Measurement 236 (2024) 115133

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Human height estimation using AI-assisted computer vision for intelligent video surveillance system

K. Iyshwarya Ratthi [a], B. Yogameena [a,*], S. Saravana Perumaal [b]

[a] Department of Electronics and Communication Engineering, Thiagarajar College of Engineering, Madurai 625015, Tamil Nadu, India
[b] Department of Mechanical Engineering, National Institute of Technical Teachers' Training and Research, Chennai 600113, Tamil Nadu, India

ARTICLE INFO

Keywords:
Human height estimation
Computer Vision
Artificial Intelligence
Human height dataset
Video Surveillance
YOLOv7
Camera Calibration

ABSTRACT

In urban areas, technological advancements have led to an increased focus on height as a critical human characteristic for surveillance purposes. Face recognition often encounters challenges due to occlusion and masks, necessitating the use of height, build, and torso. Accurately estimating human height in surveillance scenarios is complex due to camera calibration, posture variations, and movement patterns. This research introduces a novel human height estimation method for surveillance systems, along with a dedicated dataset. The process begins with camera calibration to rectify lens distortions. A deep learning-based YOLOv7-Occlusion Aware (YOLOv7-OA) target detection technique is employed to precisely locate individuals within the frame. The study assesses the impact of camera height and deflection angle on height estimation across different areas of the field of vision (FOV). The proposed method yields a mean absolute error of 0.02 cm to 0.8 cm across various FOV zones, surpassing the previous 1.39 cm benchmark findings.

1. Introduction

The pursuit of locating missing children [1] is an urgent and need-of-the-hour issue that demands immediate attention and relentless effort. Every year, countless children go missing, and this number has shown a sharp increase in the last two years [2]. Usage of CCTVs has grown rapidly in recent years, primarily attributed due to technological advancements, lower equipment costs, and greater public awareness of the advantages offered by video surveillance. Consequently, law enforcement entities employ CCTV systems to monitor communal spaces, discourage illicit activities, and safeguard assets. Authorities have strategically implemented CCTV systems to effectively monitor and discern potential security threats across various public domains [3]. The captured video footage holds immense significance and crucial evidence in investigative procedures, as it encapsulates invaluable data pertaining to the observed environment. In locating missing children, it is imperative to acknowledge that each fragment of information holds a potential impact to detect them. Computer vision has made significant strides in recent years, with its ability to derive valuable insights from visual such as images and videos. This progress has rendered computer vision an indispensable tool in the realm of person retrieval, enabling the identification and retrieval of individuals with remarkable efficacy. Computer vision algorithms and deep learning methodologies, ease law enforcement agencies possess the potential to augment their search endeavors and notably increase the likelihood of successfully locating missing children.

The traditional means of person detection predominantly rely upon the utilization of physical attributes, such as facial features and distinctive marks. Nevertheless, the precise extraction of the child's facial characteristics is hindered by inadequate image resolution or the child's pose, specifically when their back is oriented towards the camera. Additionally, in the aftermath of the COVID-19 pandemic, it is observable that individuals are donning facial masks. In such a scenario, the pursuit of alternative attributes [4,5] is undertaken to ascertain the missing child's whereabouts. When employed collectively in the process of analysis, individuals' physical, chemical, and biological attributes serve to differentiate one person from another through their unique biometric traits. It is customary to gather pertinent details encompassing their physical appearance to locate a missing child. These details typically include facial features, attire color/type, hairstyle, height [6–9] and complexion. Unlike other soft biometrics such as facial features, clothing, hairstyle, or skin tone, which may not possess adequate authentication capabilities as it can be easily masked, height has been identified as a possible credential [10–15]. The height of an individual is

* Corresponding author at: Department of Electronics and Communication Engineering, Thiagarajar College of Engineering, Madurai 625015, Tamil Nadu, India.
E-mail address: ymece@tce.edu (B. Yogameena).

https://doi.org/10.1016/j.measurement.2024.115133
Received 4 November 2023; Received in revised form 15 May 2024; Accepted 15 June 2024
Available online 21 June 2024

an important soft biometric trait for enhancing surveillance systems as it is view and distant invariant. Since height is a non-intrusive characteristic, it can be passively acquired and used for identification in circumstances in which facial recognition may not be successful due to poor image quality or occlusion. Person retrieval in video surveillance is enhanced by the consistency of height over time and various angles. In combination with other characteristics such as gender and clothing color, it enables individuals to be distinguished from one another. With the integration of height into deep learning frameworks, surveillance systems can perform robust and efficient re-identification without requiring active participation from individuals. Table 1 shows a comparative analysis of the existing human height estimation strategies. Noteworthy point to be highlighted here is that the existing works have only dealt with adults and not with children.

Height estimation is the process by which the vertical distance between an individual's head and foot is determined. The attribute of height, which can be readily perceived, offers a distinct advantage in searching for missing children. The utilization of height as a soft

Table 1
Comparative analysis of classical and deep learning based human height estimation methodologies.

Algorithm	Dataset	Pros	Limitation	Estimated error (cm)	Surveillance scenario	Year of Publication [Ref]	Height estimated [Adult/ Children]
Classical methods							
Non-linear regression.	15 pairs of points collected with ruler-based method.	Simple, direct estimation without vanishing points.	Highly sensitive to noise	1.4	✓	2015 [36]	Adult
Earth's gravity is used as a reference object along with motion cues.	29 articulated-free-fall images of 12 subjects.	They introduce a more novel field exploiting gravity.	They require a person's motion trajectory, i.e. measured under free-fall.	3.9	✓	2019 [7]	Adult
Histogram of Oriented Gradients and Haar cascade classifiers in Single view metrology.	In house dataset.	Uses environment based cues for calibration.	Head detection highly influences height estimation.	3	✓	2020 [10]	Adult
Photo-modeler software with 3D point data model obtained using laser scan	In house dataset with 16 individuals	Analyzed the effects of camera-object distance and resolution on height estimation.	LiDAR mapping is not cost effective for surveillance systems.	0.916	✓	2020 [11]	Adult
Height estimation using a 3D model generated from a single image.	IMDB-23 K	Extends the SMPL model with more accurate face modeling.	Absolute scale estimation using Inter-Pupillary Distance (as reference).	6	x	2021 [12]	Adult
Corrective image analysis with Amped Five software	Five video footages filtered to 18 frames.	High radial distortion error is corrected.	Tangential distortion is not considered.	7.6	✓	2021[37]	Adult
Semantic Description	Controlled environments	High precision in controlled settings	Limited to specific, non-variable scenes	Not specified	✓	2021[18]	Adults
2D key point estimation using 3D model fitting-SMPLX.	In house dataset	Scale indistinctness when the available reference object is absent.	3D model fitting cannot be used for low resolution surveillance images.	7.4	✓	2022 [13]	Adult
Deep learning methods							
ResNet: A deep learning algorithm is combined anthropometric measurements.	2,74,964 images of 14 individuals and IMDB-100 K	Addresses the problem when camera geometry or scene parameters are unknown	Order of magnitude is limited to a small range.	5.23	x	2018 [6]	Adult
Fully Convolutional Network with proportional relationship between human body-parts.	2136 RGB-D images with ten postures	Body parts segmentation increases accuracy and works well under different postures.	Requires a depth image captured from Kinect.	1.9	x	2020 [8]	Adult
Mask RCNN used to obtain human body information with RGB images and height estimation with depth image.	In house dataset.	Consecutive depth frames reduce the detection error in low-light environment.	Intel RealSense D435 is not cost effective for surveillance.	4.6	✓	2020 [10]	Adult
MOHE-Net that operates as object's bounding box detector (OD-Net) and HE-Net with multi-layer perceptron.	Images from vehicle mounted camera.	Works for both static and moving camera that requires only a single monocular image.	The camera is mounted on a vehicle.	5.08	✓	2021 [14]	Adult
Camera calibration attained by using priors learnt by Neural Networks.	IMDB-23 K	3D scene's scale estimation in unconstrained environment	It performs less accurate than single view metrology with Trivial object insertion	0.9	x	2021 [9]	Adult
A deep learning-based technique with no constraints on camera position.	Dataset generated using CARLA simulator	Height map of the object is determined regardless of the sensor's viewpoint.	Limitations over pedestrian surveillance scenarios.	4.19	x	2023 [15]	Adult

biometric attribute in the search process facilitates enhanced precision in the filtration of potential matches from databases and surveillance footage. By employing a height-based filtering mechanism, investigators possess the capability to effectively reduce the candidate pool to individuals whose stature closely aligns with that of the missing child. This strategic approach serves the purpose of mitigating the occurrence of false positives.

Henceforth, it is imperative to acknowledge that height can be one of several biometric identifiers used to validate a missing child's identity. In recent years, these factors have proven to be advantageous across diverse domains, including but not limited to police intelligence, person re-identification, medical treatment, three-dimensional forensic reconstruction, autonomous driving, and query-driven criminal or missing person retrieval.

Estimating human height from surveillance footage can be challenging for several reasons, where the camera's position and angle can introduce perspective distortion [10] in the footage, making it challenging to estimate the individual's height accurately. Also, it is difficult to determine an individual's height in footage when others obscure it. The contours and edges of a person's body can be difficult to distinguish when shadows or reflections result from various illumination conditions. One of the key advantages of 3D laser scanning [11] is its ability to precisely capture the conditions at a crime scene, preserving spatial relationships and distances, which is not possible with surveillance cameras. It highlights the importance of considering camera resolution and distance in forensic investigations when using PhotoModeler, a photogrammetry software tool. One of the key advantages of 3D laser scanning [11] is its ability to precisely capture the conditions at a crime scene, preserving spatial relationships and distances, which is not possible with surveillance cameras. It highlights the importance of considering camera resolution and distance in forensic investigations when using PhotoModeler, a photogrammetry software tool. This study highlights the need for careful consideration of camera setup in achieving reliable height estimation and other forensic measurements.

People wear different clothing and accessories that can affect their overall appearance and make it difficult to estimate their height accurately. Finally, human variability, where humans come in different shapes and sizes, makes it challenging to develop a one-size-fits-all height estimation algorithm. Various techniques can be employed to estimate height, encompassing camera-based [11–13], sensor-based [14], and manual approaches [7]. However, effectively tackling these challenges necessitates the utilization of computer vision algorithms to effectively manage perspective distortion, occlusion, fluctuating lighting conditions, and additional factors that exert an influence on the accuracy of height estimation. Obtaining precise calibration data for the camera and using better resolution footage helps the algorithm achieve more accurate height estimation. The widespread use of CCTVs, an automated video cognitive service- for human height estimation is put forth, specifically designed to assist in missing child recovery.

Thus, in the relentless pursuit of reuniting missing children with their families, height as a soft biometric trait offers a promising avenue for enhancing search and rescue operations. By leveraging this readily observable attribute, investigators can narrow down potential matches, accelerate the analysis of surveillance footage, and aid in identifying crucial witnesses. As biometric capabilities continue to advance, height emerges as a valuable tool within collaborative endeavors to reunite missing children with their families and foster a safer environment for them. Hence, this research introduces a novel human height estimation method for surveillance systems based on a mathematical model combined with a monocular surveillance camera, which eliminates the need for depth information, along with a dedicated dataset. A human height surveillance dataset containing camera characteristics such as tilt angle, camera height, and ground truth height of the subjects required for height estimate that matches the complexity of the real-world environment is introduced. This is the first dataset to include a wide range of collections of surveillance video with height information, covering both children and adults.

