Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 503–511
DOI: 10.15439/2023F3560
ISSN 2300-5963 ACSIS, Vol. 35

Estimation of absolute distance and height of people based on monocular view and deep neural networks for edge devices operating in the visible and thermal spectra

Jan Gąsienica-Józkowy*†, Bogusław Cyganek *†, Mateusz Knapik*†, Szymon Głogowski† and Łukasz Przebinda†
*Faculty of Computer Science, Electronics and Telecommunication,
Email: *cyganek@agh.edu.pl*
†MyLED Inc.
Email: *m.knapik@myled.pl*
Ul. W. Łokietka 14/2, 30–016 Kraków, Poland

***Abstract*—Accurate estimation of absolute distance and height of objects in open area conditions is a significant challenge. In this paper, we address these problems and we propose a novel approach that combines classical computer vision algorithms with modern neural network-based solutions. Our method integrates object detection, monocular depth estimation, and homography-based mapping to achieve precise and efficient estimations of absolute height and distance. The solution is implemented on the edge device, which enables real-time data processing using both visual and thermography data sources. Experimental evaluation on a height estimation dataset prepared by us demonstrates an accuracy of 97.06% and validates the effectiveness of our approach.**

I. Introduction

ACCURATE estimation of spatial positions and parameters of objects, such as their localization on a bird's-eye view map, absolute distance, and absolute height, is an important computer vision task with wide practical implications. In this paper, we propose a novel solution for absolute distance and height estimation that combines homography-based mapping algorithms with state-of-the-art deep learning techniques. Our approach harnesses the strengths of both classical and modern solutions to achieve highly accurate and efficient estimations under various conditions.

The proposed method integrates several key components to provide a comprehensive solution for absolute height estimation. Firstly, we capture video frames from visual and thermography cameras and input them into an object detector, specifically the YOLOv5 model [1]. This model enables robust identification and localization of objects in the monocular view. To estimate the relative depth information, we utilize a transformer-based monocular depth estimation model called DPT Levit 224 [2], [3]. This model learns to infer depth information from a single image, allowing us to determine the relative distances between objects in the scene. Additionally, we incorporate homography-based mapping techniques to establish correspondences between points in different images or views. By leveraging homography projection, we can accurately map objects from the video frame plane to the bird's-eye view 2D map, enabling easy estimation of their distance from the camera. The final stage of our approach involves polynomial regression-based estimations to compute the absolute distance and height of objects.

The proposed solution is implemented on the Arabox III-A edge device, which is based on the Jetson Nano board and offers real-time data processing capabilities. Arabox is specifically designed for fully anonymous data acquisition and is commonly used in the Digital Out-of-Home (DOOH) advertising industry.

In our experimental evaluation, we demonstrate the effectiveness of our approach and its evaluation on the prepared by us absolute height estimation dataset. The obtained results show accuracy equal to 97.06% in real-time performance, emphasizing the usefulness of our solution in a wide range of applications requiring precise absolute distance and height estimation.

The rest of this paper is organized as follows: Section 2 II provides a detailed overview of related works, including object detection, homography-based mapping, monocular depth estimation, and absolute height estimation techniques. Section 3 III presents a comprehensive description of the architecture of our solution, along with information about the necessary configuration and calibration process. Section 4 IV presents the experimental results, which are divided into indoor and outdoor experiments, accompanied by a description of the dataset used and the methodology employed. Finally, section 5 V discusses the implications of our findings and identifies potential areas for future improvement.

II. Related works

In this chapter, we provide an overview of the existing research and advancements in the field of computer vision, with a special focus on object detection, homography-based

mapping, monocular depth estimation, and absolute height estimation techniques.

A. Object detection

The task of object detection is widely used in computer vision and has a wide range of applications [4]–[6]. Currently, the best object detectors are based on convolutional neural networks (CNN). The initial success of CNN-based object detectors came with two-stage detectors like the region-based convolutional neural network (R-CNN) proposed by Girshick et al. [7] which has shown remarkable performance. This led to further advancements such as Fast R-CNN [8] and Faster R-CNN [9] - improved two-stage detectors, faster and with better accuracy. Another approach that gained popularity are one-stage detectors, exemplified by groundbreaking architectures like You Only Look Once (YOLO) [10] and Single-Shot Detector (SSD) [11]. They are faster, part of them can even work in real-time on edge devices, and currently have comparable accuracy to two-stage detectors [12].

The leading one-stage detection architecture YOLO has undergone significant improvements over time. Namely, its improved versions YOLOv2 [13] and YOLOv3 [14] introduced deeper convolutional backends, residual skip connections, residual blocks, and upsampling, resulting in one of the fastest object detection techniques while maintaining respectable accuracy. Bochkovskiy et al. presented YOLOv4 [15], which brought further enhancements to the training process, including data augmentation methods like CutMix, regularization techniques such as DropBlock, and architectural changes like the CSPDarknet53 backend network and path aggregation network with spatial attention blocks.

