journal homepage: https://journal.pktm.com.my/

International Journal of Technical Vocational and Engineering Technology

e-ISSN2710-7094, Vol. 6, No. 2 (2025)

Optimized Occlusion Handling in Human Detection by Fusion of Thermal and Depth Images for Mobile Robots

Saipol Hadi Hasim[1], T I T Nadzion[1], W T W M Rumaizi[1], S Saifuddin[2], U U Sheikh[3], A Hidayat[4]

[1]Department of Electrical Engineering, Politeknik Ibrahim Sultan, KM 10, Jalan Kong Kong, Pasir Gudang, 81700, Malaysia

[2]Politeknik Besut Terengganu, Jalan Bukit Keluang, 22200 Besut, Terengganu, Malaysia

[3]School of Electrical Engineering, Faculty of Engineering, Universiti Teknologi Malaysia, Johor Bahru, 81310, Malaysia

[4]Department of Electrical Engineering, Politeknik Negeri Padang, Jl. Kampus, Limau Manis, Kec. Pauh, Kota Padang, Sumatera Barat 25164, Indonesia

*Corresponding author email: saipolhadi@pis.edu.my

ARTICLE INFO

Article History:
Received 4 July 2025
Revised 11 September 2025
Accepted 23 October 2025
Published 30 October 2025

Keywords:
human detection,
surveillance system,
machine learning,
segmentation,
fusion modalities

ABSTRACT

In the domain of machine vision, surveillance systems serve as a security measure aimed at protecting public safety and properties. A key function of these systems is human detection. This paper introduces a human detection system that leverages thermal-depth information captured by a mobile robot in indoor settings. A novel fusion technique, termed Fusion of Thermal-Depth Information (FTDI), is proposed to enhance the segmentation process, ensuring robustness in various lighting conditions and improving processing speed. To address the challenge of occlusion, a new method known as the Occlusion Human Detector (OCHD) is introduced, which incorporates a pre-detector. This detector classifies occluded individuals using pixel codes derived from a candidate selection process. The results indicate that the proposed system achieves over 90% average accuracy across all datasets, outperforming state-of-the-art algorithms. Its innovative contribution is enhancing the classification of individuals and their occlusions. The proposed system is noted for being computationally efficient and maintaining high performance even in conditions of significant occlusion and low illumination.

1.0 Introduction

In the realm of machine vision, surveillance systems serve as a security measure focused on ensuring the security of both the property and public. One critical function of these systems is human detection. Humans often take this ability for granted, as our brains are exceptional at learning new objects and recognizing them later. We can identify objects and interpret their interactions even within complex scenes. For instance, when navigating a public space, a person can swiftly recognize and categorize various elements, such as other individuals, furniture, stairs,

pillars, walls, and signage. Humans can detect individuals under a wide range of conditions, irrespective of clothing colour, style, posture, appearance, partial occlusions, lighting variations, or background distractions. The analysis of an individual's colour and viewpoint is generally not essential to determining that an object is a person. Similarly, in surveillance systems, this challenge has been identified as one of the most difficult and computationally demanding aspects of human detection. The complexity further increases when such systems are integrated into mobile robots.

Traditional surveillance systems that rely on static cameras operate in a passive manner, primarily detecting events and triggering alarms. In contrast, active role systems, such as those based on mobile platforms, can interact with their environment, humans, and other robots, significantly enhancing their capabilities (Hacinecipoglu et al., 2020) (Bakar; et al., 2022) (Murugiah; et al., 2024) and (Gupta et al., 2024). These robots can patrol areas like airports, warehouses, and banks, monitoring valuable assets, recognizing individuals, and identifying intruders. Currently, various strategies are being developed to improve the performance of human detection systems, making them more suitable for everyday applications such as surveillance systems, driver-assistance systems, and facial recognition security systems. A range of sensors is employed, including radar, lasers, thermal cameras, visible cameras, and depth sensors (Kinect). While radar sensors, thermal cameras, and lasers tend to be expensive, vision-based systems like standard cameras and depth sensors are more affordable and provide multi-channel information and pixel-level, including depth data and colour (RGB) (Ozcan & Cetin, 2022). Since early 2015, thermal cameras have gained popularity and become more accessible for public research due to a reduction in cost. Detecting and tracking humans is essential for advanced vision-based applications, especially when they integrate autonomy. A vital aspect of any person recognition system is the capability to detect individuals under all conditions, including instances of occlusion. Unfortunately, current object recognition methods only perform adequately at lower occlusion rates but struggle with moderately to severe occlusions. This paper addresses the challenge of occlusion handling in human detection deployed on mobile robots for surveillance systems in indoor environments.

2.0 Experimental

Surveillance systems that utilize mobile robots offer significant advantages over static platforms in addressing various surveillance tasks. These mobile systems can actively interact with their surroundings, humans, and other robots, enhancing their functionality and effectiveness. However, as the robot moves, the challenges associated with obtaining reliable information cues increase. The data collected can be subject to extreme noise, and the complexity of both the background and foreground can complicate analysis. To effectively implement surveillance systems using mobile robots, several critical problems must be addressed is handling unreliable or incomplete information cues, robust object segmentation, and real-time response. Detecting human presence is a key requirement in the context of surveillance systems; however, it presents a significant challenge due to the multifaceted appearance of humans. Additional complicating factors include variations in illumination, the presence of occlusions, changes in viewpoint, and background clutter. Each of these elements can hinder the ability of the system to accurately identify and track individuals, necessitating the development of robust detection algorithms that can adapt to these challenges. Addressing these issues is essential for enhancing the effectiveness of mobile robot-based surveillance systems as mentioned by (Akram et al., 2018), (Aslan et al., 2020) and (Liu et al., 2021). Illumination presents a challenge for RGB cameras, which have limitations under various lighting conditions and computational costs. A robust surveillance system should be able to function effectively in all the mentioned situations. Furthermore, relying on a single type of sensor restricts the system's capabilities compared to utilizing sensor fusion, which can generate multi-modal features to improve the system's robustness, as noted by (Pourmehr et al., 2017) (Park et al., 2020) (Guo et al., 2020) and (W. Zhang et al., 2021).