Further, a deep learning-based YOLOv7-Occlusion Aware (YOLOv7-OA) target detection model is employed to precisely locate individuals within the frame. To assist the model in focusing on relevant regions and contextual information, a hybrid attention mechanism (HAM) is introduced into the feature extraction section of the original YOLOv7. The proposed method divides the surveillance field of view into five zones—Monitoring, Detection, Observation, Recognition, and Identification—to enhance detection efficiency and maintain precise height measurement.

1.1. Challenges

Due to the intricacy of the task and the multiple elements that determine an accurate assessment, human height estimate using computer vision presents significant obstacles. Among the major challenges are:

- Accuracy and diversity: Because of the diversity in human posture, clothes, camera angles, and other environmental factors, estimating height effectively merely based on visual clues can be difficult. While there are approaches for estimating height from images or video, their success is dependent on the input data and the complexity of the scene.
- Data Availability: A substantial database of individuals with accompanying height information would be necessary to construct an effective height-based person retrieval system. Such a database may not be readily available or may be limited in terms of diversity, making proper model training problematic.
- Privacy and Ethical Considerations: Because it includes capturing and evaluating personal physical attributes, using height as a soft biometric for person retrieval presents privacy problems. Striking a balance between privacy and the need to find missing people is a vital issue that must be addressed.

In summary, this paper contributes:

- A novel human height estimation method based on a mathematical model combined with a monocular surveillance camera is proposed, which eliminates the need for depth information.
- A human height surveillance dataset containing camera characteristics such as tilt angle, camera height, and ground truth height of the subjects required for height estimate that matches the complexity of the real-world environment is introduced. This is the first dataset to include a wide range of collections of surveillance video with height information, covering both children and adults.
- Recognizing the importance of occlusion management, the YOLOv7-Occlusion Aware is developed as an upgraded version of YOLOv7 designed specifically to address occlusion difficulties. To assist the model in focusing on relevant regions and contextual information, a hybrid attention mechanism (HAM) is introduced into the feature extraction section of the original YOLOv7.
- A detailed experimental examination of various height estimate relevant parameters such as camera height, tilt angle, horizontal and vertical fields, age, and gender is performed.
- To ensure target detection across the full FOV, CCTV's FOV is separated into five zones termed Monitoring (M), Detection (D), Observation (O), and Recognition (R), Identification (I). Zhang's calibration is used to remove any distortion in the video frames so that YOLOv7-OA can recognize and localize pedestrians in complex backgrounds.
- A comparison of several human height estimation methodologies and YOLO-based pedestrian detection algorithms is provided, including YOLOv5, YOLOv6, YOLOv7, and YOLOv8.

2. Related work

Image-Based human height estimation uses 2D images from calibrated or uncalibrated cameras to estimate a person's height. Camera calibration-based approaches estimate camera parameters to accurately relate image measurements to real-world dimensions. It is achieved using various supervised and unsupervised techniques [16] vanishing points, rectangular markers, and facial proportions. Other methods use a linear perspective principle [17] and discuss the effects of calibration that requires of a significant number of evenly distributed Ground Control Points (GCPs), which may not always be feasible in a dynamic crime scene environment. Tools like Amped FIVE and HALCON Library may have inherent limitations in how well they can handle highly distorted images or complex scenarios. These limitations could affect the accuracy and reliability of the forensic analysis. Where, age-wise height measurement is combined with photogrammetric anthropometry. They use visual geometry and feature learning with ground plane approximation analyzed using Amped Five software. These methods necessitate subjects to stand upright, limiting applicability to various poses. Besides, distance between the subject and the sensor were kept constant, which showed that they had not experimented with varying object distance. Various applications [18] of height estimation are forensics, smart surveillance systems, remote sensing, and estimation of human stature from the human face.

Uncalibrated camera-based height estimation [19] directly infers height from images without explicit camera calibration. This relies on geometry, visual cues, and statistical models. They employ Bayesian-like linear model, Histogram of Oriented Gradients and Haar features. These methods employ static uncalibrated cameras, rely on manual key points, and are less accurate than calibrated approaches. Detection discrepancies and rigidity limitations impact accuracy. They work for stationary subjects, not dynamic environments like surveillance with continuous movement. Thus, these methods are constrained in surveillance settings.

In video metrology [20,21], single-view approaches involve identifying and tracking features across multiple video frames. They proposed various methods utilizing different techniques like confidence intervals, likelihood ratios, and gravity as a reference object. These methods provided accurate 3D reconstruction, but manual supervision was often required, and lens distortion was not always considered. The existing methods have leveraged height for various applications, but they have not specifically utilized height as a differentiating factor between adults and children.

Multi-view video metrology [22,23] approaches utilized multiple images from various cameras to estimate height, combining vanishing points, gait and color features, motion information, and linear regression. However, the availability of a reference object and vanishing point were limitations in some cases. They even worked with forehead detection using Hair color histogram combined with person position estimation using LRF tracking. Since, they worked with Asian people; the histogram is taken only for people with black and gray hair color. Thus, Illumination environments, hair color restrictions, and manual annotation were also factors affecting the performance of these methods. Despite this, no person retrieval or missing child retrieval task has been addressed with height as a soft-biometric query.

3D modeling [24] could estimate human height by capturing 3D data of the person's body using 3D scanning and creating a 3D model of their body. Height is estimated by building a 3D model using laser scanning, photo-modeler software with a 3D point. They analyzed the effects of camera-object distance and resolution on height estimation. Furthermore, they have used 2D key point estimation by adopting 3D model fitting-SMPLX. Scale ambiguity existed in the absence of a known reference object. 3D modeling surpassed traditional methods such as anthropometry and photogrammetry in terms of accuracy and efficiency for estimating human height. However, implementing 3D modeling required specialized equipment and software like laser scanning and LiDAR mapping, resulting in high costs and significant time and effort for setup and maintenance. In recent years, humans' heights have been estimated by deep learning algorithms [25–27] trained on images with corresponding height measurements. They use CNN based architectures like ResNet, Mask-RCNN, FCN combined with anthropometric measures, and trivial object insertion using priors learned by Neural Networks.

As of now, this method does not address the deformation of objects when they are in motion, the weather, illumination, undesirable objects, or distortion effects caused by cameras. They use background subtraction for target detection which incurs error as few foreground pixel regions of the pedestrian are detected as background due to the presence of shadow. Also, at times when the background is more complex it is detected as foreground region. Since height estimation solely depends on the primary person detection algorithm, even minor changes in target positional coordinates can proportionally affect the height estimation. They have experimented with optical axis in straight view where the tilt angle is taken as zero degree and the measurement between the ground level and the camera is not mentioned explicitly. In real-world scenarios the surveillance cameras are mounted at a height of 10–15 feet and the tilt angle is set accordingly to capture the optimum area. The camera's FOV is critical in adequately recording individuals and their height. The target distance influences the accuracy of height estimate. As the object's distance increases, the person appears to be reducing in size, making it more challenging to estimate their height accurately. The field-of-view vs. object distance vs. human height is not addressed in the above methods.

These findings have been limited or biased based on the sample size or demographics of the study participants. The depth images could only be obtained with special stereo vision cameras like Intel RealSense D435, which were not readily available in surveillance scenarios. Additionally, they employed an increasingly complex neural network, which could complicate integration with edge devices. Hence, the proposed work focuses on using traditional surveillance cameras, rather than relying solely on depth-based cameras. Using the right dataset for a height estimate is essential in determining the efficacy of various approaches. While existing methods have utilized datasets like IMDb, which primarily provide information about actors and actresses, they do not typically focus on surveillance videos. However, it is worth noting that person retrieval incorporating height as a soft biometric trait has not yet been implemented according to available knowledge. Moreover, with query height, retrieving a person or child is not addressed yet. However, it's worth noting that while height serves as a soft biometric for person retrieval, it is just one of many factors that need to be considered. Using height alone as a query may not be sufficient for accurate and reliable person retrieval. When facial features are indistinct or obscured, the measurement of height becomes crucial in verifying the identity of a person or a child in conjunction with other soft biometrics. Additional soft biometrics like facial features, clothing, and contextual information such as time and location would enhance the effectiveness of the retrieval system. This highlights the potential for further research and development in utilizing height as a valuable characteristic in the identification and retrieval of individuals, particularly in the context of surveillance videos and missing child cases. An experimental analysis of the effects of height estimation in children, pose-variations, camera's tilt angle, low-resolution images, and gender are yet to be addressed.

3. Research materials and designs

3.1. Data collection and protocol

Existing image data collection for height estimation include W8-400 [28], RGBD-T [29,30] and [8], based on depth image. But these are not surveillance datasets, and in most cases, the height information is not available and is not accessible. Additionally, it's worth noting that these existing datasets for height estimation do not include data related to

children, further limiting their applicability for surveillance and height estimation tasks involving minors. Moreover, the subjects are either standing in the center of the frame in an upright pose or a static object is used instead of human subjects.

Thus, to evaluate a human height estimator that reflects the complexity of the real-world environment, it is not possible to use existing datasets with simple factors. Fig. 1 shows a few samples of the proposed dataset. Consequently, the initial goal was to develop an exclusive in-house dataset with height information. Hence, to evaluate the proposed "Human Height Estimator," a real-world dataset called "Sense-Height" dataset is introduced.

The dataset contains approximately 4,800 images in total where each frame is detailed with person count, tag, and the ground truth of the person. The Hikvision DS-2CD3T56G2-ISU/SL camera and the Dahua DH-IPC-HFW2531S-S-S2 camera were used to record the surveillance footage. The camera settings are tabulated in Table 2. This dataset is the first to encompass a broad spectrum of collections of surveillance footage with height information, including both children and adults. It addresses real-time environmental challenges, crucial for developing a robust human height estimation strategy. It includes images with various poses like standing upright, sitting, and bending; partial occlusions; different ground levels (flat and over a platform); and subjects across multiple motion conditions like walking, bending, and running. Also, the images are taken at various angles, depicting the subjects in different camera views. The in-house dataset contains the statistical measurements of the camera setting, namely, the tilt angle (θ), the focal length (f), the camera height (h_c), the ground truth height (h_{gt}) of the subject and the object distance (d). Thirty-five volunteers, selected to represent a diverse age group, were included in the study. This group comprised individuals ranging from 5 to 35 years old, with a subset of participants aged 5 to 13 representing children (17 participants), and the remaining participants aged 13 to 35 representing adults (18 participants). The average distribution of children's height is between 85 cm and 160 cm, whereas adults are between 160 cm and 180 cm. Among the 35 participants, there were 15 male and 20 female individuals, elucidating the gender distribution within the study. Based on the zoning idea (identification, recognition, and detection) described in Section 2.3, the subjects were asked to move randomly in the FOV. We aim to balance precision with feasibility to ensure reliable and meaningful results across different FOV zones. The samples were specifically collected with respect to an individual moving across different zones of the field of view. Therefore, for each individual involved in the test, around 40 samples were generated in total.

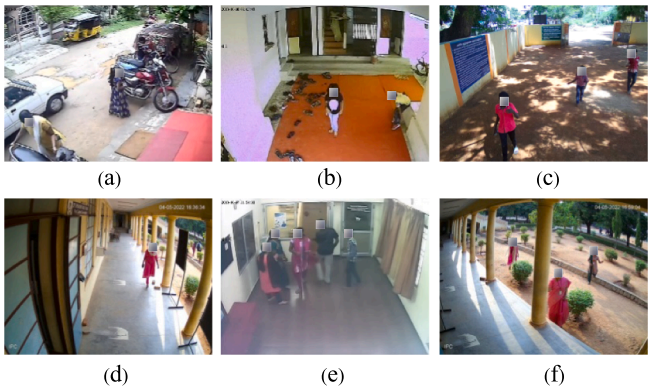

Fig. 1. In-house 'Height-Sense' surveillance dataset. A diverse array of visual data capturing individuals across different age groups, namely children and adults, gender, different camera height and object distance, was systematically acquired from a multitude of publicly accessible locations. (a) Outdoor (b) Indoor (c) Illumination effects (d) Varying object distance (up to 5 m) (e) low resolution (f) Different ground levels (3 levels of ground plane).

