that uses DarkNet-53 as its backbone to detect objects.
- **Neck (PANet):** The neck is a feature aggregator which collects feature maps from different stages of the backbone. It creates a connection between the backbone and the head. We used the Path Aggregation Network(PANet) [13] to build the neck.
- **Head (YOLO Layer):** As the acronym of YOLO stands for "You Only Look Once", it is a one-stage detector that makes the predictions for object localization and classification at the same time. The one-stage head [14] for dense prediction provides information about where the object is present.

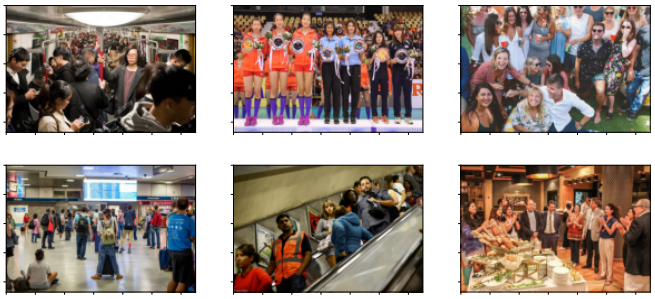


Fig. 2: CrowdHuman Dataset

First The input image is fed into CSP-Darknet53 for feature extraction and then into PANet for feature fusion, and finally, the outputs (class, location, confidence score) are obtained from the YOLO layer. Fig.3 shows an overview of Yolov5 architecture.

3) Training: We train the model with the SGD optimizer for 200 epochs on CrowdHuman Dataset. The learning rate and the momentum are 0.01 and 0.973 respectively with a weight decay of 0.0005. We use horizontal flipping and Mosaic augmentation technique to make our model more robust. Three types of losses- Bounding box regression loss (Mean Squared Error), Objectness loss (Binary Cross Entropy), and Classification loss (Cross-Entropy) are calculated. All these loss curves vs epochs are shown in Fig.4.

4) Performance evaluation: we use the crowdhuman test dataset which consists of 5000 images to evaluate our model, We used mean Average Precision (mAP) as an evaluation metric that calculates a score by comparing the predicted bounding box to the actual bounding box. We calculate mAP using the Intersection over Union(IOU) value for a given IOU threshold and obtained 0.85 mAP(0.5) on the test dataset.

C. Coordinate Transformation

To visualize the heatmap on a 2D floorplan, Coordinate transformation of each person's position from the camera image to the floorplan is required. It is much easier to visualize movement patterns presented on a 2D floorplan rather than when shown on CCTV footage. For this purpose,we use Homography Transformation [15]. Homography Transformation is a mapping between two planar projections

