Fachhochschule Aachen

Campus Jülich

Faculty of

Medical Engineering and Technomathematics

Degree Program: Applied Mathematics and Computer Science B.Sc.

FH AACHEN
UNIVERSITY OF APPLIED SCIENCES

Deep learning-based step length detection using an RGB-D camera

Seminar Thesis

by

Oskar Beaujean

First Examiner:	Prof. Dr. rer. nat. Stephan Bialonski
Second Examiner:	M. Sc. Nico Blaß
Matriculation Number:	3633683

Aachen, December 26, 2025

Affidavit

I hereby attest that I have written this seminar paper on the topic

Deep learning-based step length detection using an RGB-D camera

independently, using only the sources and aids listed, and that I properly cited all direct and indirect quotations. This paper has not been submitted to any other examination authority or published. In the course of writing this paper, the AI systems “Gemini 3”, “Grok 4” and "DeepL" were used to support linguistic revision, translate and to reflect on and clarify independently developed arguments. Neither content-related solutions nor texts produced by AI were used. I am solely responsible for the independent development of all technical statements, assessments, and conclusions. The use was in accordance with the intended purpose of the system and in compliance with data protection and copyright regulations.

Aachen, den December 26, 2025

Oskar Beaujean

Abstract

This seminar thesis explores the possibility of accurate step length estimation by combining stereo vision and AI pose detection. A software pipeline was developed using a Luxonis OAK-D Pro camera for depth estimation, the deep neural network model YOLO11n-pose for keypoint detection and a multi-step filtering process for noise reduction. Experimental validation using step intervals of 50,60 and 70 cm revealed high absolute Errors (>100 mm) with Peak detection, while a Kernel Density Estimation achieves accuracies of approximately 10 mm. These findings demonstrate potential for an autonomous, contactless pace regulation system.

Contents

1 Introduction

Intelligent control systems are essential in an increasingly automated world. An aging demographic requires an enhanced healthcare system, where physical movement is one of the most important factors influencing health. As machine learning continues to maximize technological improvements, assistive walking systems can become 'smart' through the integration of advanced control systems. An electric walker can support a person via pace regulation, assisting with everything from rehabilitation exercises to fall prevention. With the development of faster and more accurate AI detection systems, new autonomous control mechanisms are becoming a reality. A contactless regulation system, combining AI detection with 3D estimation through stereo vision, could transform a standard electric walker into a smart and safe support device.

1.1 Stereo Depth Estimation

The ability to estimate the distance of a certain object without physical measurement is found naturally in the animal kingdom. Biological systems possess innate capabilities to gauge the distance of visible objects through stereopsis. In humans, incoming reflections of light are captured by two lenses situated at a fixed distance: our eyes. The brain estimates the distance by analyzing the disparity created between the two slightly different reconstructed images.

Technological replication of this process began in the 19th century, when Sir Charles Wheatstone laid down the foundations of stereoscopy. He invented the stereoscope, a device demonstrating binocular depth vision, by presenting two pictures taken from two different perspectives to both eyes separately.

Calculating disparity requires identifying the same point in both stereoscopic images. Without a clear marker, such as a light or laser dots commonly used in active stereo, a system becomes challenged by the possibility of multiple potential candidates. With the emergence of Computer Vision in the 20th century, this so-called *Correspondence Problem* moved into the focus of research. Marr and Poggio described in their 1976 article an early algorithm to solve this exact problem [1]. With multiple iterations, the right points are singled out, strengthening candidates which have surrounding points with similar disparity. Once the point pair is identified, the disparity value can be calculated (Figure 1). This, in turn, enables the possibility to create a 3D estimation from the captured image by deriving a depth value for each pixel taken [2].