Table 2
The camera settings used for "Sense height" dataset creation.

Camera Parameters	Model DH-IPC-HFW2531S-S-S2
Sensor size	1/2.7″ (5.33 mm x 4.00 mm)
Effective Pixels	2592 (H) × 1944 (V)
Camera height (C_h)	12 feet
Focal length (f)	2.8 mm
Object distance (DORI)	64 m, 25 m, 12 m, 6 *m*
HFOV, VFOV, DFOV	97°, 71°, 129°
Pan, Tilt angle, Rotate	0°,15-30°,0°
Resolution	2592 × 1944

3.1.1. Ethical considerations

Soft biometric traits like facial biometrics can reveal a wide range of personal information, including gender, age, ethnicity, and emotional states. In contrast, height is a more straightforward measurement that does not inherently disclose as much potentially sensitive information. This reduces the risk of privacy breaches or unintended disclosure of personal attributes. Starting with comprehensive explanations of the study protocols and the diligent collection of informed consent from the participants (For the consent form, we are ready to submit when required.) the study was carried out with a strong commitment in adhering to strict ethical norms. All study protocols were clearly explained to participants, who also had the option of discontinuing at any time. Steps were taken to ensure participant comfort, providing scheduled breaks and refreshments. Prior to commencing the study, all necessary approvals were obtained from relevant authorities, such as the institutional review board with approval number [2023/TCE/ECE/RD/1], academic faculty, and other relevant departments. Unwavering adherence to these stringent ethical protocols ensured the establishment of a secure and trustworthy research environment. To protect individual identity in datasets, anonymizing faces is a common approach. Here to effectively hide faces in the proposed dataset, a blur effect is applied to the face region that ensures facial features are no longer discernible, thus preserving anonymity.

3.2. YOLOv7-OA detector

Fig. 2 illustrates the architecture of YOLOv7-OA. In recent times, significant progress in CNNs partakes about substantial enhancements in the precision and effectiveness of object detection. This progress has paved the way for a multitude of practical applications in the real world. As noted in references [31], object detection models with a deep learning foundation can be broadly divided into two types: (1) two distinct networks for detection and classification, and (2) one network for both functions. Among them, the first class of detectors is renowned for its great accuracy, with one of the most prominent instances being Faster R-CNN. However, these detectors and region proposal-based frameworks involve multiple correlated stages, resulting in increased network complexity, making them challenging to operate in real-time environments. The availability of vast amount of image data and high-speed GPUs has facilitated the training of one-stage CNN-based algorithms [32] which demonstrate significantly faster detection speeds compared to two-stage detectors on various benchmark datasets. The YOLO (You Only Look Once) algorithms revolutionized as an efficient as well as an accurate object detector in a single forward pass. Amongst them YOLOv7 [33] is the more reliable for deployment in real-time systems due to its architectural reform and the trainable Bag of Freebies, resulting in efficient inference speed and ability to handle larger datasets. YOLOv7 achieves an impressive mAP of 56.8 percent, surpassing the performance both of transformer- and convolution-based object detectors.

This superiority is attributed with the accuracy of bounding box predictions compared to other models in the same category.

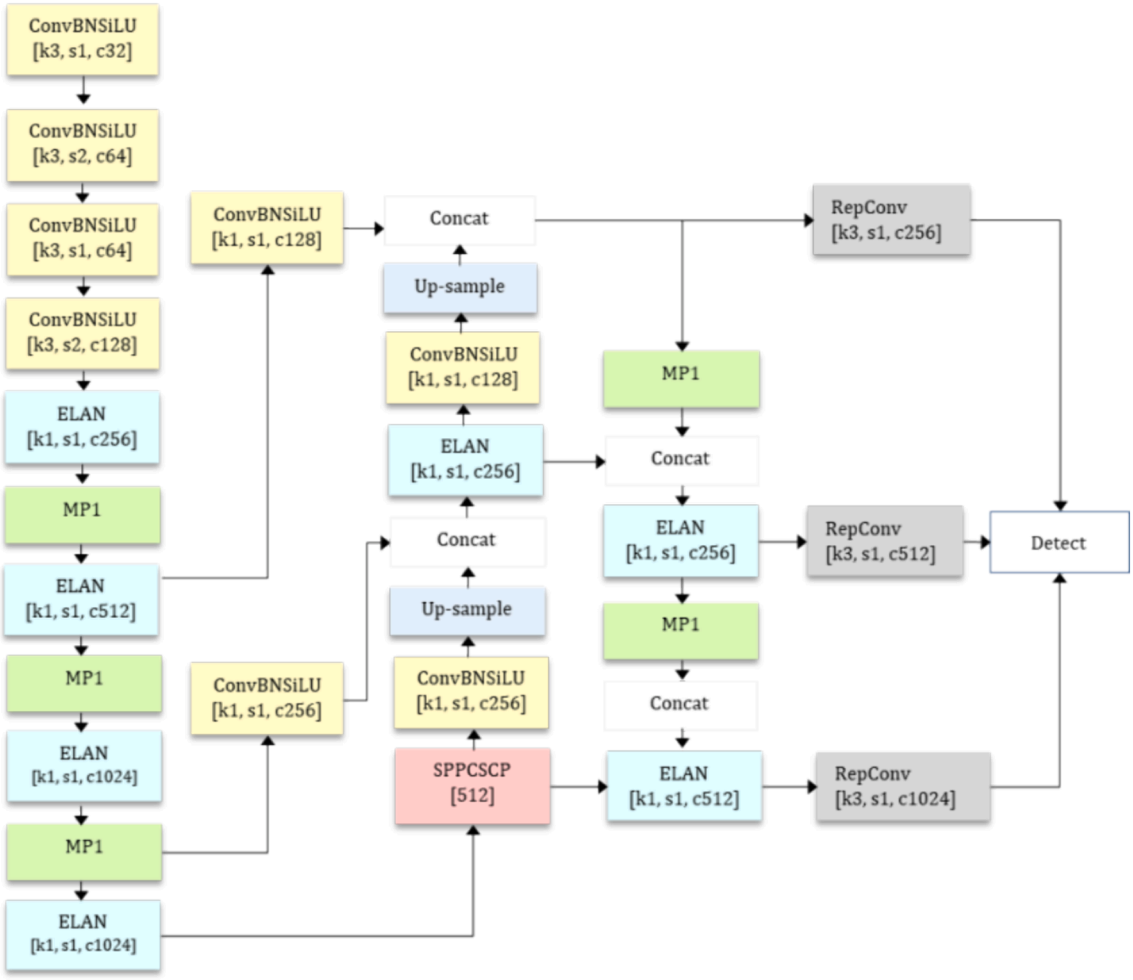

Fig. 2. Illustration of the proposed YOLOv7-OA architecture.

Consequently, we leverage the capabilities of the improved YOLOv7 object detector to identify and localize the target labeled as 'Person,' along with their associated class probabilities and bounding-box coordinates.

3.3. Target distance estimation

Fig. 3 depicts the human perspective of three-dimensional scenes. The class 'Person' is referred to as the target in this case. Using computer vision to estimate target distance entails estimating the distance

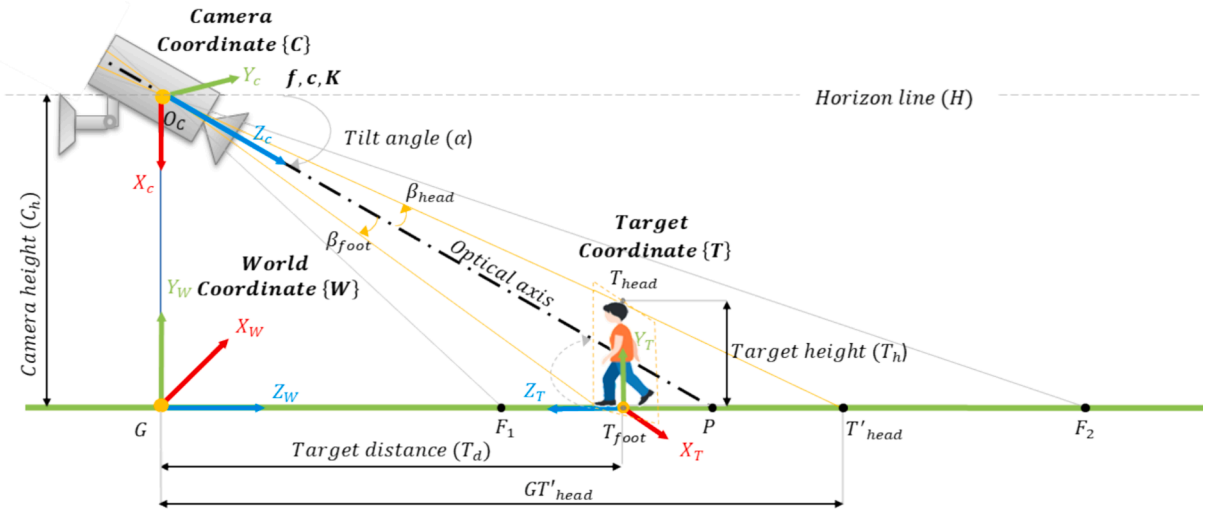

Fig. 3. Two-dimensional side view of the target distance and height estimation environment depicted following the Right-handed coordinate system to represent various coordinate frames.

between the camera and a specified target in the scene based on visual information captured by the camera. Various coordinate frames are represented using the right-handed coordinate system. Several specifications and considerations are required for proper distance estimation:

- **Contact with Ground Plane:** In the physical realm, it is observed that a target encounters the ground plane at least once in the physical realm. Furthermore, the lowest point of the target in an image matches the surface on the ground in the real world.
- **Calibration Data and distortion correction:** Relying on accurate calibration data and a comprehensive transformation between the camera's coordinate system and the real-world coordinate system is fundamental. The focal length (f), physical dimensions of the sensor ($sensorheight, S_h$) and image resolution ($imagewidthx_{max}$, $imageheighty_{max}$) can be obtained from the intrinsic parameters of the camera that are obtained from the camera's datasheet and further verified with camera calibration. These are assumed to be known parameters. Further, the extrinsic parameters such as the tilt angle ($\propto$) and the real-world coordinate of the camera relative to the horizontal plane ($cameraheight, C_h$) are also known.
- **Bounding Box Localization:** The head and foot points within image coordinates are localized using the bounding box detected by YOLOv7-OA. It is denoted as $(x_{head}, y_{head})and(x_{foot}, y_{foot})$.

Finding the head and foot points of a person is found using bounding box coordinates. Bounding boxes are rectangles that are often used to define the location of objects (in this case, a person) in an image. They defined by two points, the top-left corner and the bottom-right corner denoted as $(bx_{min}, by_{min}, bx_{max}, by_{max})$ respectively. The head of the target person is generally located at the top of the bounding box. The coordinates of the head point is estimated using the midpoint of the top side of the bounding box as:

$$(x_{head}, y_{head}) = \left(\frac{bx_{min} + bx_{max}}{2}, by_{min}\right) \tag{1}$$

The foot point is generally located at the bottom of the bounding box. Similar to the head point, the midpoint of the bottom side of the bounding box is used as

$$\left(x_{foot}, y_{foot}\right) = \left(\frac{bx_{min} + bx_{max}}{2}, by_{max}\right) \tag{2}$$

According to the perspective projection principle, the ratio between focal length and sensor size corresponds to the ratio between target distance and target height.

$$\beta_{foot} = \arctan\left(\left(\frac{S_h}{f}\right) \times \left(\frac{y_{foot}}{y_{max}} - \frac{1}{2}\right)\right)$$

$$\beta_{head} = \arctan\left(\left(\frac{S_h}{f}\right) \times \left(\frac{y_{head}}{y_{max}} - \frac{1}{2}\right)\right) \tag{3}$$

where S_h refers to the height of the sensor in the camera measured in mm, f is the Focal length of the camera measured in mm, y_{head}, y_{foot} is the head and foot coordinates of the target in the y-axis of the image plane measured in pixels represented as T_{head} and T_{foot}. y_{max} represents the total number of pixels of the in the y-axis of the image plane.

