Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.029719
Article

Multiple Pedestrian Detection and Tracking in Night Vision Surveillance Systems

Ali Raza[1], Samia Allaoua Chelloug[2,*], Mohammed Hamad Alatiyyah[3], Ahmad Jalal[1] and Jeongmin Park[4]

[1]Department of Computer Science, Air University, Islamabad, 44000, Pakistan
[2]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia
[3]Department of Computer Science, College of Sciences and Humanities in Aflaj, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia
[4]Department of Computer Engineering, Tech University of Korea, 237 Sangidaehak-ro, Siheung-si, Gyeonggi-do, 15073, Korea
*Corresponding Author: Samia Allaoua Chelloug. Email: sachelloug@pnu.edu.sa

Received: 10 March 2022; Accepted: 09 November 2022

Abstract: Pedestrian detection and tracking are vital elements of today's surveillance systems, which make daily life safe for humans. Thus, human detection and visualization have become essential inventions in the field of computer vision. Hence, developing a surveillance system with multiple object recognition and tracking, especially in low light and night-time, is still challenging. Therefore, we propose a novel system based on machine learning and image processing to provide an efficient surveillance system for pedestrian detection and tracking at night. In particular, we propose a system that tackles a two-fold problem by detecting multiple pedestrians in infrared (IR) images using machine learning and tracking them using particle filters. Moreover, a random forest classifier is adopted for image segmentation to identify pedestrians in an image. The result of detection is investigated by particle filter to solve pedestrian tracking. Through the extensive experiment, our system shows 93% segmentation accuracy using a random forest algorithm that demonstrates high accuracy for background and roof classes. Moreover, the system achieved a detection accuracy of 90% using multiple template matching techniques and 81% accuracy for pedestrian tracking. Furthermore, our system can identify that the detected object is a human. Hence, our system provided the best results compared to the state-of-art systems, which proves the effectiveness of the techniques used for image segmentation, classification, and tracking. The presented method is applicable for human detection/tracking, crowd analysis, and monitoring pedestrians in IR video surveillance.

Keywords: Pedestrian detection; machine learning; segmentation; tracking; verification

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1 Introduction

In recent years, there has been a significant influx of interest in smart surveillance systems [1], autonomous vehicles, self-driving cars, smart home systems, [2,3] and security. In particular, smart surveillance [4] can help us to improve traffic management. It can optimize the use of public resources and can be achieved with fewer human resources, while autonomous vehicles can help us to reduce the number of deaths caused by manned vehicles each year [5]. According to published surveys, road accidents and fatalities caused due to vehicle crash almost double each year. Night-time is usually considered the most vulnerable time because crimes mainly occur at night. Crime detection via real-time surveillance aids in reducing crimes for the emergency service department. However, pedestrian detection is challenging due to the non-rigidity of the human body. Some detection challenges are related to human appearance variations [6], changes in illumination, occlusions [7], movements, and background noise. As pedestrian detection at night is considered a complex problem, this paper aims to develop a vision-based surveillance system for detecting and tracking pedestrians at night.

The need to track humans is vital, and pedestrian tracking is an important part of surveillance systems [8]. Tracking can also be achieved by pose estimation [9–11]. Tracking utilized with facial recognition [12,13] can also determine specific targets and suspicious activities.

More essentially, applications for surveillance systems should be operating continuously. However, external illumination affects the standard cameras producing visible spectrum images. Thus, standard cameras are inappropriate when sufficient external illumination is absent, especially at night. Indeed, IR cameras are more appropriate for capturing images under these conditions as they can sense the radiation emitted from objects of interest, such as humans and vehicles. More importantly, the design of IR cameras has been improved, and the prices are decreasing rapidly.

In this paper, a method for detecting and tracking pedestrians has been proposed using machine learning techniques. Given the IR image sequences, we applied background subtraction using the method explained in [14]. Next, we performed segmentation to differentiate between pedestrian and non-pedestrian objects using a random forest classifier, resulting in the extraction of pedestrians, roofs of buildings, poles, and backgrounds. Then, multiple template-matching algorithms are used for pedestrian localization through a list of defined templates. Finally, pedestrians are tracked using a particle filter algorithm, which handles the association using Euclidean distance. Each pedestrian is then provided with an identification number investigated in the tracking process. Pedestrians are tracked using the particle filter technique, and finally, the pedestrians are verified using pixel values.

Our paper includes the following sections: Section 2 explains the previous work conducted on this topic. Section 3 helps to understand the methodology of the proposed system and explains individual methods in detail. Section 4 explains the results, and finally, Section 5 concludes the paper and defines future work.

2 Related Work

A video surveillance system is also known as a monitoring system for public and private situations using cameras. In the past, an introductory video surveillance system was used to acquire the signal through one or more cameras. The basic video surveillance system also converts the signal before it is displayed over a monitor or centrally recorded. Currently, the advancement of video surveillance systems has permitted the automation of data analysis. Video surveillance systems are employed to provide security by reacting to all potential occurrences [15,16].

However, these systems still lack precision for real-time reactions that may be delayed. Moreover, environmental changes pose a restriction for the system. These obstacles are problematic for video surveillance systems. By evaluating and comprehending the recorded films, processing data utilizing numerous proposed algorithms may aid in avoiding these issues through analysis and comprehension. Namely, motion detection, tracking events, anomaly detection in the scenes, crowd behavior, face recognition, and object categorization; are important processes that enable the extraction of information to manage emergencies and safeguard the safety of people.