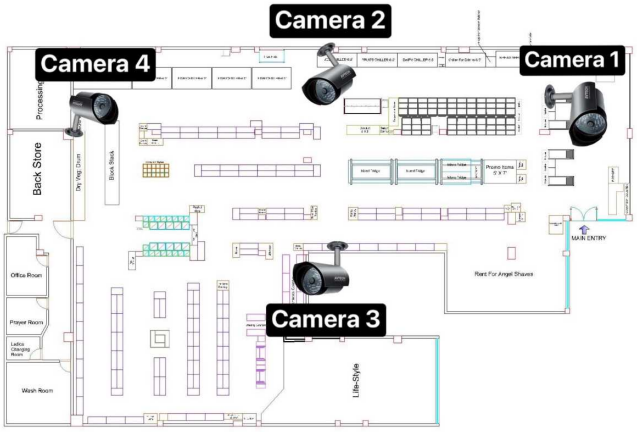


Fig. 1: FloorPlan and Camera positions of a Retail Outlet

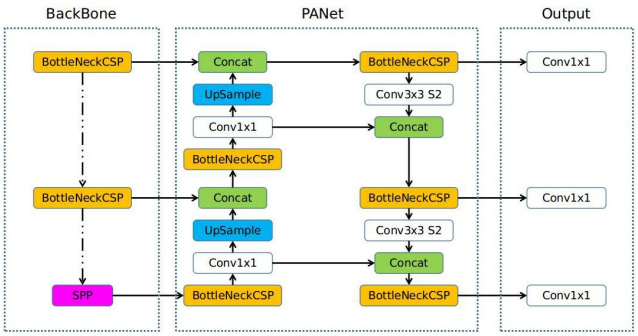
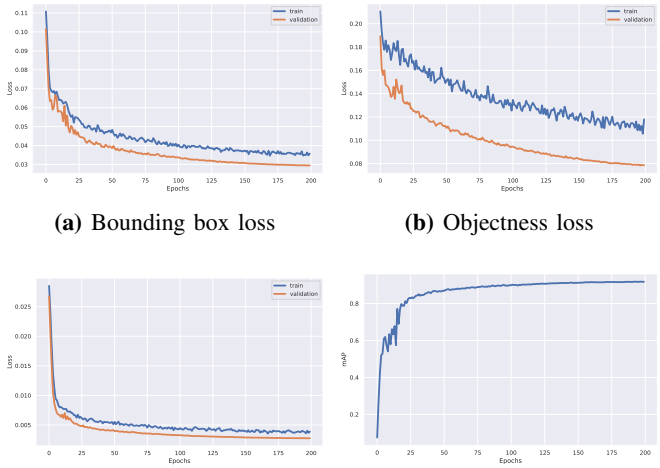


Fig. 3: YOLOv5 Model Architecture



(a) Bounding box loss **(b)** Objectness loss

(c) Classification loss **(d)** mAP(0.5)

Fig. 4: Training Losses

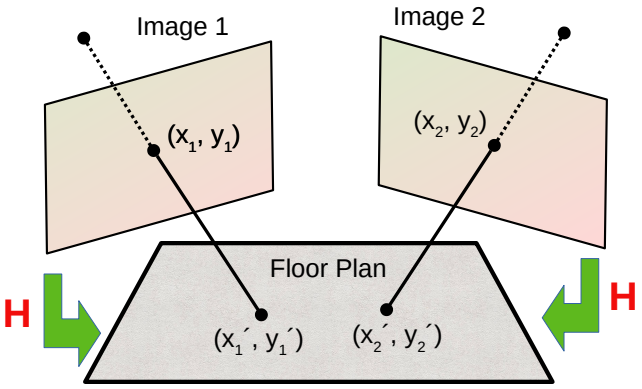
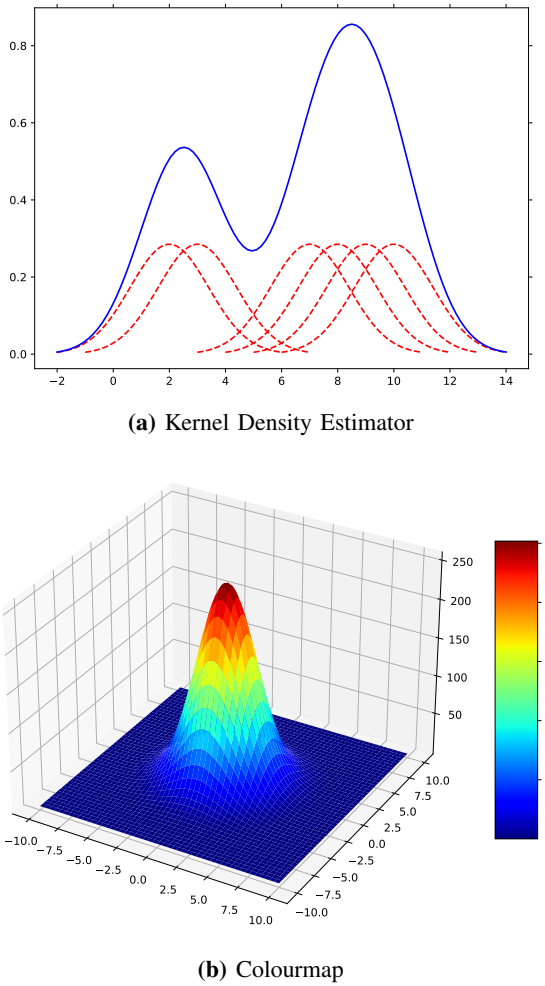