More recently, Jocher et al. presented YOLOv5 [1], which refreshed the YOLO architecture and improved its performance. The YOLO-based architecture remains a state-of-the-art object detector, with subsequent versions continually being developed and published under different names.

These advancements in object detection have significantly improved the accuracy and speed of detecting objects in various applications.

B. Homography-based mapping

Homography-based mapping is a widely used technique in computer vision that establishes correspondences between points in different images or views. It relies on the concept of a homography, which is a projective transformation that maps points from one plane to another. This mapping has numerous applications, including image stitching, augmented reality, camera calibration, and object tracking.

Works by Hartley and Zisserman [16], as well as by Cyganek and Siebert [17] provide a comprehensive overview of homography estimation algorithms. Additionally, the work of Szeliski [18] presents techniques for the robust estimation of homographies in the presence of outliers and noise. These studies serve as foundational knowledge for our use of homography-based mapping in height estimation.

C. Monocular Depth Estimation

Monocular depth estimation aims to recover depth information from a single image. This task is challenging due to the inherent ambiguity in monocular vision. Nevertheless, it plays a crucial role in various applications such as 3D reconstruction, scene understanding, and autonomous navigation.

Over the years, significant progress has been made in monocular depth estimation techniques. Early approaches were focused on hand-crafted features, superpixelation, and traditional computer vision algorithms [19]–[21]. However, with advancements in deep learning, convolutional neural networks have emerged as powerful tools for monocular depth estimation.

One notable work in this field is the pioneering study by Eigen et al. [22] where they introduced a CNN-based approach for monocular depth prediction. This work paved the way for subsequent research in deep learning-based depth estimation. Another significant contribution is the work of Laina et al. [23], who proposed a faster and lighter solution by training a fully convolutional residual network based on ResNet-50 [24]. They replaced the fully connected layers with up-convolutional blocks and modified the loss function.

Subsequently, the development of CNN-based solutions accelerated, leading to the creation of numerous works addressing this area. A few noteworthy contributions deserving special attention are listed below. Lee et al. [25] proposed a solution based on the relative depths between objects in the image. Ranftl et al. [26] presented a tool for mixing multiple datasets during monocular depth estimation training, even when their annotations were incompatible. This tool has facilitated future advancements in this field. Additionally, Ranftl et al. [2] proposed a dense vision transformer-based depth estimation architecture with a transformer backbone. Their architecture produces more fine-grained and globally coherent predictions compared to fully-convolutional architectures.

D. Absolute height estimation

Absolute height estimation is an intriguing computer vision task, but less popular than those mentioned above. To address this problem, various approaches based on image depth estimation [27], convolutional neural networks [27]–[29], and convolutional-deconvolutional deep neural networks (CDNNs) [30] have been proposed.

A notable work in this domain is the study conducted by Yin et al. [27], where they developed a four-stage estimator based on multiple CNN networks operating on a single-depth image. Their approach achieved impressive accuracy in height estimation, reaching as high as 99.1%. It is worth mentioning that their solution was limited to a controlled laboratory environment, where measurements were conducted on individuals positioned approximately 2 meters away from the camera. Despite this limitation, the achieved result is truly remarkable.

The field of absolute height estimation is relatively specialized, and fewer studies have been conducted compared to

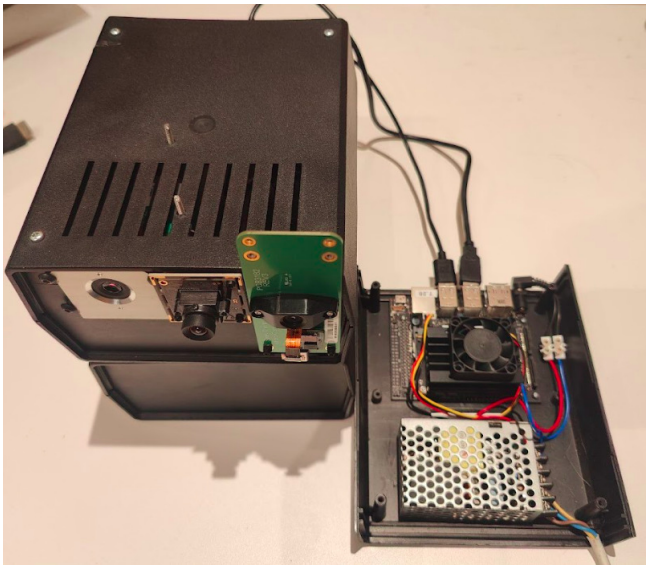


Figure 1: Arabox device - version with normal and termovision cameras and Jetson Nano board.

other computer vision tasks. Therefore, we believe that it is an interesting area for further research and development.