Video surveillance in a private scene (workspace) has not progressed much in the literature, and it continues to function with entire systems [17]. The rationale is that surveillance in the workplace is limited to counting employees, estimating hazards, and evaluating safety [18]. Fuller [19] utilized video technology to estimate the probability of dangers and the labor performed by employers. In [20], the authors promoted using video technology to analyze the association between football player injuries and conduct.

Instead of waiting for an incident to occur, deploying video surveillance systems in public areas can foresee and prevent risks, such as accidents. Shariff et al. [21] suggested a strategy for reducing the likelihood of accidents using behavior-based safety measures (BBS). Gui et al. also attempted to lower the frequency of dangerous activities [22]. Behavior recognition methods are employed to enhance the performance of employees [23]. In this regard, video sequences are the best option and the most effective source of data for enhancing the performance of factory employees in terms of safety, as they enable regular observation of events and recording along with a review of crucial times. In addition, video recordings provide businesses with opportunities to increase the security of their activities and demonstrate the most dangerous behaviors and actions.

Some authors have suggested a surveillance-based system that uses audio and video information for security, defense, and fighting terrorism [24,25]. This system depends upon a pan-tilt-zoom (PTZ) camera that can record human body video. A laser doppler vibrometer (LDV) [26,27] was also used to get audio from a distance by detecting how objects move. The system needed a theodolite to control how LDV is set up. The good thing about this system is that it uses both sound and sight to find out what is happening, making it easier to find threats. The audio may also have much information about a possible event that can hurt people, but most existing video surveillance systems only showed parts of what was being watched [28,29]. Systems that rely on what you see are sensitive to changes in the environment and noise that affect how data is sent.

Video surveillance system constraints include data transfer. To reduce this limitation, Li et al. [16,30] presented an architecture for a video surveillance system based on throughput characteristics of WiMAX (AW) and Hybrid WIFI-WiMAX (HWW) systems. The AW system allows WiMAX IP cameras to connect directly to stations. The HWW system employs WIFI IP cameras connected to customer premises equipment (CPE) that is also linked to the base station [31,32]. The remote visualization was created over the internet, but the local visualization involves connecting PCs to CPE. Ethernet cables or the wireless interface may be used for data transmission. This occurs by capturing the data (using a WiMAX or WIFI IP camera) and data conversion into digital signals. The strength of these systems is their capacity to transmit data remotely, even if the camera is located many kilometers from the base station. However, the system's drawbacks are the WiMAX wireless communication and the antenna alignment.

Inspired by advancements in other fields, video surveillance systems are continuously evolving in terms of the utilized equipment and the performed analysis, particularly with the application of machine learning and deep learning algorithms.

3 System Design and Methodology

The methodology of our system includes background extraction, semantic segmentation, object detection, object tracking, and verification. The system first extracts frames from the video that has to be analyzed. The first step of the proposed system consists of applying the background subtraction technique used to distinguish between foreground and background. After background subtraction, the image segmentation process is used to segment the video frames into different parts. Next, segmentation is performed using machine learning classifiers, and the acquired segmented result is then utilized for pedestrian detection. In segmentation, a random forest algorithm compares leaf nodes with the parent node precisely. The latter is performed using multiple template matching. After pedestrian detection, the generated result is provided to the particle filter for tracking. Finally, the pedestrian is verified by comparing the tracked object's pixel values with the pedestrian's hotspot pixel value in IR images. Fig. 1 explains the design of the proposed system methodology.

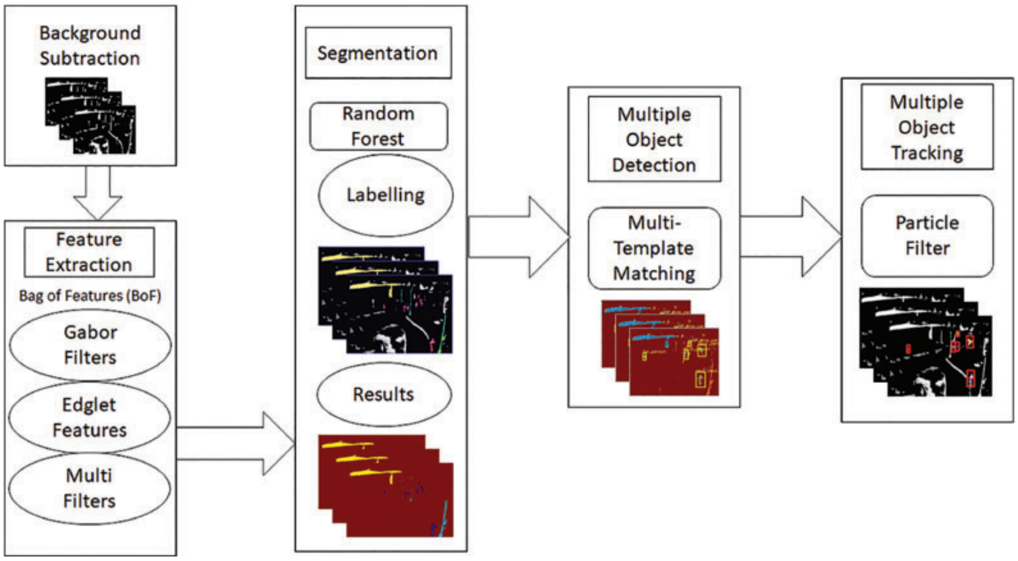

Figure 1: Proposed system architecture

3.1 Background Subtraction Using Yen Thresholding Method

Background subtraction is a process that allows an image's foreground to be extracted for further processing. It is performed using changes occurring in the foreground or calculating the difference between foreground and background pixel values. For background subtraction, Yen's thresholding technique has been adopted. This is an automatic multi-level thresholding technique.