Fig. 5: Coordinate Transformation



(a) Kernel Density Estimator

(b) Colourmap

Fig. 6: Heatmap Generation

of an image. It is represented by a 3x3 transformation matrix(H) in a homogenous coordinates space. Mathematically, the Homography matrix is represented as:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = H \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (1)$$

where (x, y) is the coordinate of a person's position in camera image, and (x', y') is the transformed coordinate in Floorplan (Fig. 5). To calculate the matrix H, we need at least 4 point pairs from the two images(camera images and floorplan). The more point pairs we provide, the better the estimate of matrix H will be.

D. Heatmap Generation

A heatmap is a visual representation of data where the values of the data are shown by colors. We implement Gaussian Kernel and Kernel Density Estimator (KDE) [16] for Heatmap generation. Using a suitable kernel, Kernel density estimates can be equipped with properties such as smoothness or continuity. Normal kernels with appropriate variance (red dashed lines) are fitted on each of the data points to measure kernel density. when Gaussian kernels overlap, their values are accumulated together (solid blue curve in Fig. 6a). Heatmap is generated based on the normalized gaussian kernels which are transformed into a continuous range from [0, 255]. As the value goes from 0 to 255, the heatmap color goes from blue to red (Fig. 6b).

A workflow of the proposed system is shown in Fig. 7.

IV. Experimental Results

We implement the system in python 3.6 on an Intel(R) Core(TM) i7-1165G7 2.80GHz processor, 8GB RAM, and 2GB NVIDIA GeForce MX330 GPU. The program is capable of running on both CPU and GPU, but the processing speed on GPU is much faster than CPU. We process the videos at 8 FPS on our system which is suitable for real-time analysis.

Our proposed system consists of two stages. The first is to capture images from the CCTV videos to detect persons.

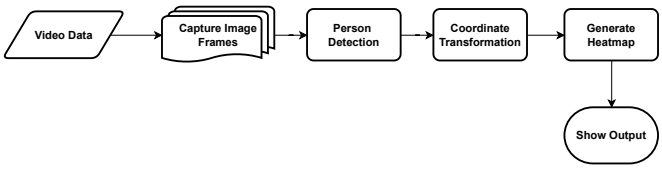


Fig. 7: Workflow of the proposed method

(a) camera 1 **(b)** camera 2

(c) camera 3 **(d)** camera 4

Fig. 8: Detected Persons in multiple cameras

Videos from CCTV cameras can be processed both online using RTSP streaming link, or offline using video files. After capturing a frame from the camera, it is fed to the yolov5 model for detecting all the persons in that image. We use a threshold of 0.3 for detecting object followed by a non-max-suppression technique with a threshold of 0.4 to keep the bounding boxes only that has higher class probabilities. Fig.8 shows the output of detecting persons associated with each camera.

Our second stage involves creating a heatmap on the floorplan to show how active people are in different areas. This is achieved by implementing Homography Transform, which converts the coordinates of all detected persons in the images into a floorplan. Heatmap generation on the Floorplan is the second part of our proposed algorithm, as depicted in Fig.9 where the black dots represent people's position over the entire floorplan and The red areas indicate people's engagement in corresponding areas.