III. System Architecture

In this chapter, we present the architecture of our method for estimating the absolute height and distance of objects, which is implemented on an edge device called Arabox III-A, developed by MyLED [31]. Our approach for height estimation relies on two sources of data: a video signal and a thermal image signal. The system flow consists of integrating object detection using the YOLOv5 neural network, monocular depth estimation transformer, homography-based mapping, and polynomial regression-based estimations. We provide a detailed explanation of our method in section III-B.

A. Arabox device

Arabox, shown in Figure 1, is a device developed for fully anonymous data acquisition in the retail industry. It can be used in both stationary stores (including those operating in the omnichannel model) and the outdoor advertising industry (particularly in digitized form, known as DOOH). The device's key component is an embedded system that includes a GPU, such as the Jetson Nano, which is responsible for encrypting and processing data from the connected cameras. Arabox also includes a carrier board, power supply, fans, and a special case. Arabox has many use cases, but below we will focus on its height estimation functionality.

B. Height estimation pipeline

The main contribution of our paper is the absolute height estimation pipeline presented in Figure 2. It utilizes two data sources: a video signal and a thermal image signal. Both signals are initially processed by the YOLOv5 object detector before being fed into separate flows. In the first flow, we project the video signal onto a bird's-eye view using homography-based mapping. This enables us to estimate the spatial position of objects, as well as their distance from the camera and height (based on initial configuration, polynomial regressions, and the height of bounding boxes estimated by YOLOv5). We perform the same process for the thermal image signal, but with a different homography matrix.

The second flow is only performed for the video signal and is based on DPT Levit 224 monocular depth estimation model. This network estimates the relative depth of the image, and its output is combined with YOLOv5 detections to calculate the average depth value for the detected bounding boxes. Using a polynomial regression model from the device configuration and the relative depths of the detected objects, we can estimate their absolute distance from the camera, as well as their absolute height, in the same way as in the first flow. Finally, we average the results from all three flows to estimate the final height of the objects. More details on each of the pipeline steps are provided in the following subsections.

1) YOLOv5 detections: This part of our pipeline comprises two YOLOv5 models, which have been trained on two datasets that we prepared - one based on visual and the other on thermal images. These datasets contain approximately 20,000 annotated photos captured in urban environments. The YOLOv5 models output class of an object (such as human, car, or bus), its anchor location represented by two coordinates, and the height and width of the bounding box. All expressed in local coordinates associated with an image plane. This information is then used in the subsequent steps of our system, i.e. in the homography-based mapping and monocular depth estimation.

2) Homography-based mapping: The goal of this step is to project the location of the detected object from a 3D photo to a 2D "map" presented from a bird's-eye view. This projection enables us to estimate the distance and height of the object in the next step. To accomplish this, we first calculate the homography matrix for a given location during the device calibration process III-C. Subsequently, we use this matrix to transform the YOLOv5 detections and project them onto a 2D plane. By knowing their positions on the 2D plane and the scale of the plane saved during the configuration process, we can accurately estimate their distance from the camera, as well as their height, using polynomial regression.

3) Monocular depth estimation: In this step, we utilize the monocular depth estimation neural network called DPT Levit 224 [2], [3]. Given an input image, the network returns a map of relative depth estimates, where lower pixel values correspond to objects that are further from the camera. To improve the accuracy of our estimations, we first crop the image to remove any visible sidewall or casing fragments before passing it to the neural network.

The DPT Levit 224 model we use was trained using a publicly available tool for mixing monocular depth estimation datasets [26]. By using this pre-trained model, we can estimate the relative depth of objects in the scene with high accuracy, even in cases where the objects are partially occluded or have