$$P_i = \sum\nolimits_{j=t_{i-1}}^{t_i-1} P_j \tag{1}$$

where Pi is the normalized probability at level j. The automatic multi-level thresholding technique is used extensively in image processing. Image thresholding methods can be divided into two categories that are parametric and non-parametric approaches. In parametric approaches, a statistical model is initially assumed, and a set of parameters that control the model's fitness are found using a histogram. In non-parametric methods, thresholds are selected by optimizing an objective function, like maximizing between-class variance. Non-parametric approaches are more error-free and vigorous than parametric ones. Fig. 2 shows the original image and the image with the background subtracted, respectively. The mean intensity of the whole image and between-class variance is approximately determined by:

$$\mu_T = \sum\nolimits_{i=0}^{L-1} i \ . \ p_i = \sum\nolimits_{k=0}^{K-1} \mu_k \omega_k \tag{2}$$

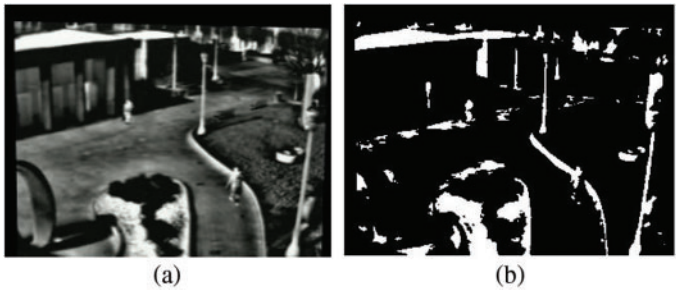

Figure 2: (a) Original image with pedestrians and background and (b) background subtraction results

3.2 Scene Segmentation: Random Forest

Segmentation is the process that labels each pixel in an image with a corresponding class. Next, a machine learning algorithm is used with the original image, and a labeled image is provided to predict the label for each pixel and define every class. We utilized a random forest classifier for class prediction. Gabor filter, pixel features, Canny, Roberts, Sobel, Scharr, Prewitt, gaussian, and median filter, were selected. These filters serve as the random forest classifier for segmentation. Fig. 3a represents the background subtracted image, while Fig. 3b shows the segment of the roof, poles, pedestrians, and background with different labeled pixels.

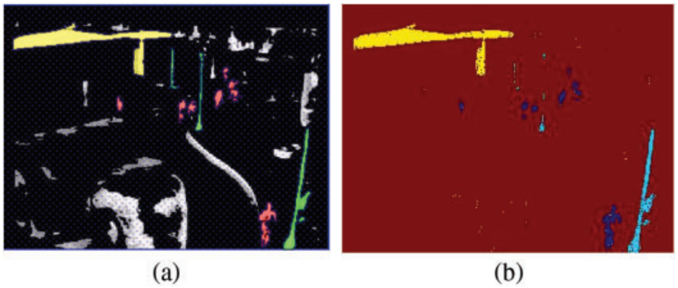

Figure 3: (a) The labels for segmentation (b) the segmented image containing the roof, poles, pedestrians, and background

We had five classes for segmentation, i.e., ground, pedestrian, poles, roof, and background. Fig. 4a represents the Gabor filters and their importance in the segmentation task, while Fig. 4b shows the rest of the features and their importance for segmentation.

The ROC curve graph shows the performance of the random forest classifier at different classification thresholds. Fig. 5 shows the ROC curve of the selected classifier used for segmentation. The average accuracy of the segmentation task is 75%. The classifier could not segment the ground region, but segmented pedestrians, roofs, poles, and ground with relatively higher accuracy as the area under the curve (AUC) is very high.

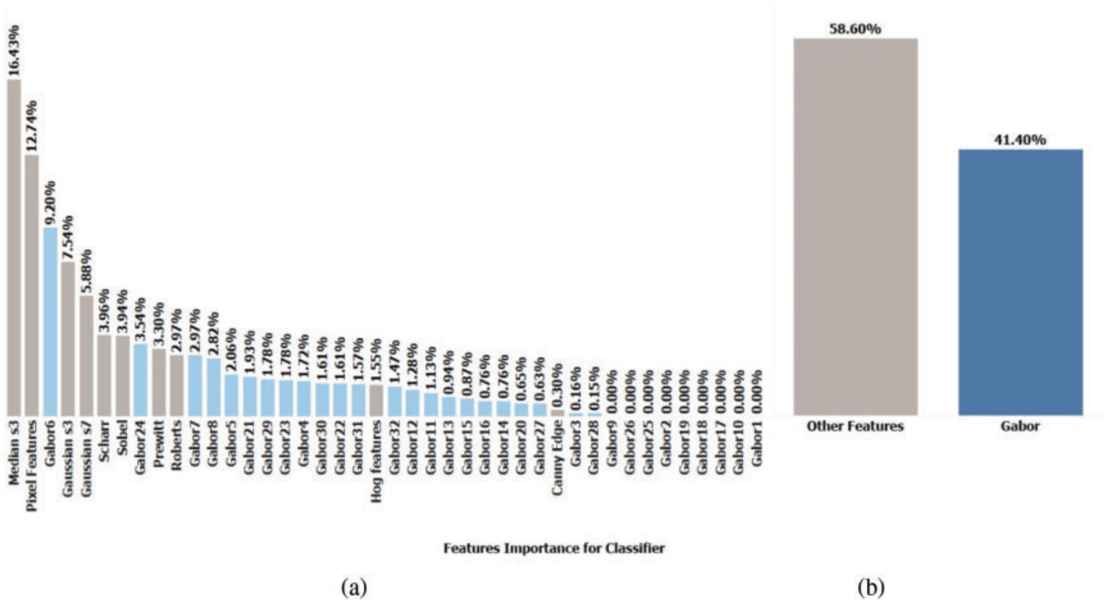

Figure 4: (a) Gabor filters and their importance in the segmentation task, (b) sobel, roberts, perwitt, gaussian, etc., and their importance in the segmentation task