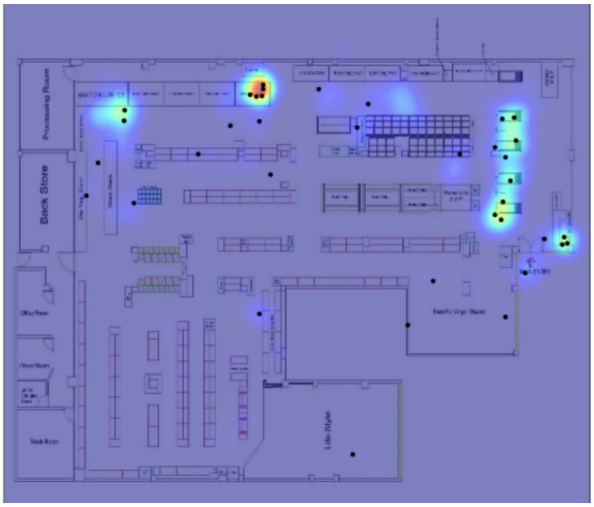


Fig. 9: Generated Heatmap on Floormap

V. Conclusions

In this paper, we present a method to detect people and find the hot zones in the floorplan in real time from a CCTV camera network. By spotting people in a retail shop area and determining what kinds of shops, brands, and products are more intriguing to customers, the suggested strategies can also be useful for controlling customer behavior in shopping centers. Thus, company management can quickly adjust the way the sale area operates by analyzing the behavior of the consumer. In the future, we hope to evaluate and separate the number of visitors to various locations as well as identify groups of people.

VI. Acknowledgement

This research is a part of our work at Advanced Chemical Industries (ACI) Limited. Video data, Floor-plan, and All the Technical support were provided by ACI Logistics Limited. We ensure that during the video processing No one's privacy was violated.

References

[1] M. Paul, S. M. E. Haque, and S. Chakraborty, "Human detection in surveillance videos and its applications - a review," *EURASIP J. Adv. Signal Process.*, vol. 2013, pp. 1–16, Dec. 2013.

[2] M. A. Ansari and D. K. Singh, "Human detection techniques for real time surveillance: a comprehensive survey," *Multimed. Tools Appl.*, vol. 80, pp. 8759–8808, Mar. 2021.

[3] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object Detection With Deep Learning: A Review," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30, pp. 3212–3232, Jan. 2019.

[4] A. J. Newman, D. K. C. Yu, and D. P. Oulton, "New insights into retail space and format planning from customer-tracking data," *Journal of Retailing and Consumer Services*, vol. 9, pp. 253–258, Sept. 2002.

[5] B. Deepak, "Real-time object detection and tracking using color feature and motion," 04 2015.

[6] E. N. Kajabad and S. V. Ivanov, "People detection and finding attractive areas by the use of movement detection analysis and deep learning approach," *Procedia Computer Science*, vol. 156, pp. 327–337, 2019. 8th International Young Scientists Conference on Computational Science, YSC2019, 24-28 June 2019, Heraklion, Greece.

[7] M. Parzych, A. Chmielewska, T. Marciniak, A. Dabrowski, A. Chrostowska, and M. Klincewicz, "Automatic people density maps generation with use of movement detection analysis," in *2013 6th International Conference on Human System Interactions (HSI)*, pp. 26–31, 2013.

[8] N. S. Punn, S. K. Sonbhadra, S. Agarwal, and G. Rai, "Monitoring COVID-19 social distancing with person detection and tracking via fine-tuned YOLO v3 and Deepsort techniques," *arXiv*, May 2020.

[9] M. Z. Khan, M. U. G. Khan, T. Saba, I. Razzak, A. Rehman, and S. A. Bahaj, "Hot-Spot Zone Detection to Tackle Covid19 Spread by Fusing the Traditional Machine Learning and Deep Learning Approaches of Computer Vision," *IEEE Access*, vol. 9, pp. 100040–100049, July 2021.

[10] G. Jocher, A. Stoken, J. Borovec, NanoCode012, Christopher-STAN, L. Changyu, Laughing, tkianai, A. Hogan, lorenzomammana, yxNONG, AlexWang1900, L. Diaconu, PetrDvoracek, P. Rai, *et al.*, "ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements," Oct. 2020.

[11] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "Crowdhuman: A benchmark for detecting human in a crowd," *arXiv preprint arXiv:1805.00123*, 2018.

[12] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, "CSPNet: A New Backbone that can Enhance Learning Capability of CNN," *arXiv*, Nov. 2019.

[13] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path Aggregation Network for Instance Segmentation," *arXiv*, Mar. 2018.

[14] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv*, Apr. 2018.

[15] F. Rovira-Más, J. Reid, and Q. Zhang, "Stereovision data processing with 3 d density maps for agricultural vehicles," *Transactions of the ASABE*, vol. 49, 07 2006.

[16] Y.-C. Chen, "A tutorial on kernel density estimation and recent advances," 2017.