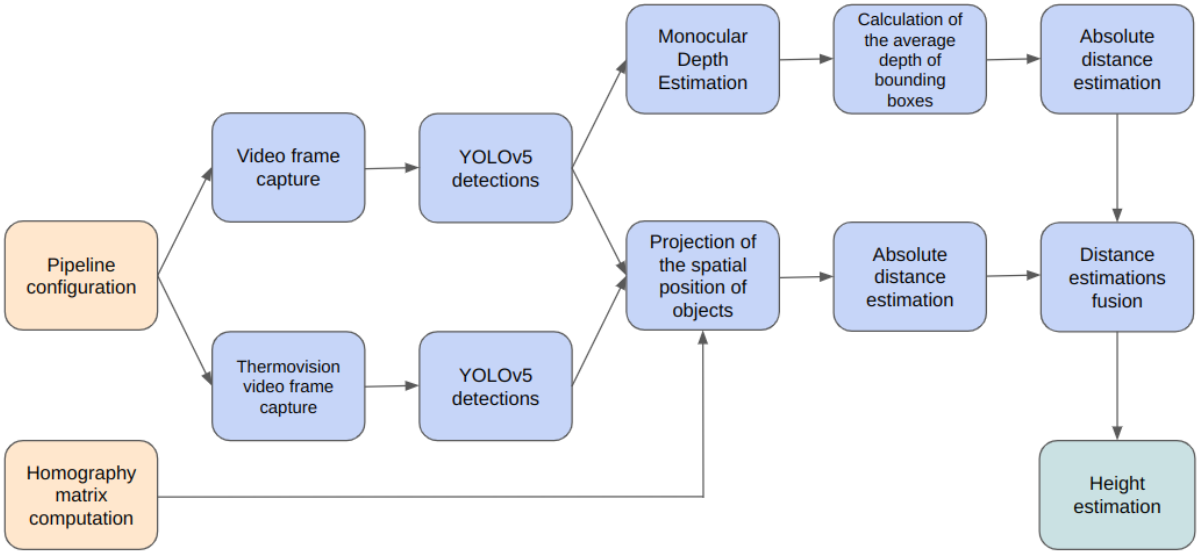


Figure 2: Block diagram of our height estimation system.

complex geometries.

4) Objects' absolute distance estimation: The estimation of object distance from the camera is performed using two methods, depending on the results of the previous step. If we estimate the distance based on the spatial position of the object obtained through homography-based mapping, the calculation is straightforward. We simply multiply the object's distance from the camera (expressed in pixels) by the scale factor included in the device configuration III-C.

On the other hand, if we estimate distance using monocular depth estimation, the process is more complex. In this case, we first calculate the average depth value for each bounding box returned by YOLOv5, and then substitute these values into a polynomial regression formula contained in the device configuration. This formula expresses the relationship between depth values and distance at a given location, enabling us to estimate absolute distance accurately. Details on how to calculate the coefficients of said polynomials are contained in the configuration section III-C.

5) Objects' absolute height estimation: After we have obtained the absolute distance of the object from the camera, we can estimate its height. However, for the estimate to be accurate, we need to calibrate the device to a specific location beforehand and calculate the coefficients of the 3rd-order polynomial accurately. This polynomial regression formula is used to determine the relationship between the distance of the object from the camera and the ratio of its height in pixels to the height in the real world. The process of calibration and calculating these coefficients is further explained in the configuration section III-C.

Once we have the coefficients and the distance value, we substitute them into the polynomial formula to obtain a height ratio. We then multiply this ratio by the height of the object in pixels obtained from the YOLOv5 detector. This calculation allows us to obtain an accurate estimate of the absolute height of the object.

Figure 3: A photo showing the process of calculating the homography matrix. The person responsible for the configurations marks the points on the image from the camera and the corresponding 2D map.

C. Configuration and calibration process

To ensure the proper functioning of the methods described, it is necessary to configure and calibrate the system. The most important parameters that we need to configure for each location where the device is to be used are: homography matrices, coefficients of the third-order polynomials used in the polynomial regression of distance and height, and the scale factor of pixels to meters.

Homography matrices are calibrated for a specific location using a simple script that requires marking several points on the original image and the 2D image, as shown in Figure 3. If the points are marked accurately, the script will return a homography matrix that allows for the projection of objects' locations from the camera's perspective onto a bird's-eye view. A separate matrix should be calculated for each camera (i.e. visual and thermal), since their parameters are different.

The polynomial regression is employed to determine the relationship between the distance of an object from the camera and the ratio of its height in pixels to its height in the real world. To compute the third-order polynomial coefficients used

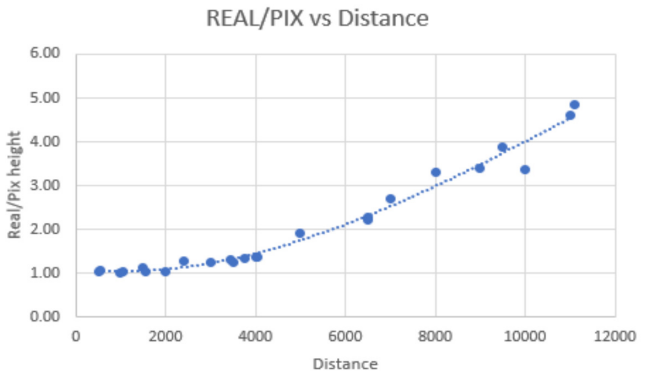


Figure 4: Regression curve calculated in the calibration phase for mapping the height of objects in pixels to the height of objects in meters.