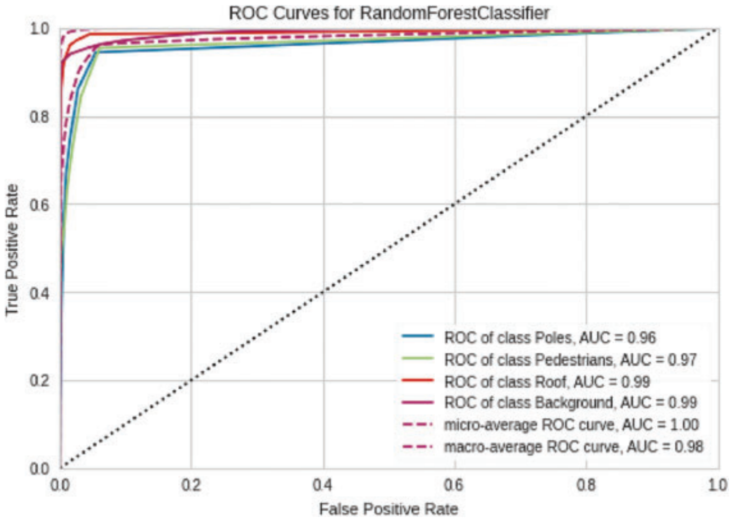

Figure 5: ROC curve for classifier used for the segmentation task

3.3 Multiple Template Matching

Multiple template matching is utilized to accomplish object detection on a particular image. It uses the list of templates as a testimonial to identify objects that are similar to the present in a particular image. It works by sliding the template across the original image. As it slides, the template is differentiated from the chunk of the image under it. It performs matching by calculating the dimensions that are similar to the template and the chunk of the original image. The multiple template matching method uses a correlation method to find the association between the original image and the list of templates. The template is found if a correlation is more than a provided threshold value. Otherwise, the template is not found. The source image is converted into a grayscale image, and it is divided into portions of $n \times n$ matrix. Then, the sliding window is used, and the template is moved across all the portions of the source image to find similarities between a portion of the source image and the template image. The similarity metric calculated is stored in a matrix, where the portion with the highest similarity value is considered the target region on the source image. If the source image has more than one portion with a high similarity value, the maximum matching value has been assigned to more than one portion of the source image. Eq. (3) is utilized to find the maximum correlation of portions of the source image for the template image.

$$Correlation = \sum_{i=0}^{N-1} \frac{(x_i - \overline{x}) \, . \, (y_i - \overline{y})}{\sqrt{\sum_{i=0}^{N-1} (x_i - \overline{x})^2 \, . \, \sum_{i=0}^{N-1} (y_i - \overline{y})^2}} \tag{3}$$

where x is the grey template image, $\overline{x}$ gives the average grey level in the template image, y provides the grey source image, $\overline{y}$ suggests the average grey level in the source image, and N represents the template image size.

In Algorithm 1, the multiple template matching algorithm takes the raw images and templates as input. It then calculates the correlation between the raw images divided into blocks of size $n \times n$, and the template images. The correlation results are saved in matrix form, and the most significant correlation is further selected as the matched region. The algorithm returns this matched region as coordinate points of a rectangular region, representing the detected template region on the original image.

Algorithm 1: Multiple Template Matching

```
Input: raw frames, templates
Output: ROI coordinates (x, y, w, h) in preprocessed frames;
begin
    1. for i in range (total frames)
    2. for j in range (total templates)
    3. convert_to_grey (frame)
    4. convert_to_grey (templates)
    5. size_of_image ← size (image);
           size_of_template ← size (templates)
           if (size_of_template image > size_of_image)
           template ← resize (template)
           else
           %Calculate similarity metric.%
```

(Continued)

Algorithm 1: Continued

```
            image_parts ← blocks (nxn, image)
            correlation (frame(i) ↔ templates(j))
            similarity_metric ← matrix (correlation).
            highest_value ← greatest (correlation);
            if (highest_value > 0.8)
             (x, y, w, h) ← matched (template)
            else
            template_not_matched
            end if
            end if
6. end for
7. end for
8. Return coordinates
```

After multiple template matching, we can detect pedestrians from our frame images. Figs. 6a and 6b show examples of the templates used. Figs. 6c and 6d display the image with detected pedestrians.