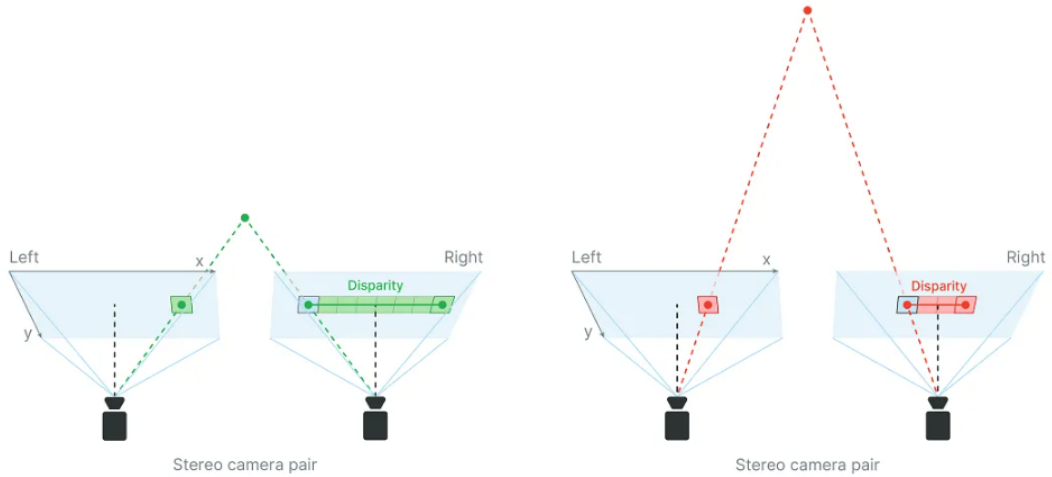
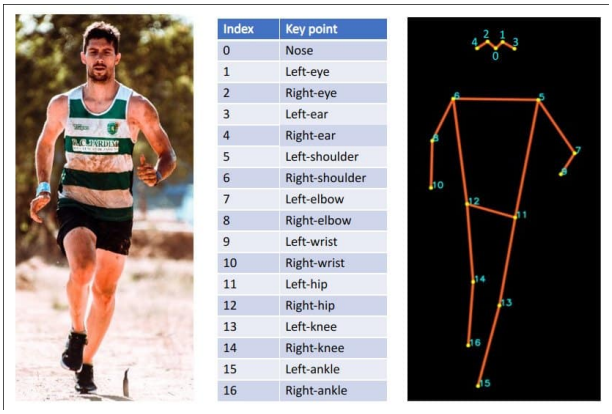


Figure 1: Stereo Camera Disparity Calculation [11]

1.2 Key Point Detection

From autonomous vehicles to surveillance, convolutional neural networks brought a significant advancement to modern computer vision. Object detection systems such as Ultralytics YOLO (You Only Look Once) [3, 4] enable real-time object detection. Introduced in 2016, YOLO reframes the process to a single regression problem, with a convolutional network predicting multiple bounding boxes with class probabilities simultaneously, reducing the computation needed to a minimum [5].

Building upon this, Deep Neural Networks for pose estimation are designed to identify anatomically important human keypoints. Combined, these points form a simplified skeletal representation, marking the most important parts of the human body [6]. Integrating both technologies, YOLO-Pose emerges as an enhanced non-heatmap bottom-up method [7]. YOLO-Pose Models are trained on the COCO Dataset [8] utilizing 200,000 images with 250,000 humans, for the detection of 17 distinct keypoints [7] (Figure 2).



Index	Key point
0	Nose
1	Left-eye
2	Right-eye
3	Left-ear
4	Right-ear
5	Left-shoulder
6	Right-shoulder
7	Left-elbow
8	Right-elbow
9	Left-wrist
10	Right-wrist
11	Left-hip
12	Right-hip
13	Left-knee
14	Right-knee
15	Left-ankle
16	Right-ankle

Figure 2: Keypoints YOLOv8-Pose [20]

2 Methodology

2.1 Hardware Specifications

The experimental apparatus consisted of an OAK-D Pro Camera that was mounted to the underside of a walker at a 10° angle facing downward. The OAK-D Pro has one RGB camera in the middle, two infrared cameras on each side, an infrared Dot Projector, and an incorporated Processor [9]. The depth perception capabilities are enhanced with the infrared dot projector providing "more texture to the scene" [11], enabling "Active Stereo". The built-in processor has a potential of 4 TOPS processing power, of which 1.4 TOPS can be used for on-board AI Calculation [12].

Figure 3: Electric Walker

Figure 4: OAK-D Pro [9]

2.2 Software Pipeline & Depth Estimation

For the control of the camera, the Luxonis DepthAI library was used [14]. It features a simple connection API and a node-based structure that can be linked into a pipeline. To achieve the depth estimation, the software was constructed in Python. A pipeline object, the RGB camera node, the two mono camera nodes, and a stereo node were initialized. The outputs of both mono camera nodes were linked to the stereo node, enabling the computation of the disparity.