Let C_h denote the height of the camera on the vertical y-axis measured from the ground plane (G) and point O is the camera's central axis. Let T_d be the distance between the camera and the target (T) on the ground plane. The camera is tilted by an angle (α) from the Horizon line (H), which is the tilt angle of the camera.

In the Fig. 3, the camera's optical center, denoted by O_c, aligns with the ground plane at point G, the base of the target at T_{foot} and a point on the horizon line, labeled H. These points form the corners of a rectangle in the plane of the image. This is deduced from the fact that the camera, when in level, has its optical axis perpendicular to the ground, forming right angles where it intersects the ground plane. Consequently, the line from G to T_{foot} is parallel to the horizon line, and the line from O_c to H is perpendicular to both, fulfilling the criteria for a rectangle. In such a configuration, opposite angles are congruent meaning they are equal in measure. Therefore, in this rectangle, angle $GT_{foot}O_c$ is equal to angle $T_{foot}O_cH$, each being right angles. This geometric principle allows us to infer the congruence of these angles from the layout denoting the angle as θ_1 as in eq. (4). Line OP is the optical axis of the surveillance camera. This angle is equal to the tilt angle α formed between the horizon and optical axis of the camera

$$\theta_1 = \alpha + \beta_{foot} \tag{4}$$

where, β_{head} and β_{foot} forms the angles between the horizontal plane through the camera's optical center and the line of sight to the top of the target and to the bottom of the target.

$$\tan(\theta_1) = \frac{O_cG}{GT_{foot}} = \frac{C_h}{T_d} \tag{5}$$

Eq. (5) relates the angles of a right triangle O_cGT_{foot} to the lengths of its sides. It defines the tangent of the angle θ_1 as the ratio of the length of the side GO_c opposite to θ_1 to the length of the side GT_{foot} adjacent to θ_1.

Therefore, the target distance T_d as given in eq. (6) typically refers to the distance from the camera to the target's location on the ground plane, specifically at the target's feet (T_{foot}). It's the horizontal distance between the camera's position and the base of the target perpendicular to the line of sight.

$$T_d = \frac{O_cG}{\tan(\theta_1)} = \frac{C_h}{\tan(\theta_1)} \tag{6}$$

where, C_h denotes the height of the camera on the vertical y-axis measured from the ground plane (G) and point O be the camera's central axis.

3.4. Target height estimation

The goal is to estimate the target's height (T_h) between the head and foot points (T_{head} and T_{foot}) obtained from the YOLOv7-OA bounding box coordinates. The target height (T_h) is defined as the vertical distance between the ground plane (G) and the highest point of the target, which is typically the top of the head (T_{head}) in a standing human subject. To find T_h, the points representing the target's head and feet in the image coordinate have to be found. In Fig. 3, the points are represented as T_{head} and T_{foot}. Let C_h denote the height of the camera on the vertical y-axis measured from the ground plane (G) and point O be the camera's central axis. Let T_d be the distance between the camera and the target (T) on the ground plane. The camera is tilted by an angle (α) from the Horizon line (H), which is the tilt angle of the camera. Line F_1F_2 represents the Field of View (FOV) of the camera. The surveillance camera was mounted at a height (C_h) of 10 m and a tilt angle of 36° to meet the real-world requirement. The camera specifications used in experimentation are detailed in Table 2. Line OP is the optical axis of the surveillance camera. Since it is a 2D view; the target person is shown as a line with projection angles β_{head} and β_{foot} between the optical axis and the head and foot point of the target person. The head point T_{head} is projected to the ground plane at point T'_{head}. According to the principle of perspective, the head point (T_{head}) projects onto the x-axis of the ground plane as point T'_{head}. As a result, the angle formed by the optical axis and the projection of the head point onto the ground plane is given by (7). β_{head} is determined in the counterclockwise from the optical axis.

$$\theta_2 = \alpha - (-\beta_{head}) = \alpha + \beta_{head} \tag{7}$$

$$\tan(\theta_2) = \frac{O_cG}{GT'_{head}} = \frac{C_h}{GT'_{head}} \tag{8}$$

Triangles OGT'$_{head}$ and $T_{head}T_{foot}T'_{head}$ satisfy the triangle similarity principle,

$$\frac{T_{head}T_{foot}}{OG} = \frac{T_{foot}T'_{head}}{GT'head} \tag{9}$$

From (9), the required target height can be estimated as,

$$T_h = T_{head}T_{foot} = \frac{T_{foot}T'_{head}}{GT'head} \times O_cG \tag{10}$$

where$T_{foot}T'_{head} = GT'head - GT_{foot} \times T_h$ is estimated by combining Eq. (9),

$$T_h = \frac{GT'head - GT_{foot}}{GT'head} \times O_cG \tag{11}$$

Solving for T_h from Eq. (9) by combining Eq. (8) and (5),

$$T_h = \left(\frac{\frac{O_cG}{tan\theta_2} - \frac{O_cG}{tan\theta_1}}{\frac{O_cG}{tan\theta_2}} \right) \times O_cG \tag{12}$$

$$T_h = \left(1 - \frac{tan\theta_2}{tan\theta_1}\right) \times O_cG \tag{13}$$

Therefore, using the known camera parameters, namely, camera height (C_h), the tilt angle (α) and the focal length (f) it is possible to estimate the target distance (T_d) and target height (T_h).

The Target height remains constant throughout the camera's FOV; however, the closer the target is to the camera, the larger it appears occupying a larger space in the proportion of the frame according to the perspective projection.

$$FOV = \arctan\left(\frac{Sensorheight(S_h)}{2 \times focallength(f)}\right) \tag{14}$$

$$\text{Image scale} = \frac{Targetheight(T_h)}{2 \times Targetdistance(T_d) \times ProportionofFrame} \tag{15}$$

where $tan(HalfVFOV)$ relates the angle subtended by the target in the camera's view and *ProportionofFrame* is the fraction of vertical frames that the target occupies. A depiction of which is given in Fig. 4. The on-filed application of implementing a missing child detection using the proposed human height estimation to enhance security is given in Table 3.

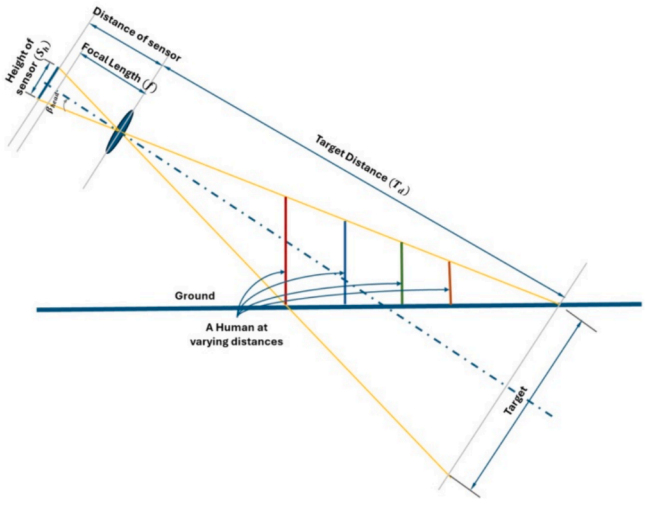

Fig. 4. The geometric relationship between camera settings and target parameters using perspective projection that shows how target distance influences target height.

Table 3
Strategic plan for implementing surveillance system on on-filed applications to enhance security and locate missing children with the proposed human height estimation.

Strategy component	Description
Objective setting	Enhance security and aid in locating lost children.
Camera Placement	Install surveillance cameras at strategic locations like entrances to capture full-body images, reducing perspective distortion.
Initial Calibration	Use known calibration patterns to set initial distortion parameters (radial and tangential).
Deployment Timing	Implement the system during less busy hours to minimize disruptions.
Algorithm Usage	Utilize a specialized algorithm to detect individuals and estimate height using reference points or scales within the image.
Physical Scales	If natural reference points are lacking, use physical scales or marks at known heights in the camera's field of view.
Performance Monitoring	Continuously monitor and adjust the system based on performance data.
Emergency Response	Quickly identify lost children or individuals matching security profiles to enhance responsiveness.

4. Experimentation

4.1. Camera calibration

Recovering the three-dimensional (3D) organization of a prospect from its 2D images is a fundamental challenge. The working environment is represented in a world coordinate plane as a three-dimensional scene. As fixed surveillance cameras are employed, single-view calibration technique is employed. This method do not require multiple varied views but can use known objects within the camera's view or specialized calibration patterns at different times or under different conditions to estimate and correct distortion parameters. The camera is calibrated here using Zhang's technique [43]. Calibration is performed using a widely established calibration pattern, an asymmetric checkerboard mounted onto a planar surface (with a square size of 7 cm in world units). Fig. 5a depicts this arrangement.

Fig. 6a depicts the frame with distortion and Fig. 6b shows frame after distortion correction. From the two sample frames shown in Fig. 6, it is evident that there exists radial distortion with an estimation of ($-0.6509, 0.4041$) as the wall present in the left FOV of the camera appears to be bend inwards similar to barrel distortion. To capture a comprehensive depiction of the model plane, it is advisable to obtain a series of images from various perspectives. This can be achieved by manipulating the plane's position. The plane was moved in various directions to ensure it occupied 80 percent of the scene. Fig. 7 shows sample frames used for calibrating the surveillance camera with multi-views of the checkerboard pattern. Next, the feature points were detected with their corresponding re-projected points and the distortion error shown in Fig. 5b through a meticulous examination of the disparities observed between the established coordinates of the calibration

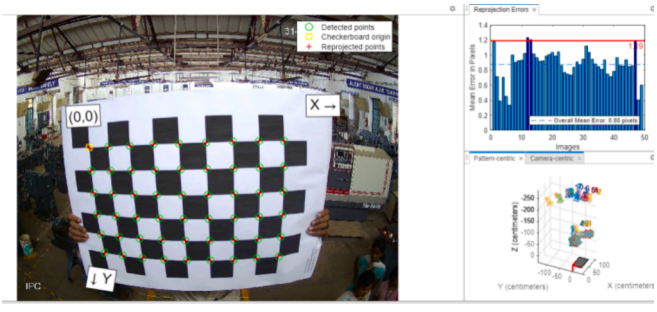

Fig. 5. Camera calibration using Zhang's [34] calibration method.

(a)Frame with distortion (b) Distortion corrected frame

Fig. 6. A frame from the raw surveillance footage with radial distortion and b shows the same frame after distortion correction.