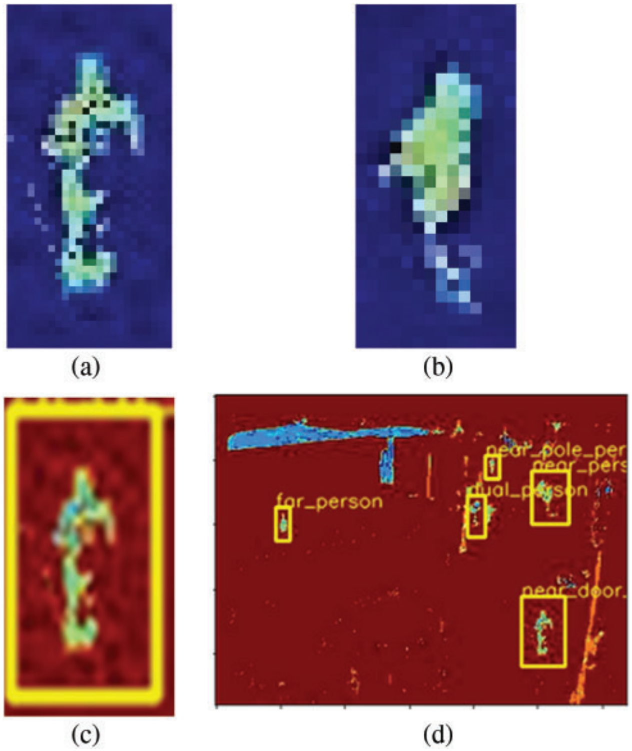

Figure 6: (a) Template used for multiple template-matching algorithms, (b) another template used for multiple template matching, (c) region where a single template is matched, and (d) detected pedestrians after multiple template-matching algorithms

3.4 Particle Filter Tracking Module

Target tracking in forward-looking infrared (FLIR) video sequences is an important but challenging topic in computer vision. Unlike visible light spectrum videos, the difficulties of tracking FLIR videos lie in the low signal-to-noise ratio (SNR), poor target visibility, completing background clutter, and abrupt motion incurred by high motion. The tracking system has two typical challenges in FLIR video tracking. The first challenge is related to modeling the target appearance to adapt to the appearance changes in case of a cluttered background. The second challenge consists of dealing with the drastic abrupt motions incurred by the sensor. Modeling the target appearance is always a challenging issue in visual tracking. The visual features can vary from low-level features to high-level semantic features. In the context of FLIR tracking, the frequently used visual features for appearance modeling include intensity histogram, standard variance histogram, and edge along with shapes. Although these features have shown their effectiveness to the scale changes and slow varying appearance, it is still hard to solve and model appearance changes incurred by intensity and size variations over a long time.

Detection of targets in the FLIR sequences is a complex problem due to the instability of targets. Hence, the thermodynamic state of the targets, atmospheric conditions, and background are the most critical factors affecting the stability of targets. Most of the time, the background outcome shapes are identical to those of the actual target, and the targets will be hidden. This paper focuses on target tracking and establishing initial target detection. The implementation focuses on the hot targets, which appear as bright spots in the scene with high contrast when compared to the neighboring background.

In a particle filter, we approximate the posterior $p\left(x_t|y_{o:t}\right)$ with a Dirac measure for finite N particles $\left\{x_t^i\right\}_{i=1\ldots N}$. Candidate particles are sampled from an appropriate proposal distribution $\widetilde{x}_t^i \sim q\left(x_t|x_{0:t-1}, y_{0:t}\right)$, and the particles are weighted according to their importance using the following ratio:

$$w_t^i = w_{t-1}^i \frac{p\left(y_t|\widetilde{x}_t^i\right) p\left(\widetilde{x}_t^i|x_{t-1}^i\right)}{q\left(\widetilde{x}_t^i|x_{0:t-1}^i, y_{0:t}\right)} \tag{4}$$

Given a general non-linear discrete-time system.

$$s\left(k+1\right) = f\left(s\left(k\right), t\left(k\right)\right) + g(s\left(k\right), t\left(k\right) w\left(k\right)) \quad k \epsilon N \tag{5}$$

where $s\left(k\right)$ $\mathrm{P_X}^n$ is the state of the system

$$z\left(k\right) = h\left(s\left(k\right), t\left(k\right)\right) + v\left(k\right) \quad k \epsilon N \tag{6}$$

$$Z\left(k\right) = \left\{z\left(0\right), \ldots, z\left(k\right)\right\} \tag{7}$$

$z\left(k\right)$ is the measurement for process noise and measurement noise densities $Z\left(k\right)$ is the set of measurements.

Algorithm 2: Particle filter tracking

```
1. Initialization:
        Draw a set of particles for the prior $p\left(X_o\right)$ to obtain $\left\{X_o^{(i)}, w_o^{(i)}\right\}_{i=1}^N$ where $w_o^{(i)}$ is the weight for
    the particle $X_o^{(i)}$, let $k = 1$.
2. Sampling
    a. For samples $i = 1, \ldots, N$
            Sample $X_k^{(i)}$ from the proposal distribution
                            $p\left(X_k^{(i)}, X_{k-1}^{(i)}\right)$
            These are sample points randomly distributed
```

(Continued)

Algorithm 2: Continued

```
    b. Evaluate the new weights
            $W_k^{(i)} = p(Z_k | W_k^{(i)}),\ i = 1, \ldots, N$
    c. Normalize the weights
            $W_k^{(i)} = \frac{W_k^{(i)}}{\sum_{j=1}^{N} W_k^{(j)}},\ i = 1, \ldots, N$
        The initial points are moved according to the evaluated weights
3. Output a set of particles $\left\{X_k^{(i)},\ W_k^{(i)}\right\}_{t=1}^{N}$ that can be used to approximate the posterior distribution as
            $P\left(X_k | Z^k\right) \approx \sum_{i=1}^{N} W_k^{(i)} \delta\left(X_k - X_k^{(i)}\right\},$
here we recursively update the posterior distribution over the current state $X_t^{(i)}$ given all the observations $Z^k = Z_1, \ldots, Z_k$ and the estimate as
            $E_{p(g|Z^k)}\left(f_k\left(X_k\right)\right) \approx \sum_{i=1}^{N} W_k^{(i)} f\left(X_k^{(i)}\right\}$
which is the Monte Carlo approximation of the integral, where $\delta\,(g)$ is the Dirac delta function.
4. Resample particles $X_k^{(i)}$ with probability $w_t^{(i)}$ to obtain $N$ independent and identically distributed random particles $X_k^{(j)}$, approximately distributed according to $P(X_k | Z^k)$
5. $k = k + 1$, go to step2.
```

Hence, Fig. 7 shows the effectiveness of particle filter algorithm for tracking diffrent number of pedestrains.