Another challenge that complicates surveillance tasks is the presence of crowded scenes and the occurrence of occlusions. A critical question arises: how can the proposed method detect individuals in occluded situations within chaotic real-life scenarios, particularly given the unpredictable movement of people? A review of the literature indicates that existing methods struggle with handling occlusions. Many of these approaches are capable of addressing occlusions, but they are often limited to scenarios involving only two or three individuals, ((Zhou & Yuan, 2019) (Chi et al., 2020) (Mo et al., 2021) (Gilroy et al., 2021) (Gilroy et al., 2021) and (Minaee et al., 2022)). Works by (Gupta et al., 2024) attempted to address occlusion involving five individuals, but challenges related to occlusion in crowded scenes remain unresolved. Moreover, earlier approaches show insufficient detection performance under moderate to severe occlusion conditions. Evidence suggests that when occlusion becomes too extreme, most of these methods tend to fail, ((Zhou et al., 2019) (Angelini et al., 2020) (Guo et al., 2020) (Mary et al., 2020) (Cao et al., 2017) (Stewart & Andriluka, 2016) and (Merad et al., 2016)).

2.1 Methodology

Figure 1 illustrates the proposed system for detecting individuals using a mobile robot. The algorithm comprises three stages, which will be described in detail in the next subsection: (i) pre-processing, (ii) region-of-interest (ROI) generation, and (iii) object classification, including the pre-detector and OCHD detector. Each stage will be described comprehensively in the following subsections.

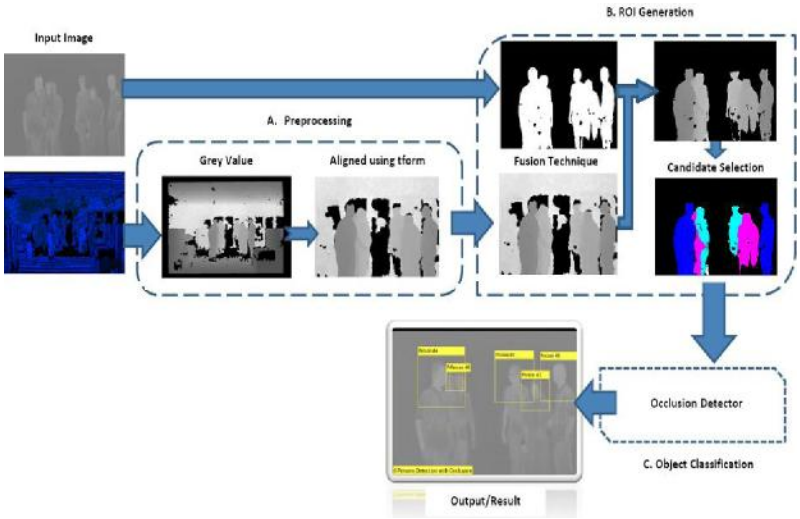


Figure 1. Proposed system procedure

2.1.1 Preprocessing

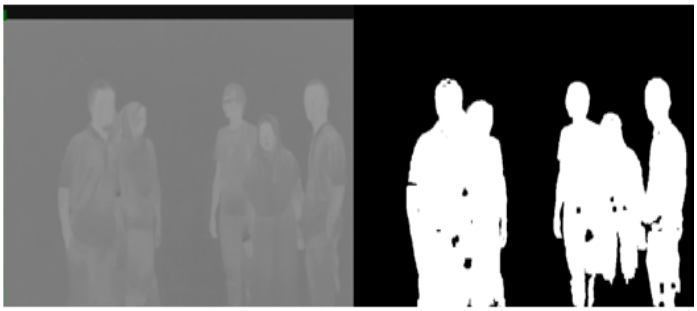
This stage involves handling the information obtained from the Kinect sensors and thermal cameras. Both raw data (depth and thermal) are manipulated to render them proper for subsequent stages. The next stage, ROI Generation, provides an initial or rough contour of the object of interest within the depth and thermal images, and the last stage, Object classification, relies on standard pattern recognition techniques based on previously extracted features. This proposed system is designed to automatically detect humans in the foreground, including those who are occluded. The object detection system incorporates several pre-trained classifiers for identifying profile frontal faces, noses, the upper body eyes, faces, and among others. However, these detectors may not be consistently effective for all applications. Consequently, training a custom classifier is essential. Cascade trainer applications, created by (Muhadi et al., 2020) and (Shoelson, 2013), can be found in the MATLAB® File Exchange and are included in the Computer Vision System Toolbox™ to facilitate the training of a custom classifier. These applications serve various purposes, including creating new cascade classifiers, training algorithms, and managing the selection and positioning of ROIs from a images collection.