Fig. 7. Sample frames used for calibrating the surveillance camera with multi-views of the checkerboard pattern.

pattern and their corresponding positions as detected in the images, the projection matrix is approximated. The projection matrix is estimated based on the relationship between the three-dimensional point P_M and image projection point P_m is given by,

$$s^*\widetilde{P_m} = A[R|t]\widetilde{P_M} \tag{16}$$

where s represents an arbitrary scaling applied to the projected image points $\widetilde{P_m}$ represents the augmented image projection point, which is a 3D vector $[u, v, 1]^T$. The third component '1′ ensures that the representation is in homogeneous coordinates, which are commonly used in computer vision and projective geometry. $\widetilde{P_M}$ represents the augmented real world point $[X, Y, Z, 1]^T$. A represents the camera's intrinsic matrix. It encapsulates the camera's focal length, optical center, and lens falsification coefficients. So, in the context of the equation, $s^*\widetilde{P_m} = A[R|t]\widetilde{P_M}$, 's' scales the homogeneous image projection point $\widetilde{P_m}$, which is transformed from the 3D world coordinate $\widetilde{P_M}$ with (A[R|t]). The result is a scaled 2D image point.

$$A = \begin{bmatrix} \propto & \gamma & U_o \\ 0 & \beta & V_o \\ 0 & 0 & 1 \end{bmatrix}, Rt = [r_1 r_2 r_3 t] \tag{17}$$

where, α, β are the scaling factors of the image axes, γ, β reflects skewness and U_0V_0 reflects the coordinates of the principal points. The estimation process effectively accounts for lens distortion and performs calculations pertaining to the camera's internal and external geometry from Zhang's calibration algorithm given in Eq. (18). An estimation of radial distortion coefficients is obtained using the linear least-squares method.

$$\widetilde{u} = u + uk_1u^2 + uk_1v^2 + \sqrt{(k_2u^2 + k_2v^2)}]$$

$$\widetilde{v} = v + vk_1x^2 + vk_1y^2 + \sqrt{(k_2x^2 + k_2y^2)} \tag{18}$$

where $k1$ and $k2$ are the coefficients of the radial distortion.

As a result, the distortions errors are eliminated. Therefore, it's important to acknowledge camera calibration as an essential pre-processing step used in computer vision to acquire precise measurements and reliable information from two-dimensional images by correcting lens distortion. Even when camera parameters (like focal length, sensor size) are known, individual variations from manufacturing processes can affect the actual performance of the lens and sensor. Over time, these components may also degrade or shift slightly, altering their characteristics. Here, calibration helps in adjusting to these changes and maintaining accuracy. In real-world scenarios, optical imperfections and assembly inaccuracies can lead the actual performance to deviate from these ideal parameters. Without the ability to reposition the camera to assess how distortions vary with angle and position, estimating these distortions accurately becomes challenging. Single-view calibration uses known geometrical properties in the scene or assumes model constraints to estimate these parameters. Single-view calibration often relies on certain assumptions about the camera model (like the lens being radially symmetric) or the scene (like the presence of parallel lines or planar surfaces).

4.2. Network setting of YOLOv7-OA detector

YOLOv7-OA is trained on the Tesla A100 GPU with 10 GB RAM written in Python-3.9.16 on google-colab platform. The training process of YOLOv7-OA is given as summarization in Algorithm 1. The model is trained using input images of dimensions 640 × 640 × 3, $Batch_{size}$ of 32, employing the $Optimizer_{StochasticGradientDescent}$ algorithm. Specific hyper-parameters are employed for the SGD algorithm. The anchor's threshold is set to 4.92 anchors/ target with 0.997 as Best Possible Recall (BPR). The specifications of the hardware are given in Table 4. This shows that the current anchors are well-suited for the proposed dataset. Warm-up training is used to ensure the model's stability and to reduce oscillations caused by high initial learning rates during training. This entails using a uniform learning rate throughout all experiments. The warm-up phase gradually increases the learning rate from 0 to a set value of 0.01, after which the cosine annealing algorithm updates the learning rate according to a specific schedule. During training, the network continuously updates the parameters to accelerate network convergence and prevents over fitting. The best results were obtained at 96 epochs. The best.pt file was selected based on the epoch which had the model's highest fitness score calculated as a weighted combination of metrics map@50 and map@50:95. The weight file saved during the training process is utilized for further analysis of the trained YOLOv7-OA. The optimal model is saved in a file called best.pt. The final output of YOLOv7-OA detector predicted using the weight file provides the localization of target class 'Person' with bounding boxes and the corresponding probability percentage of each detected target.

4.3. MDORI zones

A zone-based estimation strategy is introduced to enhance the accuracy of height estimation, accounting for different object distances. The scene environment is divided into five zones based on the European Union standard (EN 62676–4:2015) for security applications. The five

Table 4
Training setting of YOLOv7-OA.

Specifications	Parameters
GPU	NVIDIA A100-SXM
Driver Version	525.105.17
CUDA Version	12.0

different zones are Identification (zone 1), Recognition (zone 2), Observation (zone 3), Detection (zone 4) and Monitoring (zone 5). It is detailed in Table 5. They are divided based on the pixel density and the widest FOV estimated based on the Eq. (19):

$$\text{Widest FOV} = (TotalHorizontalpixelresolution)/PPM \tag{19}$$

where, the horizontal pixel resolution represents the width of 2D image captured by the camera in pixels. Pixel density (PPM) is a measure of detail in surveillance images, crucial for different levels of monitoring within MDORI zones. The widest FOV is estimated using the Eq. (15) allowing for optimal camera placement and configuration to meet the requirements of each zone. This ensures that the surveillance system captures the necessary detail for effective monitoring, detection, observation, recognition, and identification. IP video design tool [35] is used for simulating and analyzing video feeds and camera setups. The scene environment designed using the IP video design tool is shown in Fig. 8.

In the real world, the distance between each zone is the same, but when based on the 2D imaging system, the pixel information of each zone gets varies depending on the projections. Based on the above analysis, the subject's height is estimated when they move into zone 3 (Observation). Thus, this measure ensures that the height estimation correlates with the actual height of the subject.

4.4. Performance measures

4.4.1. Camera calibration

Re-projection error is global measure of geometric error which measures the difference in distance between the points detected in the image and the points re-projected back onto the image. CCTV cameras are prone to radial distortion, which generates skewness since beams of light bend towards the edges of the lens. Thus, the re-projection error as given in Eq. (20) and the radial distortion were computed for N images with M calibration points in each image.

$$E = \sqrt{\left(1/N^{*}\sum (u-u')^2 + (v-v')^2\right)} \tag{20}$$

where: (u, v) represent the original 2D image coordinates and (u', v') are the re-projected 2D image coordinates of the point estimated using the calibrated camera parameters.

4.4.2. YOLOv7-OA performance indicators

To quantitatively evaluate the model's effectiveness the YOLOv7-OA model, the following well-established evaluation metrics are used, following the widely acknowledged technique of Caltech.

Precision calculates provides an estimate of the accuracy of the YOLOv7-OA model in correctly identifying the target in the image with respective to positive outcomes. It is defined as given by:

$$Precision(P) = TP/(TP+FP) \tag{21}$$

The Recall represents the ratio of occurrences that are truly positive and are correctly identified as positive by the classification model to the total number of occurrences of classes present in the image given by Eq. (18):

$$Recall = TP/(TP+FN) \tag{22}$$

where, TP = True Positive: the number of correctly detected instances of children and adults. FP = False Positive: the number of incorrectly detected instances of children and adults. "False Negative (FN)" refers to instances where the proposed model failed to detect children or adults present in the image. Using the trapezoidal rule, computations were performed, thereby establishing the average Precision (AP) for each class. This process is demonstrated as:

$$AP = \int_0^1 P(R)dR \tag{23}$$

Finally, the average Precision (AP) summarizes how well a model ranks positive instances higher than negative instances as presented in Eq. (24):

$$AP = 1/n\sum_{k=1}^{n}(P(k).\Delta r(k)) \tag{24}$$

where: n = is the total number of positive instances in the dataset. $P(k)$ is the precision at the k-th positive instance in the ranked list. $\Delta r(k)$ is the change in recall at the k-th positive instance compared to the previous positive instance. The AP provides an estimate of the overall accuracy of the YOLOv7-OA model in detecting the instances across all the classes.

Table 5
Description of VFOV Zones.

Zones (Description)	PPM	Widest VFOV (m)
Identification: Positively identify a person beyond a reasonable doubt.	432	6
Recognition: Recognize a known individual.	216	12
Observation: Observe specific characteristics of a person.	103	25
Detection: Detects the presence of a person.	40	64
Monitoring: Used to monitor or crowd control	Below 40	Beyond 64

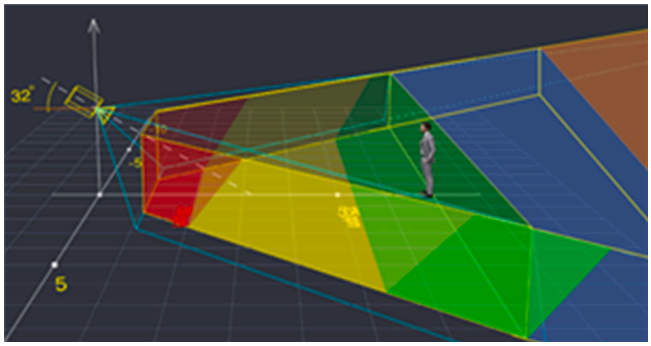

Fig. 8. The scene environment generated with IP video design tool [35].

4.4.3. Human height estimation performance indicators

The proposed human height estimation method is evaluted using estmated and relative error. The estimated error is the difference between the predicted height (T_h) and ground truth height of the target GT_h. It is denoted by e_h and defined as:

$$e_h = T_h - GT_h \tag{25}$$

where T_h is the predicted target height and GT_h is the ground truth height of the target.

The relative error gives a sense of the size of the error compared to the ground truth height. It is expressed in percentage to get a clear identification of the error magnitude relative to the ground truth. The relative error is given by r_h and defined as:

$$r_h = (T_h/e_h) \times 100\% \tag{26}$$

where T_h is the predicted target height and e_h is the estimated error in target height estimation.

The maximum and minimum erros is measured using the largest and smallest error in absolute terms among the estimated errors e_{h_i}, given as:

$$E_{h_{max}} = max(|E_{h_1}|, |E_{h_1}|, \cdots |E_{h_n}|) \tag{27}$$

$$E_{h_{min}} = min(|E_{h_1}|, |E_{h_1}|, \cdots |E_{h_n}|) \tag{28}$$

The standard deviation (σ_h) of the human height estimation error is obtained as given by:

$$\sigma_h = \sqrt{\sum_{i=1}^{n}(e_{h_i} - \mu/n)} \tag{29}$$

where n is the total number of height measurement samples, μ is the mean error given by:

5. Results and discussion

5.1. Comparative analysis of YOLOv5, v6, v7 and v8

YOLOv7-OA demonstrates efficient performance across a wide range of video frame rates, from 5 frames per second (fps) to 270 fps. It achieves an impressive AP rate of 56.8 percent on ImageNet dataset, indicating its accuracy in object detection tasks. YOLOv7 surpasses the performance of its predecessors, YOLOv5 and v6, successor v8, which are convolution-based object detectors. Running the YOLOv7 model on the same dataset is 50 percent more cost-effective with high speed and accuracy improvements. The hidden layer parameters in the neural network are reduced by up to 40 percent, indicating model optimization without compromising performance. It achieves a 1.5 times higher average precision compared to previous versions with significant reduction in parameters and computational time (75 percent fewer parameters and 36 percent less time). As evident from Fig. 9a, YOLOv7 exhibits the highest confidence score of 0.84 when detecting individuals at a distance zones compared to YOLOv5, YOLOv6, and YOLOv8 which achieve scores of 0.71, 0.66, and 0.45 respectively. In contrast YOLOv8 exhibits the lowest confidence score. It struggles to detect individuals in more distant zones. In contrast, YOLOv7 achieves the highest confidence score of 0.84. Regarding the speed of the four object detectors, YOLOv6 demonstrates the lowest average FPS of 12.82, while YOLOv8 and YOLOv5 follow closely behind with 16.3 and 16.98. Surprisingly, YOLOv7 outperforms all, topping the average FPS charts with an impressive 20.17 FPS.