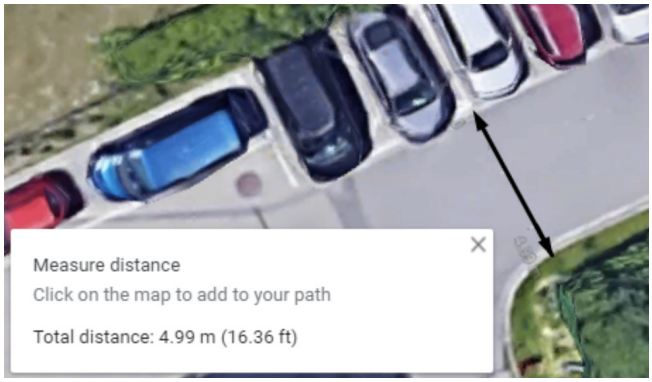


Figure 5: An example of distance measurement for device calibration using Google Maps.

for this regression, the following steps are proposed: Firstly, several YOLOv5 detections of a person with a known height should be made at different distances from the camera, and the height returned by YOLOv5 in pixels should be recorded. Next, a plot similar to the one shown in Figure 4 should be created, and a third-order polynomial regression should be computed on it. The obtained coefficients should then be saved in the device configuration. The height estimation module will then multiply the distance of the object from the camera by these coefficients and then by the height of the YOLOv5 prediction. This will provide an estimate of the height of the given object. A similar process should also be performed for the monocular depth estimation module and its depth-to-distance regression.

The final parameter needed to calibrate the device to a specific location is the scale that determines how many centimeters in the real world correspond to one pixel on the 2D map. This parameter can be easily calculated by measuring the distance between two characteristic objects on a Google Maps and then checking how many pixels on our 2D map they correspond to, as shown in Figure 5. Once all the parameters have been calibrated and configured, the device is ready for use in estimating the absolute height and distance of objects for chosen location. In the future we want to improve and automate the configuration process.

IV. Experimental part

To validate the effectiveness of our method, we conducted an experiment using a small dataset comprising videos of 11 individuals with known heights. The videos were captured in two distinct locations: one in an open environment and the other inside the building. By utilizing this dataset, we evaluated the performance of our system following the methodology outlined in section IV-B and achieved an estimation accuracy of 97.06%.

The experiment aimed to assess the system's ability to accurately estimate the height of individuals in different environmental conditions and validate the effectiveness of our proposed approach. In the following sections, we will discuss the details of the gathered dataset, our methodology and we will present the results obtained from our evaluation.

A. Dataset

The dataset comprises 10 recordings, each featuring a different individual with a known height. The recordings were captured using two types of cameras: a regular vision camera (model ELP-USB500W05G-FD100) and a thermal imaging camera (model SEEK Thermal MS202SP Micro Core).

The dataset includes videos from two distinct locations: indoors, specifically in an office space, with a total of three recordings, and outdoors, in a parking lot, with a total of seven recordings. The individuals participating in the recordings had heights ranging from 160cm to 185cm. While the dataset may not be extensive, we believe it provides sufficient variety to validate and confirm the effectiveness of our absolute growth estimation method. Sample frames from videos used in our dataset are presented in Figure 6.

B. Methodology

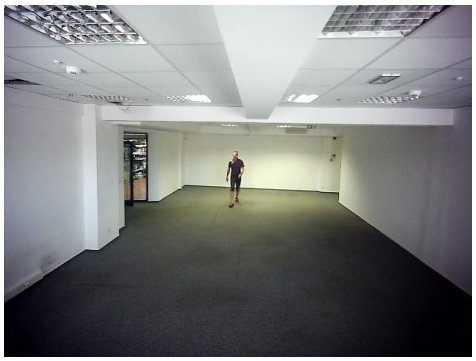
To validate the performance of our method for estimating the absolute height of individuals, we employed the following methodology.

For each video in our dataset, our model conducted height estimations on every frame in which the YOLOv5 detection model detected a person. The estimated heights were stored in a temporary table, and the measurements were averaged at the end of the video, using the Formula 1. Where, h_a represents the averaged height measurement result, h_i represents the result from a single frame, and N is the number of frames in which the person was measured.

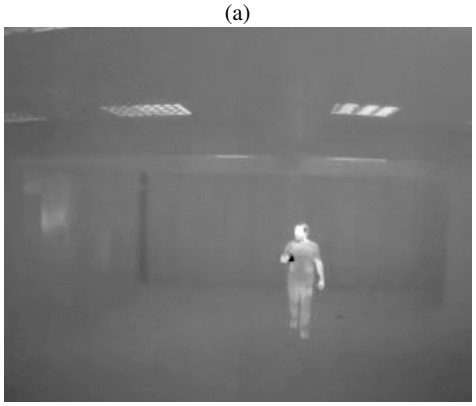
$$h_a = \frac{\sum_{i=1}^{N} h_i}{N} \tag{1}$$

These estimates were then compared against the known actual heights of the individuals (h_e) to calculate the percentage errors using Formula 2.

$$\delta = \left|\frac{h_a - h_e}{h_e}\right| * 100\% \tag{2}$$

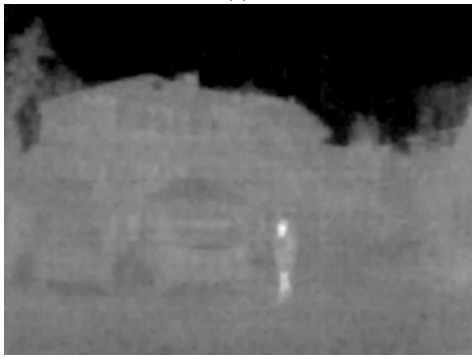


(a)