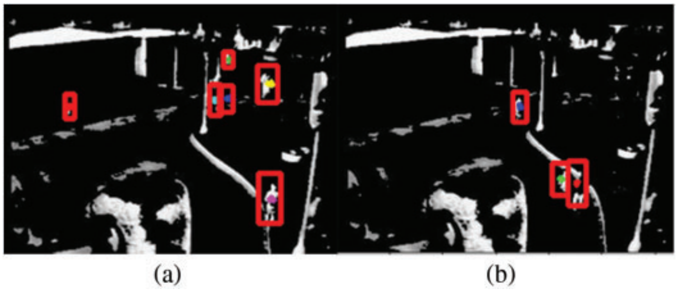

Figure 7: (a) Particle filter tracking result with six tracked pedestrians, (b) particle filter tracking result with three tracked pedestrians

4 Experimental Results and Analysis

The experimental section is divided into two sections. The first section describes the acquired dataset, while the second section evaluates the performance of the proposed system over the Ohio State University (OSU) Thermal Pedestrian Database. Furthermore, the proposed system is assessed with segmentation accuracy, precision, recall, F1-score, and support. Detection accuracy is represented by the accuracy of detection in multiple sequences. Finally, the system's tracking accuracy is calculated using the multiple-object tracking accuracy model.

4.1 OSU IR Dataset Description

The dataset utilized is the OSU Thermal Pedestrian Database. It is part of the OCTVBS benchmark dataset. The number of sequences in this dataset is 10. The proposed system used the first sequence, which has 1057 images, where each image is an 8-bit grayscale bitmap with 360 × 240 resolution. The dataset shows pedestrian intersections at Ohio state university with multiple pedestrians on campus. Our sequence contains a minimum of two pedestrians and a maximum of 10 pedestrians in a single frame. The scenery of the images has ground, different pedestrians, street lights, and campus buildings.

4.2 Performance Analysis of Proposed System

The segmentation process was achieved using the random forest classifier to segment the background-subtracted images. The segmentation gives segmented region classes of background, roofs, poles, and pedestrians. Table 1 shows the classification report for our classifier, and Table 2 represents the confusion matrix over the dataset. It is clear that random forest achieved better results for the roof and background classes.

Table 1: Classification report of RF model over OSU IR dataset

Number of objects	Precision	Recall	F1-score	Support
Poles	0.62	0.19	0.29	108
Pedestrians	0.82	0.36	0.50	154
Roof	0.92	0.83	0.87	354
Background	0.98	1	0.99	14744
Accuracy	0.83	0.60	0.76	1566

Table 2: Confusion matrix for RF classifier over OSU IR dataset

Classes	Poles	Pedestrians	Roof	Background
Poles	20	6	2	80
Pedestrians	2	55	1	96
Roof	0	0	294	60
Background	10	6	23	14705

In Table 3, we observed pedestrian detection using a multiple template matching algorithm. Due to the changing shape of the pedestrians, pedestrian detection works only when a new pedestrian enters a frame. It enhances the accuracy by up to 87%. For a given time interval, if we made detections through a multiple template matching algorithm and if the shape of the pedestrians did not match the template in the list of templates, then the algorithm failed to detect the pedestrian. The accuracy result of the pedestrian detection method is given in Table 3.

Table 3: Detection accuracy of MTM model

Frame sequences	Pedestrians	Ground truth	Detected	Accuracy
Sequence 1 (80 frames)	2	160	160	100%
Sequence 2 (121 frames)	3	363	363	100%
Sequence 3 (12 frames)	4	48	48	100%

(Continued)

Table 3: Continued

Frame sequences	Pedestrians	Ground truth	Detected	Accuracy
Sequence 4 (24 frames)	5	120	108	90%
Sequence 5 (42 frames)	6	252	242	96%
Sequence 6 (31 frames)	8	248	234	94%
Sequence 7 (202 frames)	9	1818	1573	87%
Sequence 8 (364 frames)	10	3640	3030	83%
Average accuracy				87%

In Fig. 8, pedestrian tracking is done using a particle filter, and the results are calculated using a multiple-object tracking algorithm (MOTA). The results of MOTA are presented in Fig. 8, where the false negatives and false positives are reduced to minimum levels.