2.2 Segmentation

The ROIs containing individuals in the thermal image are selected using approach described in (Panetta et al., 2021). The values of threshold are determined from the histogram using the Image Viewer Application within Image Processing Tools. This application offers various features for displaying images, optimizing the properties, numbers, and axes of the image objects for improved visualization. Research by (Lou et al., 2021) indicated that images of thermal post-thresholding are referred to as thermal of interest (TOI). In this context, the analysis is conducted based on the lower threshold value, t_l and the highest threshold value, t_h, as expressed in the following formula:

$$TOI(x,y) = \begin{cases} 0: & t_l \geq tml(x,y) \geq t_h \\ tml(x,y): & otherwise \end{cases} \quad (1)$$

Where *TOI(x, y)* denotes the thresholder image, while *tml(x, y)* represents the original thermal image pixels. If the measured value of thermal *tml(x, y)* is outside the t_h range or falls below the t_l range, it is set to zero; otherwise, the value of thermal is retained. Figure 2 illustrates the process of segmentation for an image of thermal.



(a) (b)

Figure 2. (a) Image of thermal; (b) Extraction of TOI based on tl = 135 and th = 160.

Generally, the Kinect sensor is set up to provide valid range data between approximately 1.5 to 10 meters (Malik et al., 2019) (Hacking et al., 2019) (Luchao et al., 2018) (G. Zhang et al., 2017) (Zhao et al., 2017) and (Marchal & Lygren, 2017). However, the range within 3 to 8 meters are considered valid for human, the value beyond 8 meters are inaccurate, and below 3 meters are unsuitable due to the thermal camera's 18mm focal length and a horizontal field of view of 29.9°. As a result, ranges below 3 meters and above 8 meters need to be excluded. This exclusion is achieved using the FTDI technique to generate the depth of interest (DOI) with the TOI images serving as a template against the depth_r images. This technique can be represented as,

$$Output = A \cdot B$$

$$DOI(x,y) = \begin{cases} 0: & TOI(x,y) \text{ or } depth\_r(x,y) \text{ or both equal to zero} \\ depth\_r(x,y): & \text{both bigger than zero} \end{cases} \quad (2)$$

Where Output stands for *DOI(x, y)* - the thresholder depth image, *A* corresponds to *TOI(x, y)* - the thresholder thermal image, and *B* corresponds to *depth_r(x, y)* - the registered depth. One of the objectives of this work is to detect multiple persons within the 3 to 8-meter range including occluded persons. While the depth of interest has been generated, accurately determining the total of candidates, especially for occluded individuals, remains challenging. The candidate selection process consists of two steps: (i) clustering all objects, with each cluster stored as a single object (split-restore region), and (ii) filtering to eliminate small regions (filter-merge region). To obtain distinct regions, an output region *Oi* is created for each candidate *i*, (3). Ensuring that each output region *Oi* contains only one candidate. Each output is then filtered to remove small adjoining regions, with regions smaller than 500 pixels considered noise and subsequently eliminated (Malik et al., 2019) and (Gonzalez et al., 2017).

$$Oi=pli<pi<phi \quad (3)$$
$$Ri=opening(Oi,A) \quad (4)$$

2.3 Object Classification

This stage involves three phases of detection: (i) Pixel Code Features – This phase converts the pixel values of all regions *Ri* into individual codes. (ii) Pre-detector – It generates the bounding box coordinates (BBC), specifically *x, y, height, and width*, which are utilized in the Occlusion Human Detection (OCHD) stage. (iii) OCHD Detector – This stage executes detection by pixel codes based on the BBC obtained from the pre-detector. Within the first phase, human detection is performed by focusing on structures of upper body, and the BBC are saved. This is accomplished by modifying algorithm based on (Nasir et al., 2019) through system object detectors, with properties set to 30x30 pixels with 1.05 scale factor (minimum size), following the framework of Viola-Jones. The image size is based on the search window and scale factor using the Viola-Jones algorithm, as illustrated in Figure 3.

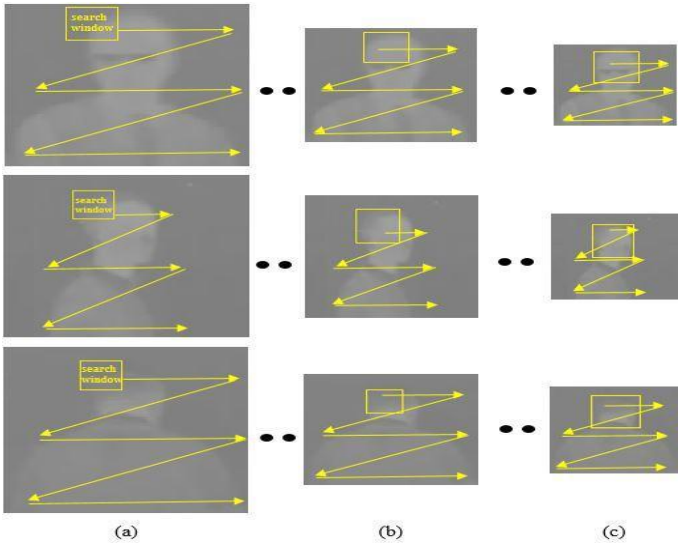


Figure 3. Image size based on the search window and scale factor.
(a) Maximum size; (b) actual size; (c) minimum size