(b)

(c)



(d)

Figure 6: Sample images from the dataset

To provide a comprehensive evaluation, we stored all the results, as well as results from every module of our system in Table I and Table II, respectively. These tables serve as a consolidated record of the estimated heights, actual heights, and corresponding absolute errors for each video. Additionally, they contain the estimated heights from each component of the pipeline, namely results from the homography-based mapping using the vision data (HBM vision), homography-based mapping using the thermal data (HBM thermo), monocular depth estimation (MDE), and the fusion module. The fusion module results are calculated as presented in formula 3.

$$Fusion = \frac{\frac{HBE_{vision}+HBE_{thermovision}}{2} + MDE}{2} \quad (3)$$

As can be observed, the fusion formula is not a simple arithmetic average, as the module based on monocular depth estimation carries the greatest weight. This is because the homography-based mapping modules provide similar information, whereas the monocular depth estimation module offers distinct and additional insights. By utilizing this methodology, we can quantitatively assess the accuracy and reliability of our height estimation method across the entire dataset. The percentage error values obtained will allow us to analyze the performance of our system and identify areas for improvement.

In the subsequent sections, we will present the detailed results obtained from our evaluation and discuss the implications of these findings for the effectiveness of our proposed method.

C. Results

The results of our experiments are presented in three separate tables. Table I displays the measurements conducted indoors for three individuals, while Table II showcases the measurements carried out in an open area for eight individuals. Finally, Table III provides a weighted average summary of the results obtained from all experiments.

The average accuracy achieved in each experiment is as follows: 97.73% for Experiment 1, 96.77% for Experiment 2, with an overall weighted average accuracy of 97.06%. These accuracy percentages represent the degree of agreement between the estimated heights and the actual heights of the individuals. A more detailed description of the experiments and their results is provided below.

1) Experiment 1 - indoor area: In the first experiment conducted indoors, specifically in an office space; we recorded the heights of three individuals ranging from 173 cm to 186 cm; the maximum distance from the camera in which they could walk was around 12 meters. The system performed around 370 measurements for each person and then averaged them to obtain the final results presented in Table I. The average accuracy of the absolute height estimations obtained in this experiment was 97.73%. The best-performing module is based on homography mapping with a signal from the video camera with an error of only 1.14%. On the other hand, the worst-performing module is also homography-based mapping, but with a signal from the thermal camera - with a percentage

Table I: Indoor experiment results

	HBM vision	HBM thermo	MDE	Fusion	Ground Truth	Number of frames
Person 1	185cm	180cm	188cm	185cm	186cm	346
Error 1	0.54%	3.23%	1.08%	0.54%	-	346
Person 2	179cm	167cm	170cm	172cm	178cm	350
Error 2	0.56%	6.18%	4.49%	3.37%	-	350
Person 3	177cm	167cm	164cm	168cm	173cm	412
Error 3	2.31%	3.47%	5.20%	2.89%	-	412
Avg. Error	1.14%	4.29%	3.59%	2.27%	-	-

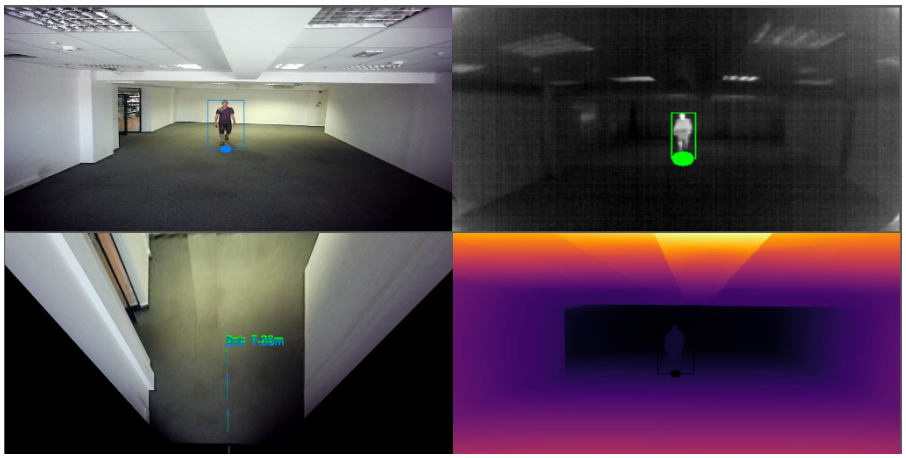


Figure 7: Visualization of the described method: Upper left - detection on a video signal, upper right - detection on a thermal image, bottom left - projection of a person's detections into bird's eye view, bottom right - monocular depth estimation output

error equal to 4.29%. A sample visualization of our method's work on data from the indoor experiment was presented in Figure 7.