The data from the calculated disparity are transformed to depth values through the Triangulation equation:

$$z = \frac{f \cdot b}{d} \tag{1}$$

where f is the focal length, b is the baseline distance between the cameras (calculated internally inside the camera), and d is the disparity [2]. The resulting frames from the RGB camera alongside the 2D depth data array were saved for later usage.

2.3 Keypoint Detection (YOLO-Pose)

To estimate the step length of the participant, each RGB frame is given to the YOLO-Pose model to detect the human pose keypoints. Of particular interest were the 15th and 16th keypoints, located at the left and right ankle. These two points are best suited for step distance estimation. The model outputs, in case of a human detection, an array where the keypoints are stored. If the 15th and 16th keypoints are detected, the X and Y pixels from each point are saved alongside the confidence score.

2.4 Depth Extraction

In the next part, the detection data is loaded, and the keypoint pixels are used as indices to retrieve the depth value from the corresponding data array. To ensure more reliable results, the noise of the depth map must be minimized. To achieve that, a "Region of Interest" (ROI) around the given coordinates was utilized rather than a single data point [15].

For the ROI size, a 9×9 area was chosen as a starting ground. If less than one-third of the values are valid, the region slowly grows iteratively up to a maximum size of 20×20. This process reduces the possibility of too few valid values while minimizing the risk of including values from outside the desired region. During the depth estimation process, inside the OAK-D, the disparity of each pixel has a certain confidence score; if it did not reach a certain threshold, the disparity value was set to zero. From the array, the median value is taken, to reduce the noise interference even further.

2.5 Step Length Calculation

Before the step length can be calculated, the 2D coordinates need to be transformed to 3D coordinates. The camera intrinsic horizontal and vertical focal lengths (f_x, f_y) and principal point (c_x, c_y) are extracted beforehand through a calibration node. With the depth value z, the 3D x and y coordinates can be calculated with the triangulation equations [5]. For a 2D point (u,v) with $c : camera$ and $w : world$:

$$x_c = \frac{(u - c_x) \cdot z_c}{f_x}, \quad y_c = \frac{(v - c_y) \cdot z_c}{f_y} \tag{2}$$

Because the camera is situated at a slight angle of $\theta = 10°$, the y and z coordinates need to be adjusted:

$$y_w = y_c \cdot \cos(\theta) - z_c \cdot \sin(\theta), \quad z_w = y_c \cdot \sin(\theta) + z_c \cdot \cos(\theta) \tag{3}$$

All coordinates are saved, and the absolute difference between both z values is taken as the step length.

2.6 Data Filtering

For a refined step length estimation, certain filter methods were implemented into the software in a multi-step approach.

- **Biomechanical Threshold:** A maximum step length of 1000 mm was set. The average step length of a human is calculated with a factor of 0.43 from their height; this threshold would only exclude humans averaging around 230 cm. This makes subsequent filter implementations cleaner.
- **Hampel Filter:** Implemented for strict outlier identification and replacement. A strict statistical threshold is calculated using the Median Absolute Deviation (MAD). The MAD value is multiplied by 1.4826 to make it comparable with standard deviation. A threshold factor of 3 was chosen (keeping 99.87% of original data points), filtering out data points further away than the threshold and replacing them with the median. This is used iteratively.
- **Savitzky-Golay Filter:** This transforms the data with polynomial regression in a moving window. It leads to a stricter outlier correction while generally preserving curvature, peaks, and valleys.

3 Experiment

The experiments were conducted in a windowless corridor with constant artificial illumination. The camera's infrared floodlight and dot projector were set to maximum intensity to enable active stereo. Both the RGB and Mono cameras captured data at 20 frames per second. For the detection the YOLO11n-Pose was chosen. To validate the experiment a ground truth measurement was needed. Three walkways were created, each had 11 parallel tape markings on the ground.

- Walkway 1: 50 cm distance.
- Walkway 2: 60 cm distance.
- Walkway 3: 70 cm distance.