The minimum size property limits the object sizes that can be detected; ideally, these parameters can be adjusted to decrease computation time when the size of the target object is known beforehand. These properties are influenced by the scale factor, which constrains the dimensions of the search windows. This detector employs Histogram of Oriented Gradients (HOG) features as suggested by (Dalal et al., 2006) and uses a cascade of classifiers trained via boosting. The OCHD detector consists of two stages: (i) person prediction and (ii) verification and refinement. During the first stage, a scanning process is used to locate new persons within the BBC region. The scanner checks for new codes along both the y-axis and x-axis of the BBC to identify any discrepancies from the reference codes. This scanning technique is adapted from the Multi-Block Local Binary Pattern (MB-LBP) method introduced by (Cruz et al., 2016), also extends the Local Binary Pattern (LBP) technique described by (Cruz et al., 2016). Unlike LBP, which relies on individual pixels, MB-LBP applies the LBP operator to blocks of pixels, ensuring that all blocks conform to a 3x3 layout (i.e., they must be of uniform size). The authors highlighted that MB-LBP acts as a faster and more effective texture classifier, with texture defined as a function representing spatial variations in pixel intensity within an image. The original expression for MB-LBP labels is given by;

$$LBP(x_p, y_p) = \sum_{p=0}^{p-1} s(i_n - i_p)2^n \quad (5)$$

Where i_n and i_p represents the gray level of the neighboring and central pixels, (x_p, y_p) is the pixel of the image, and n is the number of neighboring pixels. This process is illustrated in Figure 4, expressed as;

$$Pc_n(xi, yi) = \sigma_{0}^{1} \mathbb{X}(\blacksquare\blacksquare\blacksquare) \quad (6)$$

Where Pc_i is the pixel code of i, (x_i, y_i) is the pixel of the image, i denotes the current pixel, and $s(pc_i)$ is described by expression (7).

$$s(pc_i) = \begin{cases} 0, & \blacksquare\blacksquare\blacksquare = 0 \\ 0, & \blacksquare\blacksquare\blacksquare = \blacksquare\blacksquare\blacksquare \\ \blacksquare\blacksquare\blacksquare, & \blacksquare\blacksquare\blacksquare \neq \blacksquare\blacksquare\blacksquare \end{cases} \quad (7)$$

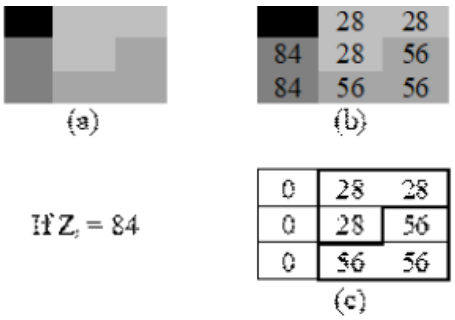


Figure 4. Scanner process. (a) is the 3x3 scanner window; (b) pixel codes; (c) the assignment of 0 if the pixel value is 0 or same to Z_r

When the Pc_n value differ from the value of Z_r, the Pc_n is stored in a new matrix, which will be used along with Z_r for the next iteration as the reference code. A new matrix is created to gather the coordinates, including the values of maximum x and y, the number of rows in *BBC* and Pc_n which will be used in subsequent processes. This matrix, designated as New Persons Coordinates (*NPC*), is represented as;

$$NPC_{m,j} = [xj, yj, x\text{-}maxj, y\text{-}maxj, qj, Pcnj] \quad (8)$$

Where m is the total number of Pc_n and j is the number of rows in NPC. Each row of the output matrix NPC comprises a six-element vector, where x_j and y_j are coordinate for Pc_{nj}, q is the number of rows in the BBC, Pc_{nj} is the code of a new person, and *x-max/y-max* denote the maximum value of x-axis/y-axis for the current *BBC*. Concurrently, the *BBC* matrix is updated by adding the pixel code (Pc), with the new matrix represented as;

$$BPC_{n,i} = [xi, yi, hi, wi, Pci] \quad (9)$$

The next stage is verification, which occurs during the scanning of each BBC. Here, identical pixel codes detected across different BBCs may emerge. The NPC matrix will be compared to the BPC matrix to remove coordinates associated with the same candidate.

If z is equal to 0, it indicates that there are no new individuals present in the bounding box of the *BPC*. This process involves examining each row of the *NPC* matrix one by one, with z representing the current row index of the *NPC*, and comparing it against each bounding box of the *BPC*. Start by initializing d as the number of current *BPC* and obtaining the value q from current *NPC*. If the value of q and d for NPC_z are the same, the process will skip this instance and move to the next d. The condition for comparison is as follows: $x_z \geq x_d.min$ && $x_z \leq x_d.max$ while ensuring $Pc_{nz} = Pc_d$. If these conditions are met, NPC_z is reset to 0, represented NPC_z as;

$$NPC(z, :) = [0, 0, 0, 0, 0, 0] \quad (10)$$

This process focuses on removing any coordinates from the matrix where $NPC_z = 0$, and the remaining coordinates are organized according to the standard image matrix defined in equation (3.35), referred to as *NPCR*. Additionally, the height and width parameters will be set to 30x30 pixels with 1.05 scale factor (minimum size), in accordance with the settings of the Viola-Jones algorithm. The *BBC* and *NPCR* matrices will then be combined into a single matrix. Ultimately, this will yield the final matrix coordinate of the bounding box, which indicates the total number of humans detected by both detectors. The final matrix is named *FBB* and represented *FBB* as;

$$NPCR = [xz, yz, height, width] \quad (11)$$
$$FBB = [BBC; NPCR] \quad (12)$$

Current mobile robots emphasize the necessity for a reliable and fast system in human detection. Yet, techniques from static applications cannot always be directly applied to mobile platforms due to factors like platform movement, increased noise, computational constraints, and environmental uncertainty. Additionally, the challenges intensify when the robot moves, leading to unreliable information cues. Therefore, key issues for surveillance systems using mobile robots include managing unreliable or incomplete information, ensuring real-time responses, and robust object segmentation.

3.0 Results and Discussion

This study presents the results of the proposed human detection system using a Kinect sensor and thermal camera deployed on the mobile robot, utilizing the suggested pre-detectors and occlusion detectors. An experiment demonstrates multi-person detection using the occlusion detector, showcasing its performance. Both detectors are evaluated in the experiment, and potential improvements are discussed. In crowded environments, multiple individuals may overlap, necessitating a robust detector capable of identifying all persons in the images. The previously generated BBC coordinates are used to identify new individuals in the current BBC through pixel codes. The pixel codes in DOI images can pinpoint a person's location. The proposed occlusion detector identifies overlapping individuals by comparing pixel codes and the positions of bounding boxes in each image. Consequently, detecting occluded persons without a depth image may result unsuccessful.