In Fig. 9b, captured under low-illuminated and lower-resolution conditions compared to Fig. 9a, YOLOv7 demonstrates superior performance compared to other models. It excels in detecting a person in a bending position with the highest confidence score of 0.78, outperforming YOLOv8, YOLOv6, and YOLOv5, with confidence score of 0.61, 0.33, and 0.30, respectively. This highlights YOLOv7′s capability to detect individuals in diverse poses. In terms of handling occlusion, YOLOv8 faces challenges in distinguishing individuals when occluded and often detects them as a single entity as seen in Fig. 9b, whereas other models successfully identifies and addresses occlusion. Among these models, YOLOv6 performs the best, achieving a confidence score of 0.77, followed by YOLOv7 with 0.73 and YOLOv5 with 0.61. However, it's worth noting that YOLOv6 has the slowest processing speed among all models. Though YOLOv8 is a successor of YOLOv7, it is evident from the results that they have limitations in detecting objects in low-light environment, low resolution and when they are partially occluded.

Considering these various scenarios, YOLOv7 emerges as the most well-rounded choice, effectively balancing performance under different lighting conditions, lower resolutions, and maintaining a favorable trade-off between speed and accuracy. But, compared when it comes to occlusion, YOLOv7 suffers negligibly compared to YOLOv6. Recognizing the importance of occlusion handling, YOLOv7 is trained on a diverse dataset that includes images with various occlusion levels. This augmentation helps improve the model's robustness in detecting partially obscured targets. YOLOv7-Occlusion Aware is introduced as a modified version of YOLOv7, specifically designed to handle occlusion challenges. A hybrid attention mechanism (HAM) is incorporated into the feature extraction part of the original YOLOv7 to help the model focus on relevant regions and contextual information. It is believed that the combination of data augmentation and the HAM technique could enhance occlusion handling. The experimental findings show that YOLOv7-OA is effective in handling occlusion circumstances, with average detection accuracy (mAP) of 98.86 %. This shows high precision in recognizing targets, even partially obscured.

Notably, the YOLOv7-OA model improves performance while lowering the amount of model parameters, making it computationally efficient. In a variety of fields, including surveillance, security, and human pose estimation, YOLOv7-OA can be used to estimate target height by recognizing occluded individuals. The HAM is introduced as part of the feature extraction section of YOLOv7. YOLOv7-OA achieved average detection accuracy (AP) of 98.86 % and reduced the number of model parameters from 37.62 MB to 33.71 MB, making it superior at recognizing persons with less than 95 % occlusion. The YOLOv7-OA model recognizes occluded individuals accurately, giving an excellent approach for precisely calculating target height.

Fig. 9. Detection of 'Person' class by YOLOv5, v6, v7 & v8. The target distance in Frame 1 is 17.25 m and Frame 2 between 4 and 7 m as there are four targets each at a different distance from the camera.

5.2. Comparative study with SOTA models

This section compares the proposed human height estimator with other state-of-the-art algorithms, that includes Shi et al. [27] based on a parametric way with monocular cameras, Matveev et al. [26] based on geometric scene-related parameters, Kainz et al. [8] based on calibrated scaling factor, Dea'k et al. [19] based on the anthropometric measurement, and Criminisi et al. [16] based on the linear perspective principle. Fig. 10 depicts the human height estimation at various FOV. Fig. 11 a comparative analysis of several classical state-of-the-art (SOTA) methods for estimating human height from diverse positions (e.g., standing, walking, bending, sitting, and running/jumping) is presented in the figure. The evaluation of each approach is conducted by quantifying the error in height estimation, a metric that exhibits substantial variation among various human poses and thus reflects the distinct obstacles that each pose poses for height estimation methods. Cameras in real life surveillance systems are generally installed at heights ranging from 5 m to 15 m with typical coverage area of 150 m and tilt angle ranging from 15 to 60 degrees.

However, the height of camera used in the experiments in state-of-the-art studies is comparable to the person's height and range is between 10 % of actual installations. Also, horizontal range is quite restricted. Hence, the selected parameters may be responsible for the low error in person height estimation. Deak et al. rely on average body proportions, where slight variations in body proportions significantly affect height estimation accuracy. Factors like posture, clothing, camera angle, and camera resolution introduced calculation errors. It only provided a rough estimate but not a precise measurement. It has been inferred from the analysis that the accuracy of body proportion-based height estimation is highly dependent on the camera angle and distance to the person. Shi et al. proposed a parameterized strategy for detecting the distance as well as the pedestrian's height. Combining an enhanced block differencing method with global contour detection aided in removing complex backgrounds. It enhanced the accuracy of extracting the actual contour of the pedestrians. Changes in camera perspective and varying distances introduced significant errors. Also, it has shown an inability to handle occlusion or partial body visibility. In Kainz's height estimation using a calibrated scaling factor, environmental and operational factors pose considerable challenges. Lighting conditions and video quality impede the algorithm's ability to detect a moving target, further complicating height measurements.

Criminisi et al. suffers to accurately measure a person's Height when he walks, sits, or bends expect when the subject is present at the center of the image at a specific location. Matveeve et al. used Tsai camera calibration for reliable height estimation. Even minor errors in the camera parameters, such as focal length or sensor size, introduced errors in the height estimation results. Also, it incurs errors in complex geometries or uneven height distributions. Our method outperforms other approaches by a considerable margin. Notably, our method achieves an average relative error of approximately 0.90, which shows that it can accurately determines the height from the given image data with the complex dataset.

(a) Before calibration (b) Calibrated image (c) Height estimation

Fig. 10. The height estimation results obtained inn various zones of the FOV. Row 1 represents Zone 1, and the corresponding rows represent Zone 3, 2 and 1.

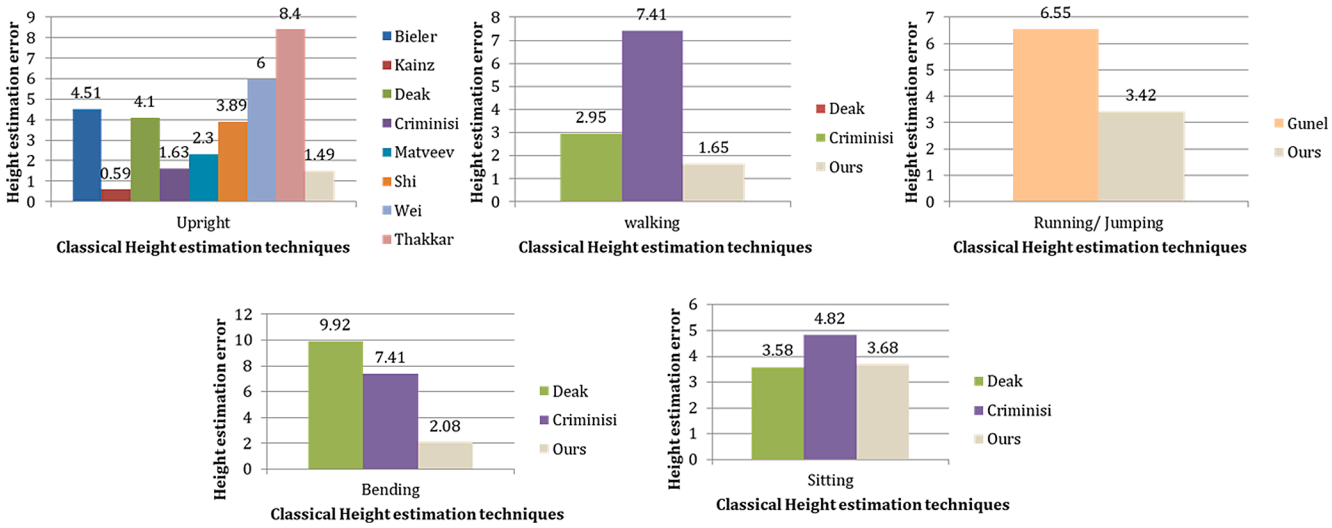

Fig. 11. A comparative study of classical methods for estimating human height from different poses.

The proposed method performs well in upright and walking poses, with low error rates. Further, False positives in height estimation occurs when an individual is seated or kneeling, making it difficult for the system to accurately determine their height. To further reduce the occurrence of false positives, the removal of outliers from the estimation process is done. An average predicted height is observed in each zone. Outliers are the height measurements that significantly deviate from an average predicted height of children in each zone. By excluding these outliers, the accuracy of the height estimation can be improved. Despite these challenges, the proposed method maintains lower error rates in running/jumping poses compared to other SOTA methods.

5.3. Influence of camera height and tilt angle

Human height estimation involves dealing with certain variables that influence accuracy. Two crucial factors are the camera height C_h and the tilt angle (θ). The camera height was measured with a measuring tape held between the ground plane and the camera's optical center.

The tilt angle was measured with a digital protractor. The ground truth camera height and tilt angle are 274 cm and 24.8 $^\circ$. To comprehensively understand the impact of camera height and tilt angle, an extensive experimental analysis was conducted. This analysis involved two distinct scenarios: in the first case, the camera height was fixed while the tilt angle was varied, and in the second, the tilt angle was fixed while the camera height was altered.

In the first case, the camera height is fixed at 274 cm and the tilt angle is varied between the ranges of $\pm 1^\circ$ with an interval of 0.1 $^\circ$. From the analysis, it is evident that a discrepancy in tilt angle of up to $\pm 0.4^\circ$ is acceptable. During this interval, the estimation process consistently yields accurate results with relatively negligible error of 0.0023 cm. However, beyond this range, the estimated error values starts to increase noticeably to ± 1.4833 cm. In the same way, from the ground truth Camera height of 275 cm, a variation of $\pm 10cm$ with an interval of $1cm$ was analyzed. It is evident from the analysis, that a discrepancy of $\pm 2cm$ is acceptable with a negligible error of $\pm 0.1333cm$. Beyond the acceptable range, the height estimation error increases proportional to the Camera height. Fig. 13 depicts the influence of camera height and tilt angle on target height estimation. Fig. 12 presents two graphs illustrating the influence of camera setup parameters on the relative error in height estimation across various zones. The first graph shows the relative error in height estimation across various zones for varied camera heights of 8 feet, 10 feet, and 12 feet. As the distance from the camera increases, there is a general trend of increasing error for all camera heights, with the 8 feet height demonstrating the highest variability and peaks in error. Conversely, the 12 feet height setting generally exhibits lower error rates, suggesting that a higher camera placement may be more stable for height estimation. The second graph evaluates height estimation error for various camera tilt angles of 15, 20, 25, 30, and 35 degrees. The relative error in height estimation remains within a tighter range, with the 15 and 20-degree angles showing less error variation across different distances. As the tilt angle increases, the error becomes more variable, especially at 30 and 35 degrees. Both graphs display an overall trend where the error in height estimation becomes more significant as the distance increases. Camera height plays a critical role in the stability of the measurement over distance, with higher camera placements offering potentially more reliable results. A moderate

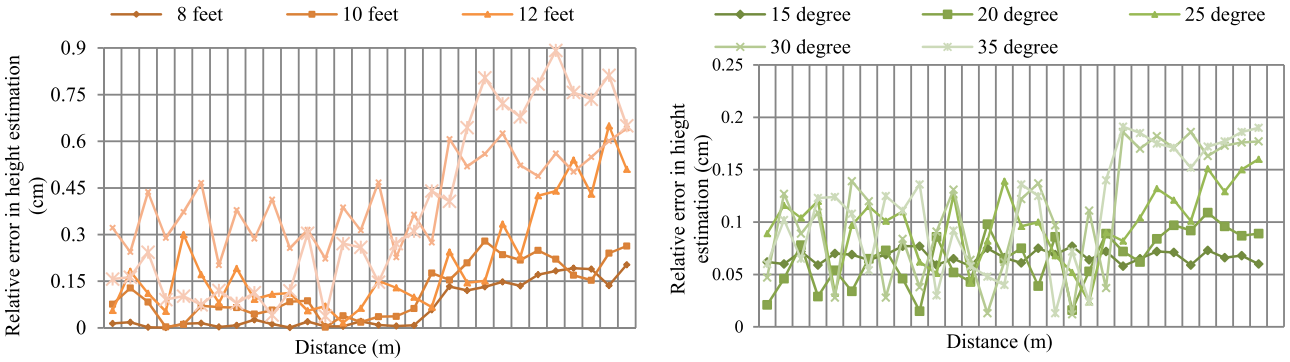

Fig. 12. Influence of camera height and tilt angle on human height estimation.