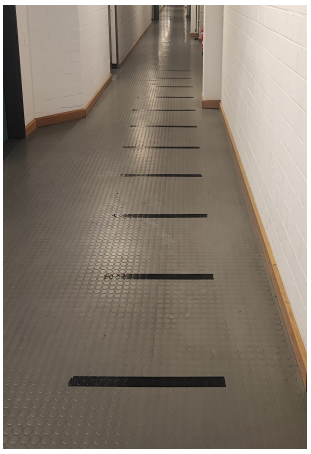


Figure 5: Walkway with 70 cm distanced tape marker

With each step the participant needed to touch the 48 mm wide tape with their heel, otherwise the measurement was invalid and was repeated. In total each walkway was completed 10 times, with each iteration capturing 10 step length measurements when both heels touched the tape. With each step a maximum error of 48 mm could be made, but the average step length would be the measured distance between the tapes. This was an attempt to establish a limited ground truth while also creating a realistic usage of the walker. Due to the quite difficult task of placing the foot inside the 48 mm space a slower, more deliberate movement was observed. After each step a short pause was intended so more frames could be captured, resulting in more data points for later examination.

4 Results

4.1 Keypoint Detection Accuracy

The foundation of the step length estimation is reliable detection results. The YOLOn-pose model delivers a fast and accurate keypoint estimation. Although the exact ankle location varied, the model predicted the correct area with over 70% confidence in nearly all 15,000 captured frames. In total an average of 90.7 % confidence is given to the right ankle keypoint and 89.4 % for the left.

4.2 Signal Processing and Noise Reduction

First a representative plot from the processed data will be examined and the multi step filter approach described in the methodology was applied. Figure 6 shows one of the ten 50 cm measurements, plotting the step length per image frame.

Depth map errors can cause relatively high peaks, which can be seen in Figure 6. Even with ROI value extraction noise is a common occurrence. This is largely caused by larger noisy areas, rather than only a few of the pixels. First the data points are filtered using a maximum possible distance threshold of 1,000 mm (Figure 7).

After thresholding, the graph is displays a clearer rhythmic pattern. The ten steps create approximately twelve local minima separated by a larger accumulation of points in the 400 and 600 mm range. The larger number of points is due to slower movement while placing the foot and the short stop to capture the maximum step length. Numerous outliers are visible, ranging from short dips to high spikes near the threshold limit.

To combat the measurement noises the Hampel Filter is applied iteratively up to maximum of 10 repetitions. In the first iteration many outliers are removed and replaced (Figure 8).