The benchmarking process involves comparing four algorithms as follows, method by (Hacinecipoglu et al., 2020) using a depth camera to focuses on human detection with a two-stage cascaded structure that includes extreme points in the head-shoulder descriptor (HSD) and an edge map. These systems were evaluated using a mobile platform dataset provided by (Munaro & Menegatti, 2014). The study by (Ozcan & Cetin, 2022) presents a real-time human detection method using a depth camera, and a CNN descriptor combined with a physical radius-depth detector. These systems were evaluated using own dataset and COCO dataset. Works by (W. Zhang et al., 2020) addresses human detection with occlusion handling in complex environments, utilizing a depth camera along with three features namely Depth map, Multi-order depth template, and Height difference map (DMH). Authors also assessed their system utilizing a mobile platform dataset supplied by (Choi et al., 2013).

The experiments evaluated detection performance under moderate and strong occlusion conditions using images from the MRV-MP dataset, which contains 500 images—262 with moderate occlusion and 238 with strong occlusion. The datasets have a total of 786 and 714 NGT data points, respectively, with each detected individual counted as a false alarm. Table 1 and Figure 5 present quantitative results, showing that the proposed system performs well, achieving an average precision of 0.934 for moderate occlusion and 0.935 for strong occlusion. It outperforms three other algorithms, attaining accuracy rates of 96.25% for moderate occlusion

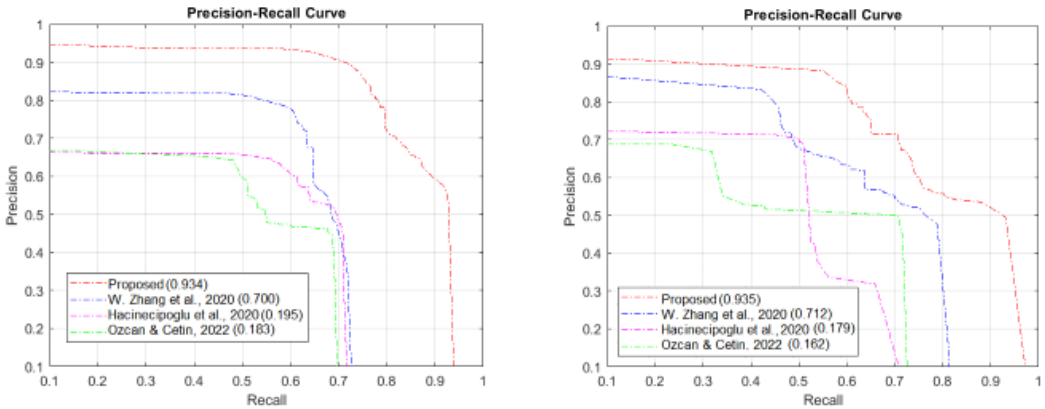
and 96.41% for strong occlusion, compared to the previous systems' accuracy rates of 44% to 75%.

Here are the possible reasons for the performance differences observed in the compared methods:- Method by (W. Zhang et al., 2020), This approach utilizes a two-stage detection process to capture information of geometric structure from the depth map through DMH, along with CHL (Candidate Head-top Locating) using Convolutional Neural Networks (CNN). While this method achieves commendable results in terms of average precision and recall, it struggles with the classification of cases involving moderate and strong occlusion. The primary issue lies in the CHL method focusing exclusively on the positions of simply occluded persons, thereby neglecting candidates in instances of severe occlusion. This limitation hampers its overall efficacy in addressing complex occlusion scenarios.

Table 1. Moderately and strongly occlusion of detection metric results.

Dataset/ Modality	Moderately			Strongly		
	P	R	A	P	R	A
Proposed	93.41%	99.24%	96.25%	93.54%	99.44%	96.41%
W. Zhang et al., 2020	69.95%	73.16%	74.82%	71.06%	73.95%	75.59%
Hacinecipoglu et al., 2020	19.49%	9.67%	45.71%	17.93%	7.28%	45.98%
Ozcan & Cetin, 2022	18.33%	11.20%	44.50%	16.22%	7.70%	44.92%

Method by (Hacinecipoglu et al., 2020), This method employs a 3D head contour and surface model in depth images for detecting upper-body individuals. However, it relies solely on depth information, making the algorithm susceptible to inefficiencies due to significant noise present in the depth data. Moreover, the segmentation of depth information can be challenging, particularly in the presence of various pose variations and frequent moderate to strong occlusions of upper-body persons. As a result, the method yields lower precision and recall rates. Method by (Ozcan & Cetin, 2022), This approach focuses on the 3D geometric properties of the human head, utilizing radius and depth values for effective top-head detection and incorporating a CNN descriptor to assess the validity of human location. Nevertheless, the method called Physical Radius-Depth fails to accurately generate features for head-top human detection. Additionally, employing deep learning object-based detectors tends to yield an excessive number of region proposals, leading to potential confusion and decreased detection accuracy.



(a) (b)

Figure 5. The OCHD detector Performance based on Precision-Recall Curve.
(a) moderately dataset and (b) strongly dataset

The experiments were conducted to assess the computational efficiency of the proposed system compared to three state-of-the-art algorithms, achieving an average processing time of 58.8 milliseconds per frame on the specified dataset. This performance indicates that the

proposed system operates efficiently relative to the three benchmark algorithms, as detailed in Table 2, which presents a comprehensive comparison of the run-time performance.