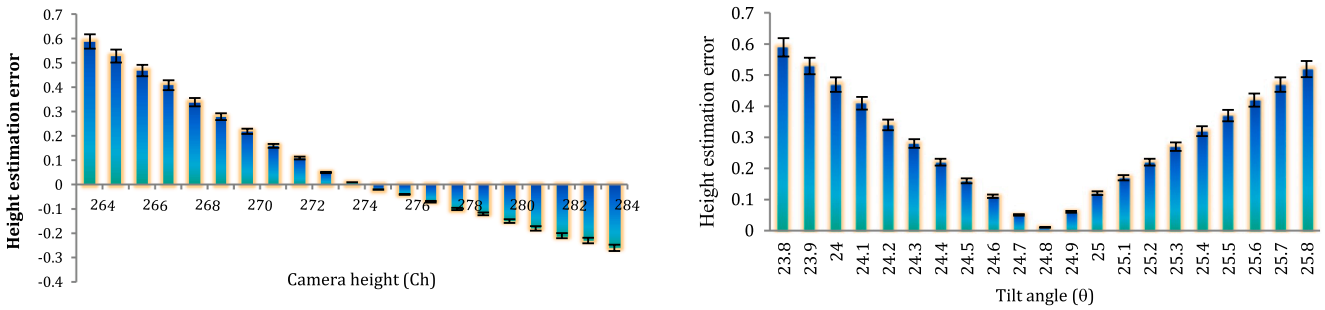

Fig. 13. Influence of camera height and tilt angle on human height estimation with respect to object distance.

camera tilt is preferable for maintaining lower and more consistent error rates. The data from these graphs are critical for informing the optimal setup of cameras for height estimation tasks. For practical applications, it would be important to choose a balance between camera heights and tilt angle that minimizes the overall relative error, especially at the distances at which height estimation is most critical for the application at hand. It is observed from Fig. 14 that the distortion is typically more pronounced towards Peripheral zones ('More Left' and 'More Right'). Target that are in the far field of the camera (Monitoring zone) appears smaller due to perspective, which can make height estimation more challenging and potentially less accurate, leading to higher error rates in these zones. As subjects move to the peripheral zones, they move closer to the boundaries of the camera's field of view. This leads to geometric distortion where the shapes and sizes of targets are misrepresented, causing errors in height measurement.

5.4. Influence of horizontal field of view and vertical field of view

During the experimentation process, the focus was on capturing the target individual as they walked within the camera's FOV. To achieve this, the subject was directed to walk a distance of approximately 200 cm while remaining within the camera's scope. In order to ensure a detailed analysis, the frames obtained were systematically filtered. This filtering process was based on the distinct markings on the ground plane within the FOV at intervals of 20 cm. The target's position within the camer's FOV (HFOV and VFOV) significantly affects the human height estimation in video surveillance system. From Fig. 14, the experimental analysis is evident that when the target is either towards the extreme left or right of the HFOV (Peripheral zones), the error in height estimation ranges between ± 1.87 cm. Whereas when the target is present directly proportional to the camera i.e. in the exact middle of the HFOV (central zones) the error is negligible, with ± 0.29 cm. The data also revelas that the error in heigt estimation is influenced by the target distance in the VFOV. When the target is farthest away (far field – Monitoring zon) from the camera i.e. Zone 5, the error ranges between $\pm 0.5cm$ and $\pm 0.73cm$. When the target is Zone 3 (Observation zones) from the camera, the error in height estimation is the lowest with $\pm 0.04cm$ followed by Zone 2 and 1 with $\pm 0.9cm$ and $\pm 1.24cm$. In Zone 3, the angle of view is optimized, reducing the extent of foreshortening. This allows for accurate proportionality in the height estimation. Extreme camera distances (very close or very far) can cause perspective distortion, affecting the accuracy of height estimation. In Zone 1, the camera is too close, leading to significant perspective distortion and foreshortening, making accurate height estimation challenging. Whereas in Zone 5, the camera is too far away, causing the person to appear small in the frame resulting in errors due to the lack of visible details for accurate object detection. Hence, Zone 3′s moderate distance minimizes the distortion.

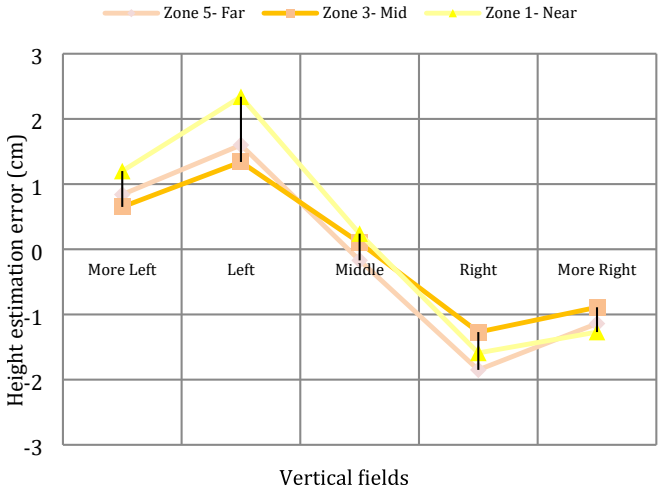

Fig. 14. Influence of the vertical and horizontal fields on height estimation.

In practical terms, these findings is applied to optimise the placement and orientations of the cameras. Therefore, when target cross the specified green zones, a more precise height estimations can be achieved, which can have implications in various applications like security, monitoring, and analytics. Heights measured in the frames extracted from the HeightSense dataset footage had a total mean absolute error of 0.42 cm considering error obtained with the twenty targets (0.749,−0.282,−0.244,−0.495,0.399,−1.061,0.805,−0.351,0.147,-−0.115,0.674,−0.950,−0.149,−0.177,0.523,−0.507,−0.080,-−0.405,0.019,0.269), with a minimum and maximum error of −1.061 and 0.805 cm and a standard deviation of 0.507 cm.

5.5. Influence of gender and age in height estimation

Based on the gender and age of the participants, a detailed experimentation was carried on to analyze their influence over height estimation. From the 35 participants 28 were choosen for the analysis of the influence of gender and age over height estimation. The partcipants were choosen in such a way that equal importance was given for age and gender where they were divided into 4 groups each with 7 participants. Table 6 summarises the performance analysis height estimator and the influence of age and gender at different FOV. From, the analysis it is evidend that the effect of gender and age has only a minor influence on height estimation accuracy, as, it does not hinder with the parameters of height estimation in anyways.

5.6. Influence of image scale on height estimation in HFOV

This image scale (I_s) represents the how many millimeters in real world, each pixel in the image represents at near-far fields. It is shown in Fig. 4. The image scale per pixel influences the human height estimation at different zones. The influence of Image scale on human height estimation at different zones of the FOV of the camera is given in Table 7. It is observed from the image scale estimation, that at near-field central zone (Identification zone) the height measurements are more precise, where the scale is approximately 4.44 mm per pixel. Whereas, at far-

Table 6
Performance analysis of the influence of age and gender on height estimation.

Tag	Actual Height (cm)	Gender (M/F)	Distance 5.28 *m*		Distance 7.28 *m*		Distance 10.28 *m*		Distance 15.28 *m*	
			Estimated error (cm)	Relative error (cm)	Estimated error (cm)	Relative error (cm)	Estimated error (cm)	Relative error (cm)	Estimated error (cm)	Relative error (cm)
C1	149.8	M	−0.51	−0.0034	−0.99	−0.0066	3.26	0.0218	5.26	0.0351
C2	130.3	M	1.62	0.0201	−1.59	−0.0122	−4.03	−0.0309	−4.31	−0.0331
C3	147.2	M	−0.7	−0.0387	1.65	0.0112	8.14	0.0553	3.36	0.0228
C4	164.4	F	2.53	0.0154	1.03	0.0063	−2.95	−0.0179	−4.49	−0.0273
C5	152.7	F	−4.2	−0.0275	−2.98	−0.0195	−0.08	−0.0005	−2.35	−0.0154
C6	126.7	F	1.36	0.0107	1.65	0.0130	2.72	0.0215	4.86	0.0384
A7	182.9	F	0.57	0.0031	1.41	0.0077	−2.14	−0.0117	−3.67	−0.0201
A8	171.5	M	−1.36	−0.0079	−1.88	−0.0110	1.62	0.0094	−5.57	−0.0325
A9	168.3	M	1.86	0.0111	−1.84	−0.0109	−1.25	−0.0074	−4.12	−0.0245
A10	189.3	F	0.63	0.0033	0.33	0.0017	−2.36	−0.0125	−3.78	−0.0200
A11	163.1	F	−2.57	−0.0158	1.26	0.0077	2.91	0.0178	−2.75	−0.0169
A12	170.6	M	1.74	0.0102	1.19	0.0070	−1.56	−0.0091	−1.47	−0.0086

Table 7
Influence of Image scale on human height estimation on different zones of the FOV of the camera.

VFOV	HFOV	Target distance (*m*)	Scale per pixel (*mm*)	Height estimation errors (*mm*)	Influence on Height measurement	Image scale
Identification (*I*)	Central	6.6	4.44	± 1.89	High accuracy due to reduced distortion and better focus.	Objects appear larger and more detailed.
	Peripheral	7.2	3.92	± 2.65	Challenging due to lens distortions which can warp object appearance.	
Recognition (*R*)	Central	12.3	8.88	± 1.32	Clearer images support better feature extraction and classification, enhancing recognition accuracy.	Scale is consistent and proportional. High fidelity in scale aids in feature recognition.
	Peripheral	14.7	6.32	± 1.53	Shape and size distortions may lead to misrecognition; algorithms need to compensate for distortions.	
Observation (*O*)	Central	25.4	18.50	± 0.47	Ideal for analyzing detailed attributes necessary for precise identification due to high image fidelity.	Scale remains constant, facilitating consistent observation.
	Peripheral	28.8	18.03	± 0.82	Degraded quality of object features; features may appear smeared or stretched, reducing effectiveness.	
Detection (*D*)	Central	64.6	47.35	± 1.02	Stable and reliable view with consistent quality, crucial for accurate long-term analysis.	Scale is consistent and proportional. High fidelity in scale aids in feature recognition.
	Peripheral	65.2	33.72	± 4.32	Variable image quality and potential vignetting can mislead analytical models.	
Monitoring (*M*)	Central	Above 64	54.31	± 5.26	Supports detailed and prolonged examination without significant quality degradation.	Scale remains constant, facilitating consistent observation.
	Peripheral	Above 72	43.72	± 7.32	Lower image quality limits detailed monitoring tasks; it requires repositioning cameras to bring critical areas into focus.	

field (Detection zone), image scale is approximately 47.35 mm per pixel. The coarser scale shows that each pixel covers a wide area, which results in blurring of finer details and leads to a less precise height estimation.