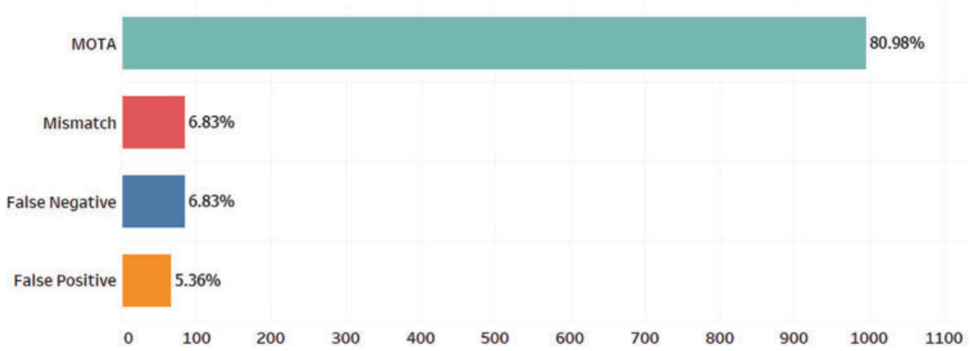

Figure 8: Tracking results of multiple object tracking

Table 4 illustrates the comparison with benchmarked techniques for pedestrian detection on the same dataset. EBV is the Edgelet-boosting-VS, HSV is HOG-SVM-VS, and HSI is HOG-SVM-IR. The proposed method outperformed the benchmarked techniques in terms of detection rate, while pedestrian detection based on Haar-Like features with an Adaboost filter is ranked as the second-best technique.

Table 4: Detection comparison with state of the art

Techniques	Detection rate (%)
EBV [33]	79.68%
HSV [33]	79.16%
HSI [33]	78.70%
Shape context based adaboost cascade [34]	70%
Haar-like features with adaboost filter [35]	86%
Proposed system	87%

5 Conclusion

In this paper, we have designed and implemented a pedestrian detection and tracking system that can accurately detect and predict the movements of pedestrians during the night by using multiple machine learning algorithms. First, the background and foreground are segmented via Yen's thresholding technique. After background subtraction, the image is segmented into multiple classes using a machine-learning classification model. Then, pedestrian detection and pedestrian tracks are applied over the segmented frames. Further, the pedestrian detector module uses a multiple template matching algorithm. It accurately detects pedestrians with an 87% accuracy rate in all the frames. In comparison, the tracking algorithm accurately tracks the pedestrians with an 81% accuracy rate through a particle filter for tracking in the frames and keeping the count.

We plan to enhance the model in the future by experimenting with deep learning algorithms [36–38] for further improvements. The accuracy rates will also be improved by using deep learning techniques and different models.

Acknowledgement: Thanks to Professor Ahmad Jalal for supporting experimentation sections at Air University, Islamabad.

Funding Statement: This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2018-0-01426) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). Also, this work was partially supported by the Taif University Researchers Supporting Project Number (TURSP-2020/115), Taif University, Taif, Saudi Arabia. This work was also supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R239), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Conflicts of Interest: The authors declare they have no conflicts of interest to report regarding the present study.

References

[1] K. A. Joshi and D. G. Thakore, "A survey on moving object detection and tracking in video surveillance system," *International Journal of Soft Computing and Engineering*, vol. 2, no. 3, pp. 44–48, 2012.

[2] A. Jalal, M. A. K. Quaid and M. A. Sidduqi, "A triaxial acceleration-based human motion detection for ambient smart home system," in *Int. Bhurban Conf. on Applied Sciences and Technology (IBCAST)*, Islamabad, Pakistan, pp. 353–358, 2019.

[3] A. Jalal, S. Kamal and D. S. Kim, "Detecting complex 3D human motions with body model low-rank representation for real-time smart activity monitoring system," *KSII Transactions on Internet and Information Systems*, vol. 12, pp. 1189–1204, 2018.

[4] F. Zhu, Y. Lv, Y. Chen, X. Wang, G. Xiong *et al.,* "Parallel transportation systems: Toward IoT-enabled smart urban traffic control and management," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, pp. 4063–4071, 2019.

[5] R. D. Brehar, M. P. Muresan, T. Mariţa, C. C. Vancea, M. Negru *et al.,* "Pedestrian street-cross action recognition in monocular far infrared sequences," *IEEE Access*, vol. 9, pp. 74302–74324, 2021.

[6] A. Jalal, A. Nadeem and S. Bobasu, "Human body parts estimation and detection for physical sports movements," in *Int. Conf. on Communication, Computing and Digital Systems (C-CODE)*, Islamabad, Pakistan, pp. 104–109, 2019.

[7] B. Leibe, E. Seemann and B. Schiele, "Pedestrian detection in crowded scenes," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, US, pp. 878–885, 2005.

[8] P. Viola, M. Jones and D. Snow, “Detecting pedestrians using patterns of motion and appearance,” in *Proc. Ninth IEEE Int. Conf. on Computer Vision*, Nice, France, pp. 734–741, 2003.

[9] A. Jalal, S. Kamal and D. Kim, “Depth silhouettes context: A new robust feature for human tracking and activity recognition based on embedded HMMs,” in *Int. Conf. on Ubiquitous Robots and Ambient Intelligence (URAI)*, Seoul, Korea, pp. 294–299, 2015.

[10] A. Jalal, N. Sarif, J. T. Kim and T. -S. Kim, “Human activity recognition via recognized body parts of human depth silhou-ettes for residents monitoring services at smart home,” *Indoor and Built Environment*, vol. 22, no. 1, pp. 271–279, 2013.

[11] A. Jalal and Y. Kim, “Dense depth maps-based human pose tracking and recognition in dynamic scenes using ridge data,” in *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, Seoul, Korea, pp. 119–124, 2014.

[12] M. Mahmood, A. Jalal and H. A. Evans, “Facial expression recognition in image sequences using 1D transform and gabor wavelet transform,” in *Int. Conf. on Applied and Engineering Mathematics (ICAEM)*, Taxila, Pakistan, pp. 1–6, 2018.