2) Experiment 2 - outdoor area: In the second experiment, we conducted measurements in an open area, specifically a small parking lot. Seven individuals with heights ranging from 160 cm to 185 cm participated in this experiment; the maximum distance from the camera in which they could walk was around 20 meters. For each person, a varying number of measurements, ranging from 865 to 1968, were conducted and averaged. Final results are presented in Table II. The average accuracy of the estimations obtained in the second experiment was 96.77%. The best-performing module in this experiment was the monocular depth estimation model, with an average percentage error equal to 3.03%, whereas the worst-performing method was once again thermovision homography mapping with an error equal to 4.39%.

3) Results summary: Summarizing the results of the aforementioned experiments, we achieved a weighted average accuracy of 97.06%. Among the different modules used, the height estimation module based on homography projecting yielded the highest accuracy of 97.35%. The other modules, namely the monocular depth estimation module and homography-based mapping working on thermal imaging, achieved slightly lower accuracies of 95.89% and 95.64%, respectively.

Looking for reasons for such results, the lower accuracy of the model working on thermal imaging data can be attributed to the less accurate detections of the YOLOv5 on the thermal images. The thermal images dataset, on which the YOLOv5 model was trained, was smaller than the traditional dataset, which can correspond to weaker results. Notably, the detections from the thermal-based model were often 10-15% higher in the vertical axis, which was not observed in normal data.

Regarding monocular depth estimation, certain challenges were encountered due to the background conditions. For instance, if a person passed in front of a car, the model believed that person to be closer than if they were at the same distance but there was no car in the background. Despite this limitation, the results achieved in this experiment were considered very good, taking into account the difficulty of the scenery.

Experiment no. 2 presented slightly weaker results due to the more complex scene and higher maximal distance in which individuals could walk. Particularly, beyond 15 meters, the system encountered significant challenges in accurately estimating the distances and therefore absolute heights.

Moving forward, we aim to expand our dataset and conduct experiments in a larger number of testing locations with a more diverse group of individuals. This will further validate and enhance the proposed method. Additionally, we will focus on improving other aspects of our method, which will be discussed in detail in the following section.

V. Conclusion and Future Works

Presented in this paper a method for absolute distance and height estimation that incorporates a combination of visual and

Table II: Outdoor experiment results

	HBM vision	HBM thermo	MDE	Fusion	Ground Truth	Number of frames
Person 4	187cm	188cm	185cm	186cm	185cm	865
Error 4	1.08%	1.62%	0%	5.40%	-	865
Person 5	171cm	187cm	169cm	174cm	179cm	1044
Error 5	4.47%	4.47%	5.56%	2.79%	-	1044
Person 6	173cm	189cm	179cm	180cm	174cm	937
Error 6	0.57%	7.94%	2.87%	3.45%	-	937
Person 7	167cm	180cm	172cm	173cm	170cm	1255
Error 7	1.76%	5.88%	1.18%	1.76%	-	1255
Person 8	157cm	179cm	171cm	170cm	168cm	1968
Error 8	6.55%	6.55%	1.79%	1.19%	-	1968
Person 9	172cm	171cm	173cm	172cm	167cm	1080
Error 9	2.99%	2.40%	3.59%	2.99%	-	1080
Person 10	151cm	157cm	150cm	152cm	160cm	1015
Error 10	5.63%	1.88%	6.25%	5.00%	-	1015
Avg. Error	3.29%	4.39%	3.03%	3.23%	-	-

Table III: Results summary

	HBM vision	HBM fusion	MDE	Fusion
Avg. Error	2.65%	4.36%	4.11%	2.94%

thermal imaging data, and which employs advanced technologies such as object detection, homography-based mapping, and monocular depth estimation, constitutes a significant scientific contribution to the field of spacial position estimation in real conditions.

With an accuracy of 97.06%, our method demonstrates promising results, making it suitable for applications on edge devices. However, we acknowledge that there is room for improvements. In our future endeavors, we aim to enhance the accuracy of our method and streamline the configuration and calibration processes.

Moving on to future works, one of our primary objectives is to expand our dataset by incorporating additional locations and involving a more diverse range of participants. This expansion would provide valuable insights into the performance of different modules of our height estimation method and their effectiveness in various environmental conditions. By evaluating our method on a more diverse dataset, we can identify areas for improvement and optimize its performance accordingly.

Another improvement of the proposed method will be streamlining the configuration and calibration process. At the moment, it takes an experienced operator about 30 minutes to configure the device for a new location. We would like to streamline this process and automate it further, especially the part related to the calculation of the polynomial regression coefficients of the distance and height estimation modules.

Additionally, we plan to extend our method with new modules. These could include methods such as monocular depth estimation based on thermal imaging, a human pose estimation [32] module, and the utilization of the object segmentation [33] methods for obtaining more accurate data for calculating the average depth of objects with monocular depth estimation module. By incorporating these new modules, we aim to enhance the capabilities and versatility of our method in estimating absolute distance and height.