5.7. Influence of crowded environment on human height estimation

Fig. 15. Human height estimation using YOLOv7-OA in a densely packed scene. We conducted experiments in environments with different crowd density to evaluate the robustness and effectiveness of our proposed human height estimation method. The results demonstrate that our YOLOv7-Occlusion Aware (YOLOv7-OA) detection model and the hybrid attention mechanism (HAM), performs well even in challenging conditions. Despite the occlusions and overlaps typical in such scenarios, our method accurately detected and located individuals. The hybrid attention mechanism enabled the model to focus on pertinent features and contextual information, improving the accuracy of height measurement even when individuals were partially obscured. The ability to handle moderate-density environments enhances the applicability of our approach in real-world surveillance scenarios, where such conditions are common. We believe these results demonstrate the practical value and reliability of our algorithm. In the future, this research can be extended to high-density crowded scenarios.

6. Conclusion

The proposed method employs a comprehensive approach to addressing the issues of accurate human height estimation in surveillance settings. It starts with camera calibration, which is a necessary first step in correcting lens-induced distortions. Individuals are precisely found within the camera's field of view using the deep learning-based target detection approach YOLOv7-OA, paving the way for height estimate based on the perspective principle. There is also an examination of the interaction between camera height and deflection angle in terms of accurate estimation of human height across various zones inside the camera's FOV. The approach has been shown effective through thorough experimentation on real-world datasets. With a mean absolute error ranging from 0.02 cm to 0.8 cm in different field of view zones, the technique consistently outperforms existing alternatives. This level of precision far superiors the current state-of-the-art performance, which is 1.39 cm based on ground truth data. In summary, the research presents a novel human height estimator for surveillance systems as well as a useful human height surveillance dataset. By addressing the challenges of

Fig. 15. Human height estimation using YOLOv7-OA in a densely packed scene.

camera calibration, target detection, and perspective-based height estimation, it provides a robust and practical solution to improve urban surveillance capabilities, ultimately assisting in more effective missing child and person retrieval and tracking within complex urban environments.

7. Future work

False positives in height estimation can occur when a person is seated or kneeling, making it difficult for the system to accurately determine their height. To address this problem, a future work can be implemented by incorporating posture recognition techniques into the height estimation process. By accurately detecting and recognizing different body postures, the system can adapt its estimation algorithms accordingly. One of the best approaches is to rely on upper body measurements and apply known anthropometric ratios to extrapolate the full standing height. This approach takes into account the unique human body proportions and can compensate for the seated position. Additionally, analyzing the surrounding environment, including chairs or seating furniture in the vicinity, can provide additional cues to confirm the seated posture.

CRediT authorship contribution statement

K. Iyshwarya Ratthi: Conceptualization, Data curation, Methodology, Software, Validation, Writing – original draft. **B. Yogameena:** Conceptualization, Investigation, Methodology, Supervision, Visualization, Writing – review & editing. **S. Saravana Perumaal:** Conceptualization, Methodology, Software, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

The authors would like to express their sincere gratitude to the All-India Council for Technical Education (AICTE), Anna University, and Thiagarajar College of Engineering for their invaluable support through the Doctoral Fellowship. This support has been pivotal in successfully completing this research work.

References

[1] C. Galiano López, J. Hunter, T. Davies, A. Sidebottom, Further evidence on the extent and time course of repeat missing incidents involving children: a research note, The Police J: Theory, Practice and Principles 96 (1) (2023), https://doi.org/10.1177/0032258x211052900.
[2] L. Boulton, J. Phoenix, E. Halford, A. Sidebottom, Return home interviews with children who have been missing: an exploratory analysis, Police Pract. Res. 24 (1) (2023), https://doi.org/10.1080/15614263.2022.2092480.
[3] M. Shorfuzzaman, M.S. Hossain, M.F. Alhamid, Towards the sustainable development of smart cities through mass video surveillance: a response to the COVID-19 pandemic, Sustain. Cities Soc. 64 (2021).
[4] B. Hassan, E. Izquierdo, T. Piatrik, Soft biometrics: a survey, Multimed Tools Appl. (2021), https://doi.org/10.1007/s11042-021-10622-8.
[5] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, C. Busch, Demographic bias in biometrics: a survey on an emerging challenge, IEEE Trans on Technol. Soc. 1 (2) (2020) 89–103.
[6] Günel, S., Rhodin, H., Fua, P. (2018). What Face and Body Shapes Can Tell About Height. *CoRR, abs/1805.1*.
[7] Di Bieler, S.G. Gunel, P. Fua, H. Rhodin, Gravity as a reference for estimating a person's height from video, Proce. IEEE Int. Conference on Comp. Vision (2019), https://doi.org/10.1109/ICCV.2019.00866.
[8] F. Yin, S. Zhou, Accurate estimation of body height from a single depth image via a four-stage developing network, Proce. IEEE Computer Society Conference on Comp. Vision and Pattern Recognition (2020), https://doi.org/10.1109/CVPR42600.2020.00829.
[9] O. Kainz, M. Dopiriak, M. Michalko, F. Jakab, P. Fecil'ak, Estimating the height of a person from a video sequence, 2021 19th Int. Conference on Emerging eLearning Technol. Applications (ICETA) (2021) 150–156, https://doi.org/10.1109/ICETA54173.2021.9726680.
[10] Zhu, R., Yang, X., Hold-Geoffroy, Y., Perazzi, F., Eisenmann, J., Sunkavalli, K., Chandraker, M. (2020). Single View Metrology in the Wild. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12356 LNCS*. Doi: 10.1007/978-3-030-58621-8_19.
[11] D.S. Lee, J.S. Kim, S.C. Jeong, S.K. Kwon, Human height estimation by color deep learning and depth 3D conversion, Appl. Sci. (Switzerland) 10 (16) (2020), https://doi.org/10.3390/app10165531.
[12] A.M. Olver, H. Guryn, E. Liscio, The effects of camera resolution and distance on suspect height analysis using PhotoModeler, Forensic Sci. Int. 318 (2021), https://doi.org/10.1016/j.forsciint.2020.110601.
[13] N. Thakkar, H. Farid, On the feasibility of 3D model-based forensic height and weight estimation, IEEE Comp. Society Conference on Comp. Vision and Pattern Recognition Workshops (2021), https://doi.org/10.1109/CVPRW53098.2021.00106.
[14] N. Thakkar, G. Pavlakos, H. Farid, The reliability of forensic body-shape identification, IEEE Comp. Society Conference on Comp. Vision and Pattern Recognition Workshops (2022), https://doi.org/10.1109/CVPRW56347.2022.00014.
[15] J. Wei, J. Jiang, A. Yilmaz, MOHE-Net: monocular object height estimation network using deep learning and scene geometry, Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives 43 (B2-2021) (2021), https://doi.org/10.5194/isprs-archives-XLIII-B2-2021-557-2021.

[16] I.S. Kim, H. Kim, S. Lee, S.K. Jung, HeightNet: monocular object height estimation, Electronics (Switzerland) 12 (2) (2023), https://doi.org/10.3390/electronics12020350.
[17] A. Criminisi, I. Reid, A. Zisserman, Single view metrology, Int. J. Comput. Vis. 40 (2) (2000) 123–148.
[18] F. Tosti, C. Nardinocchi, W. Wahbeh, C. Ciampini, M. Marsella, P. Lopes, S. Giuliani, Human height estimation from highly distorted surveillance image, J. Forensic Sci. 67 (1) (2022) 332–344.
[19] Q. Li, L. Mou, Y. Hua, Y. Shi, S. Chen, Y. Sun, X.X. Zhu, 3DCentripetalNet: Building height retrieval from monocular remote sensing imagery, Int. J. Appl. Earth Obs. Geoinf. 120 (2023) 103311.
[20] A Déak, O Kainz, M Michalko, F Jakab. (2017) Estimation of human body height from uncalibrated image. In 2017 15th International Conference on Emerging eLearning Technologies and Applications (ICETA), pages 1–4. IEEE.
[21] J. Shao, S.K. Zhou, R. Chellappa, Robust height estimation of moving objects from uncalibrated videos, IEEE Trans. Image Process. 19 (8) (2010) 2221–2232.
[22] Yu. Chai, X. Cao, A real-time human height measurement algorithm based on monocular vision, in: In 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), IEEE, 2018, pp. 293–297.
[23] J. Jung, I. Yoon, S. Lee, J. Paik, Object detection and tracking-based camera calibration for normalized human height estimation, J. Sensors 2016 (2016).
[24] K. Koide, J. Miura, Identification of a specific person using color, height, and gait features for a person following robot, Rob. Auton. Syst. 84 (2016) 76–87.
[25] M. Johnson, E. Liscio, Suspect height estimation using the Faro Focus3D laser scanner, J. Forensic Sci. 60 (6) (2015) 1582–1588.
[26] Tejeda, Yansel González, Helmut A. (2022) Mayer. "Effect of Gender, Pose and Camera Distance on Human Body Dimensions Estimation." In Image Analysis and Processing. ICIAP 2022 Workshops: ICIAP International Workshops, Lecce, Italy, May 23–27, 2022, Revised Selected Papers, Part II, pp. 179-190. Cham: Springer International Publishing.
[27] I. Matveev, K. Karpov, I. Chmielewski, E. Siemens, A. Yurchenko, Fast object detection using dimensional based features for public street environments, Smart Cities 3 (1) (2020) 93–111.
[28] Z. Shi, Xu. Ziming, T. Wang, A method for detecting pedestrian height and distance based on monocular vision technology, Measurement 199 (2022) 111418.
[29] T.v. Nguyen, J. Feng, S. Yan, Seeing human weight from a single RGB-D image, J. Comput. Sci. Technol. 29 (5) (2014), https://doi.org/10.1007/s11390-014-1467-0.
[30] C. Pfitzner, S. May, A. Nüchter, Body weight estimation for dose-finding and health monitoring of lying, standing and walking patients based on RGB-D data, Sensors (Switzerland) 18 (5) (2018), https://doi.org/10.3390/s18051311.
[31] SMIT PATEL. (2020). *Heights and Weights Dataset*. Kaggle. https://www.kaggle.com/datasets/burnoutminer/heights-and-weights-dataset.
[32] J. Cao, Y. Pang, J. Xie, et al., From handcrafted to deep features for pedestrian detection: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 44 (9) (2022) 4913–4934.
[33] E. Arkin, N. Yadikar, X. Xu, et al., A survey: object detection methods from CNN to transformer, Multimed Tools Appl 82 (2023) 21353–21383.
[34] Wang, Chien, Alexey Bochkovskiy, Hong Liao. (2022) "YOLOv7: Trainable Bag-of-freebies Sets New State-of-the-art for Real-time Object Detectors." *ArXiv*. Accessed October 5, 2023. /abs/2207.02696.
[35] Burger, Wilhelm. (2016). Zhang's Camera Calibration Algorithm: In-Depth Tutorial and Implementation. 10.13140/RG.2.1.1166.1688/1.
[36] JVSG. (2023). *Video Surveillance Design Apps | JVSG*. https://www.jvsg.com/.
[37] S. Li, V.H. Nguyen, M. Ma, C.B. Jin, T.D. Do, H. Kim, A simplified nonlinear regression method for human height estimation in video surveillance, EURASIP J. Image and Video Processing 2015 (2015) 1–9.