Table 2. Comprehensive comparison of the run-time performance against algorithms.

Algorithms	Run-time/Frame
Proposed	58.80msec/frame
W. Zhang et al., 2020	60.75msec/frame
Hacinecipoglu et al., 2020	60.90msec/frame
Ozcan & Cetin, 2022	61.52msec/frame

The fusion method utilized by the proposed system proves to be highly effective, yielding accurate Depth of Interest (DOI) results. In contrast, the state-of-the-art algorithms often require multiple filtration stages to produce their DOI outputs, resulting in a significant computational burden. While the proposed system demonstrates commendable efficiency, there remains room for enhancement to achieve real-time detection capabilities. The OCHD detector particularly excels in identifying individuals across varying levels of occlusion, namely low, moderate, and strong occlusions - highlighting the system's robustness in challenging scenarios. Overall, the proposed system not only exhibits competitive computational efficiency but also shows promising improvements in detection accuracy under diverse occlusion conditions. Qualitative results, as illustrated in Figure 6, showcase the advancements in detection performance facilitated by the Pre-detector, left column. The left column (a–c) shows the pre-detector used to obtain BBC coordinates, while the right column (d–f) presents the OCHD detector aimed at improving the human detection rate.

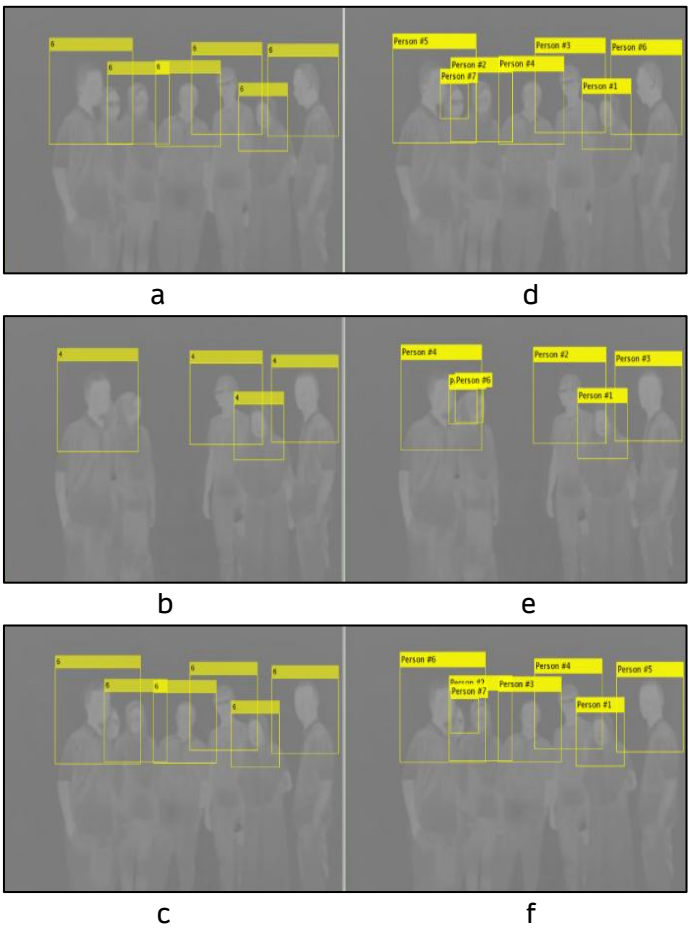


Figure 6. Advancements in detection performance facilitated by the Pre-detector.

4.0 Conclusion

This study investigates several methods aimed at enhancing human detection performance, particularly for occluded individuals, in surveillance systems deployed on mobile robots. The proposed system is based on the infrared spectrum, utilizing data from thermal and depth cameras to improve detection capabilities. A major innovation in this work is the development of a novel hybrid approach that integrates detectors trained on both unoccluded and occluded individuals. This method effectively utilizes the advantages of Fusion Thermal-Depth Information (FTDI), enabling the system to precisely determine the positions of individuals in images by combining thermal and depth data. By addressing the limitations associated with traditional computer vision (CV) sensors, this system represents a significant advancement in detection technology. The system's design takes into account the challenges posed by the movement of mobile robots and the limitations of infrared sensors. The results suggest that detection methods can be further refined to better address occlusion issues and improve human detection accuracy in cluttered environments. However, it is acknowledged that the proposed system is not yet fully optimized for real-world deployment, and additional improvements are needed to enhance detection performance across a broader range of conditions. Despite its current limitations, the system has shown significant progress and yielded promising results. Ongoing development is crucial to enhance visibility in low-light conditions, adverse weather, and environments affected by smoke or other obscurants. One of the notable strengths of the proposed system is its computational efficiency, which ensures effective performance even in poor lighting and crowded scenes, thanks to its reliance on depth and thermal information. Overall, while improvements are still needed, the proposed system lays a solid foundation for advancing human detection technology in dynamic surveillance environments.

Acknowledgements

The authors would like to express their sincere appreciation to the management and lecturers of the participating Politeknik Ibrahim Sultan and Universiti Teknologi Malaysia for their support in facilitating this research and innovation. Special thanks are also extended to the students who actively provided valuable feedback. The authors are grateful to the 6th International Conference on Applied Sciences, Information and Technology (ICo-ASCNITech) 2025 organizing committee for the opportunity to present and publish this work.

Author Contributions

Hashim S. H.: Conceptualization, Statistical Analysis, Data Collection, Writing-Original Draft Preparation, Software, Project Administration; **U U Sheikh**: Supervision, Writing-Review and Editing; **A Hidayat**: Supervision, Writing-Review and Editing; **T I T Nadzion**:Data Curation, Methodology Design, Validation; **W T W M Rumaizi**: Problem Statement, Literature Review, Writing-Review and Editing; **S Saifuddin**: Statistical Analysis, Writing-Review and Editing, Referencing, Validation.