In conclusion, our article shows that the accurate estimation of absolute distance and height from the monocular view is possible with high accuracy by using a hybrid solution based on object detection, homography-based mapping, and monocular depth estimation. Furthermore, we recognize the potential for further development and propose future improvements in this task.

ACKNOWLEDGMENTS

This work was supported by the National Centre for Research and Development, Poland, under the grant no. POIR.01.01.01-00-1116/20.

REFERENCES

[1] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, TaoXie, J. Fang, imyhxy, K. Michael, Lorna, A. V, D. Montes, J. Nadar, Laughing, tkianai, yxNONG, P. Skalski, Z. Wang, A. Hogan, C. Fati, L. Mammana, AlexWang1900, D. Patel, D. Yiwei, F. You, J. Hajek, L. Diaconu, and M. T. Minh, "ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference," Feb. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6222936

[2] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," 2021.

[3] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze, "Levit: a vision transformer in convnet's clothing for faster inference," *CoRR*, vol. abs/2104.01136, 2021. [Online]. Available: https://arxiv.org/abs/2104.01136

[4] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. N. Asghar, and B. Lee, "A survey of modern deep learning based object detection models," *CoRR*, vol. abs/2104.11892, 2021. [Online]. Available: https://arxiv.org/abs/2104.11892

[5] J. Gąsienica-Józkowy, M. Knapik, and B. Cyganek, "An ensemble deep learning method with optimized weights for drone-based water rescue and surveillance," *Integrated Computer-Aided Engineering*, vol. 28, pp. 221–235, 2021, 3.

[6] M. Knapik and B. Cyganek, "Driver's fatigue recognition based on yawn detection in thermal images," *Neurocomputing*, vol. 338, pp. 274–292, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219302280

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2014.

[8] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2016.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016.
[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 21–37. [Online]. Available: https://doi.org/10.1007%2F978-3-319-46448-0_2
[12] M. Knapik and B. Cyganek, "Fast eyes detection in thermal images," *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 3601–3621, Jan 2021. [Online]. Available: https://doi.org/10.1007/s11042-020-09403-6
[13] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517–6525, 2017.
[14] ——, "Yolov3: An incremental improvement," *ArXiv*, vol. abs/1804.02767, 2018.
[15] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020.
[16] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. New York, NY, USA: Cambridge University Press, 2003.
[17] B. Cyganek and J. Siebert, "An introduction to 3d computer vision techniques and algorithms," pp. 459–474, 01 2009.
[18] R. Szeliski, "Image alignment and stitching: A tutorial," *Found. Trends. Comput. Graph. Vis.*, vol. 2, no. 1, p. 1–104, jan 2006. [Online]. Available: https://doi.org/10.1561/0600000009
[19] J. Michels, A. Saxena, and A. Y. Ng, "High speed obstacle avoidance using monocular vision and reinforcement learning," in *Proceedings of the 22nd International Conference on Machine Learning*, ser. ICML '05. New York, NY, USA: Association for Computing Machinery, 2005, p. 593–600. [Online]. Available: https://doi.org/10.1145/1102351.1102426
[20] A. Saxena, S. Chung, and A. Ng, "Learning depth from single monocular images," *Advances in neural information processing systems*, vol. 18, 2005.
[21] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic photo pop-up," *ACM Trans. Graph.*, vol. 24, pp. 577–584, 2005.
[22] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," 2014.
[23] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," 2016.
[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.
[25] J.-H. Lee and C.-S. Kim, "Single-image depth estimation using relative depths," *Journal of Visual Communication and Image Representation*, vol. 84, p. 103459, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320322000190
[26] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," 2020.
[27] F. Yin and S. Zhou, "Accurate estimation of body height from a single depth image via a four-stage developing network," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8264–8273.
[28] D.-s. Lee, J.-s. Kim, S. C. Jeong, and S.-k. Kwon, "Human height estimation by color deep learning and depth 3d conversion," *Applied Sciences*, vol. 10, no. 16, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/16/5531
[29] P. Alphonse and K. Sriharsha, "Depth estimation from a single rgb image using target foreground and background scene variations," *Computers & Electrical Engineering*, vol. 94, p. 107349, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790621003207
[30] L. Mou and X. X. Zhu, "Im2height: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," 2018.
[31] M. sp. z o.o., "Myled sp. z o.o." 2021, accessed on 05-22-2023. [Online]. Available: https://myled.pl/
[32] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," 2022.
[33] A. M. Hafiz and G. M. Bhat, "A survey on instance segmentation: state of the art," *International Journal of Multimedia Information Retrieval*, vol. 9, no. 3, pp. 171–189, jul 2020. [Online]. Available: https://doi.org/10.1007%2Fs13735-020-00195-x