[13] J. Wang, D. Chen, H. Chen and J. Yang, “Pedestrian detection and tracking in infrared videos,” *Pattern Recognition Letters*, vol. 33, no. 6, pp. 775–785, 2012.

[14] A. Jalal and A. Shahzad, “Multiple facial feature detection using vertex-modeling structure,” in *Proc. of ICL*, Melbourne, Australia, pp. 1–7, 2007.

[15] T. Kongsvik, J. Fenstad and C. Wendelborg, “Between a rock and a hard place: Accident and near-miss peporting on offshore service vessels,” *Safety Science*, vol. 50, pp. 1839–1846, 2012.

[16] H. Li, M. Lu, S. Hsu, M. Gray and T. Huang, “Proactive behavior-based safety management for construction safety improvement,” *Safety Science*, vol. 75, pp. 107–117, 2015.

[17] D. Cook, J. Augusto and V. Jakkula, “Ambient intelligence: Technologies, applications, and opportunities,” *Pervasive and Mobile Computing*, vol. 5, pp. 277–298, 2009.

[18] T. McSween, *Values-Based Safety Process: Improving Your Safety Culture With Behavior-Based Safety*. John Wiley & Sons, New Jersey, USA, 2003.

[19] C. Fuller, “An assessment of the relationship between behaviour and injury in the workplace: A case study in professional football,” *Safety Science*, vol. 43, pp. 213–224, 2005.

[20] N. Bahr, *System Safety Engineering and Risk Assessment: A Practical Approach*. CRC Press, Boca Raton, USA, 2014.

[21] A. Shariff and N. Norazahar, “At-risk behaviour analysis and improvement study in an academic laboratory,” *Safety Science*, vol. 50, pp. 29–38, 2012.

[22] C. Gui, L. Kai, L. Jiao, S. Hua and Z. Jian, “Risk management and workers safety behavior control in coal mine,” *Safety Science*, vol. 50, pp. 909–913, 2012.

[23] A. Laureshyn, “Application of automated video analysis to road user behaviour,” *Denna Side Pa Svenska*, Thesis, Lung Unversity, Lung, Sweden, pp. 1–123, 2010.

[24] T. Lv, H. Zhang and C. Yan, “Double mode surveillance system based on remote audio/video signals acquisition,” *Applied Acoustics*, vol. 129, pp. 316–321, 2018.

[25] Y. Qu, T. Wang and Z. Zhu, “Remote audio/video acquisition for human signature detection,” in *Computer Vision and Pattern Recognition Workshops*, Miami, FL, pp. 66–71, 2009.

[26] H. Zhang, T. Lv and C. Yan, “The novel role of arctangent phase algorithm and voice enhancement techniques in laser hearing,” *Applied Acoustics*, vol. 126, pp. 136–142, 2017.

[27] M. Kyriakidis, R. Hirsch and A. Majumdar, “Metro railway safety: An analysis of accident precursors,” *Safety Science*, vol. 50, pp. 1535–1548, 2012.

[28] S. Andriulo and M. G. Gnoni, “Measuring the effectiveness of a near-miss management system: An application in an automotive firm supplier,” *Reliability Engineering & System Safety*, vol. 132, pp. 154–162, 2014.

[29] S. Lubobya, M. Dlodlo, G. Jager and A. Zulu, “Throughput characteristics of wimax video surveillance systems,” *Procedia Computer Science*, vol. 45, pp. 571–580, 2015.

[30] M. Yang, J. Tham, D. Wu and K. Goh, “Cost effective IP camera for video surveillance,” in *Industrial Electronics and Applications (ICIEA)*, Xi’on, China, pp. 2432–2435, 2009.

[31] B. Gumaidah and H. Soliman, “Wimax network performance improvement through the optimal use of available bandwidth by adaptive selective voice coding,” *International Journal of Modern Engineering Sciences*, vol. 2, pp. 1–16, 2013.

[32] J. Changjiang, W. Jin, L. Mingfu and L. Zhichuan, “A design and implementation of mobile video surveillance terminal base on arm,” *Procedia Computer Science*, vol. 107, pp. 498–502, 2017.

[33] S. Kamal, A. Jalal and D. Kim, “Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified HMM,” *Journal of Electrical Engineering and Technology (JEET)*, vol. 11, no. 6, pp. 1857–1862, 2016.

[34] A. Jalal, S. Kamal and D. Kim, “Shape and motion features approach for activity tracking and recognition from kinect video camera,” in *IEEE 29th Int. Conf. on Advanced Information Networking and Applications Workshops*, Gwangju, Korea, pp. 445–450, 2015.

[35] L. Zhang, B. Wu and R. Nevatia “Pedestrian detection in infrared images based on local shape features,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, USA, pp. 1–8, 2007.

[36] W. Wang, J. Zhang and C. Shen, “Improved human detection and classification in thermal images,” in *IEEE Int. Conf. on Image Processing*, Alaska, USA, pp. 2313–2316, 2010.

[37] W. Sun, L. Dai, X. R. Zhang, P. S. Chang and X. Z. He, “RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring,” *Applied Intelligence*, vol. 52, pp. 8448–8463, 2022.

[38] W. Sun, G. C. Zhang, X. R. Zhang, X. Zhang and N. N. Ge, “Fine-grained vehicle type classification using lightweight convolutional neural network with feature optimization and joint learning strategy,” *Multimedia Tools and Applications*, vol. 80, no. 20, pp. 30803–30816, 2021.