Conflicts of Interest

The manuscript has not been published elsewhere and is not being considered by other journals. All authors have approved the review, agree with its Submission and declare no conflict of interest in the manuscript.

5.0 References

Akram, B. A., Zafar, A., Akbar, A., Wajid, B., & Chaudhry, S. A. (2018). Change Detection Algorithms for Surveillance in Visual IoT: A Comparative Study Visual Internet of Things. *Mehran University Research Journal of Engineering and Technology*, *37*(1), 77–94. https://hal.archives-ouvertes.fr/hal-01676639/.

Angelini, F., Member, S., Fu, Z., Long, Y., & Member, S. (2020). 2D Pose-based Real-time Human Action. *IEEE Transactions on Multimedia*, *22*(6), 1433–1446. https://doi.org/https://doi.org/10.1109/TMM.2019.2944745.

Aslan, M. F., Durdu, A., Sabanci, K., & Mutluer, M. A. (2020). CNN and HOG based comparison study for complete occlusion handling in human tracking. *Measurement: Journal of the International Measurement Confederation*, *158*, 107704. https://doi.org/10.1016/j.measurement.2020.107704.

Bakar;, B. H. A., Kamaruddin;, S., & Mustapha;, M. F. (2022). Hoo Filter with Fuzzy Logic Estimation Based SLAM to Refrain Finite Escape Time: An Analysis in Different Mobile Robot Movement. *International Journal of Technical Vocational and Engineering Technology (IJTvET)*, *3*(2), 1–10.

Cao, C., Wang, Y., Kato, J., Zhang, G., & Mase, K. (2017). Solving Occlusion Problem in Pedestrian Detection by Constructing Discriminative Part Layers. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 91–99. https://doi.org/10.1109/WACV.2017.18.

Chi, C., Zhang, S., Xing, J., Lei, Z., Li, S. Z., & Zou, X. (2020). PedHunter: Occlusion robust pedestrian detector in crowded scenes. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 10639–10646. https://doi.org/10.1609/aaai.v34i07.6690.

Choi, B., Mericli, C., Biswas, J., & Veloso, M. (2013). Fast human detection for indoor mobile robots using depth images. *Robotics and Automation (ICRA), 2013 IEEE International Conference On*, 1108–1113. https://doi.org/10.1109/ICRA.2013.6630711.

Cruz, J., Shiguemori, E., & Guimarães, L. (2016). A comparison of Haar-like, LBP and HOG approaches to concrete and asphalt runway detection in high resolution imagery. *Journal of Computational Interdisciplinary Sciences*, *6*.

Dalal, N., Triggs, B., Schmid, C., Dalal, N., Triggs, B., Schmid, C., Detection, H., & Oriented, U. (2006). Human Detection using Oriented Histograms of Flow and Appearance. *European Conference on Computer Vision (ECCV '06)*, 428–441. https://doi.org/10.1007/11744047.

Gilroy, S., Jones, E., & Glavin, M. (2021). Overcoming Occlusion in the Automotive Environment - A Review. *IEEE Transactions on Intelligent Transportation Systems*, *22*(1), 23–35. https://doi.org/10.1109/TITS.2019.2956813.

Gonzalez, R. C., Woods, R. E., & Eddins, S. L. (2017). Image Processing Toolbox User's guide. In *The MathWorks, Inc.* (Vol. 3).

Guo, Z., Liao, W., Xiao, Y., Veelaert, P., & Philips, W. (2020). Deep learning fusion of RGB and depth images for pedestrian detection. *30th British Machine Vision Conference 2019, BMVC 2019*, 1–13.

Gupta, C., Gill, N. S., Gulia, P., Yadav, S., Pau, G., Alibakhshikenari, M., & Kong, X. (2024). A Real-Time 3-Dimensional Object Detection Based Human Action Recognition Model. *IEEE Open Journal of the Computer Society*, *5*(November 2023), 14–26. https://doi.org/10.1109/OJCS.2023.3334528.

Hacinecipoglu, A., Konukseven, E. I., & Koku, A. B. (2020). Pose invariant people detection in point clouds for mobile robots. *International Journal of Mechanical Engineering and Robotics Research*, *9*(5), 709–715. https://doi.org/10.18178/ijmerr.9.5.709-715.

Hacking, C., Poona, N., Manzan, N., & Poblete-Echeverría, C. (2019). Investigating 2-D and 3-D proximal remote sensing techniques for vineyard yield estimation. *Sensors (Switzerland)*, *19*(17). https://doi.org/10.3390/s19173652.

Liu, T., Luo, W., Ma, L., Huang, J. J., Stathaki, T., & Dai, T. (2021). Coupled network for robust pedestrian detection with gated multi-layer feature extraction and deformable occlusion handling. *IEEE Transactions on Image Processing*, *30*, 754–766. https://doi.org/10.1109/TIP.2020.3038371.

Lou, A., Guan, S., Kamona, N., & Loew, M. (2021). *Segmentation of Infrared Breast Images Using MultiResUnet Neural Networks*. 1–6. https://doi.org/10.1109/aipr47015.2019.9316541.

Luchao, T., Li, M., Hao, Y., Liu, J., Zhang, G., & Chen, Y. Q. (2018). Robust 3D Human Detection in Complex Environments with Depth Camera. *IEEE Transactions on Multimedia*, *20*(9), 2249–2261.

Malik, M. H., Qiu, R., Gao, Y., Zhang, M., Li, H., & Li, M. (2019). Tomato segmentation and localization method based on RGB-D camera. *International Agricultural Engineering Journal*, *28*(4), 278–287.

Marchal, G., & Lygren, T. (2017). *The Microsoft Kinect : validation of a robust and low-cost 3D scanner for biological science*. https://doi.org/10.13140/RG.2.2.12069.40167.

Mary, S. P., Ankayarkanni, Nandini, U., Sathyabama, & Aravindhan, S. (2020). A Survey on Image Segmentation Using Deep Learning. *Journal of Physics: Conference Series*, *1712*(1). https://doi.org/10.1088/1742-6596/1712/1/012016.

Merad, D., Aziz, K. E., Iguernaissi, R., Fertil, B., & Drap, P. (2016). Tracking multiple persons under partial and global occlusions: Application to customers' behavior analysis. *Pattern Recognition Letters*, *81*(September), 11–20. https://doi.org/10.1016/j.patrec.2016.04.011.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2022). Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(7), 3523–3542. https://doi.org/10.1109/TPAMI.2021.3059968.

Mo, X., Sajid, U., & Wang, G. (2021). Stereo Frustums: a Siamese Pipeline for 3D Object Detection. *Journal of Intelligent and Robotic Systems: Theory and Applications*, *101*(1). https://doi.org/10.1007/s10846-020-01287-w.

Muhadi, N. A., Abdullah, A. F., Bejo, S. K., Mahadi, M. R., & Mijic, A. (2020). Image segmentation methods for flood monitoring system. *Water (Switzerland)*, *12*(6), 1–10. https://doi.org/10.3390/w12061825.

Munaro, M., & Menegatti, E. (2014). Fast RGB-D people tracking for service robots. *Autonomous Robots*, *37*(3), 227–242. https://doi.org/10.1007/s10514-014-9385-0.

Murugiah;, N., Mahendra;, V., Purushotaman;, P., Rassiah;, K., & Karudin;, A. (2024). Review on AutoSeed: Smart Seeding Tray Robot with IoT Integration. *International Journal of Technical Vocational and Engineering Technologv (IJTvET)*, *5*(2), 32–38.

Nasir, A. F. A., Ghani, A. S. A., Zakaria, M. A., Majeed, A. P. A., & Ibrahim, A. N. (2019). Automated Face Detection Using Skin Color Segmentation and Viola-Jones Algorithm. *Mekatronika*, *1*(1), 58–63.

Ozcan, A., & Cetin, O. (2022). A Novel Fusion Method With Thermal and RGB-D Sensor Data for Human Detection. *IEEE Access*, *10*(June), 66831–66843. https://doi.org/10.1109/ACCESS.2022.3185402.

Panetta, K., Shreyas Kamath, K. M., Rajeev, S., & Agaian, S. S. (2021). FTNet: Feature Transverse Network for Thermal Image Semantic Segmentation. *IEEE Access*, *9*, 145212–145227. https://doi.org/10.1109/ACCESS.2021.3123066.

Park, Chen, Cho, Kang, & Son. (2020). Detección de personas basada en CNN mediante imágenes infrarrojas para sistemas de advertencia de intrusión nocturna.pdf. *Sensors (Switzerland)*, *20*(1).

Pourmehr, S., Thomas, J., Bruce, J., Wawerla, J., & Vaughan, R. (2017). Robust sensor fusion for finding HRI partners in a crowd. *Proceedings - IEEE International Conference on Robotics and Automation*, *i*, 3272–3278. https://doi.org/10.1109/ICRA.2017.7989373.

Shoelson, B. (2013). *Cascade Trainer: Specify Ground Truth, Train a Detector*. http://www.mathworks.com/matlabcentral/fileexchange/39627-cascade-trainer--specify-ground-truth--train-a-detector.

Stewart, R., & Andriluka, M. (2016). End-to-end people detection in crowded scenes. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, 2325–2333. https://doi.org/10.1109/CVPR.2016.255.

Zhang, G., Liu, J., Li, H., Chen, Y. Q., & Davis, L. S. (2017). Joint Human Detection and Head Pose Estimation via Multistream Networks for RGB-D Videos. *IEEE Signal Processing Letters*, *24*(11), 1666–1670. https://doi.org/10.1109/LSP.2017.2731952.

Zhang, W., Guo, X., Wang, J., Wang, N., & Chen, K. (2021). Asymmetric adaptive fusion in a two-stream network for rgb-d human detection. *Sensors (Switzerland)*, *21*(3), 1–17. https://doi.org/10.3390/s21030916.

Zhang, W., Wang, J., Guo, X., Chen, K., & Wang, N. (2020). Two-Stream RGB-D Human Detection Algorithm Based on RFB Network. *IEEE Access*, *8*, 123175–123181. https://doi.org/10.1109/ACCESS.2020.3007611.

Zhao, J., Zhang, G., Tian, L., & Chen, Y. Q. (2017). Real-time human detection with depth camera via a physical radius-depth detector and a CNN descriptor. *Proceedings - IEEE International Conference on Multimedia and Expo*, *March*, 1536–1541. https://doi.org/10.1109/ICME.2017.8019323.

Zhou, C., Yang, M., & Yuan, J. (2019). Discriminative feature transformation for occluded pedestrian detection. *Proceedings of the IEEE International Conference on Computer Vision*, *1*, 9556–9565. https://doi.org/10.1109/ICCV.2019.00965.

Zhou, C., & Yuan, J. (2019). Multi-label learning of part detectors for occluded pedestrian detection. *Pattern Recognition*, *86*, 99–111. https://doi.org/10.1016/j.patcog.2018.08.018.