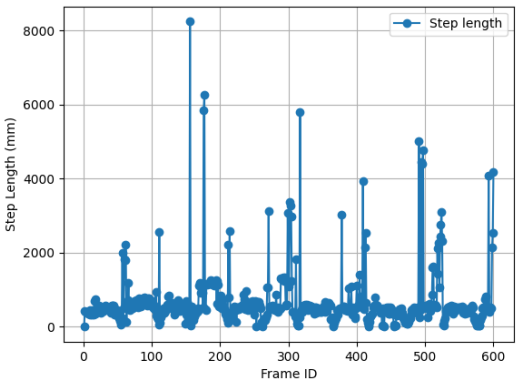


Figure 6: Step length per Image Frame

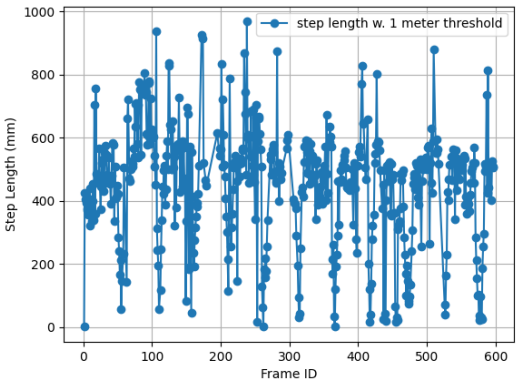


Figure 7: Step length with 1 meter threshold

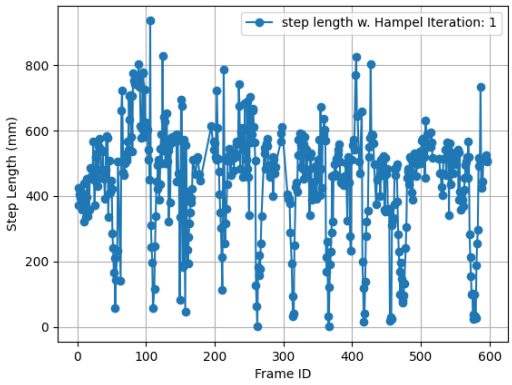


Figure 8: Hampel Filter applied. First Iteration

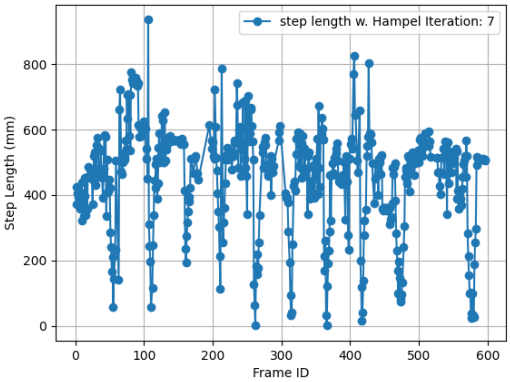


Figure 9: Hampel Filter applied. 7th Iteration

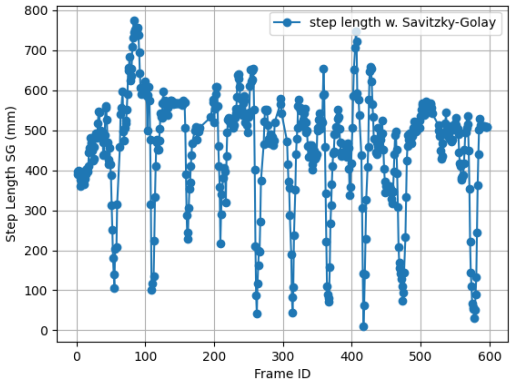


Figure 10: Savitzky Golay Filter applied. Polynomial order 2, window size 7

Nevertheless, several clusters reaching up to 800 mm are unaffected. This is normal due to the overall contribution of the clusters to the calculation of the median. After seven iterations no additional outliers can be detected (Figure 9). One can observe that correct data is also removed. The local minima around frame 530 has been cut, leading to complications later during peak estimation. It becomes clear that a modest approach for the selection of the sigma threshold is most effective.

The final method applied is the Savitzky Golay Filter, using a window size of seven and polynomial order of two all data points of the graph are transformed (Figure 10). The influence of certain outliers is minimized, while other spikes are still too dominant to be removed. Furthermore, many fluctuations in the data are removed and smoothed.

4.3 Step Length Estimation Validation

4.3.1 Analysis of Peak Detection

Finally, peak detection is performed. Except that the first and last each local maximum represent one of the ten steps done during the measurement (Figure 11). The detected peaks range from 572.4 to 773.4 mm, averaging around 649.4 mm. The prominence and distance parameters control the peak detection process. Because the second to last minima was cut out during the filter process only 9 peaks are returned. Nevertheless, the identification of the individual steps is mostly successful. measurements. There is a large error when comparing the Peak values and

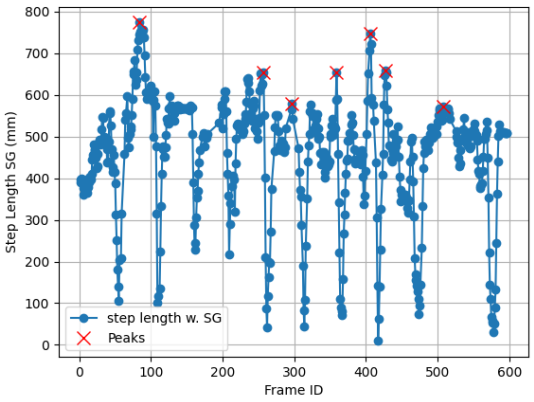


Figure 11: Peak Detection: Distance 1, Prominence 350

the ground truth of 50 cm. This is a common occurrence with all other filtered measurements (Table 1).

Table 1: Comparison of Ground Truth with Step length estimation through Peak detection

Ground Truth	Steps recorded	Peaks Detected	Median Peak	Absolute Error
500 mm (avg.)	100	89	619.0 mm	119.0 mm
600 mm (avg.)	100	102	727.8 mm	127.8 mm
700 mm (avg.)	100	106	806.6 mm	106.6 mm

The absolute error between ground truth and Median Peak detected ranges between 106.6 to 127.8 mm, showing a clear inability to estimate the step length through peak detection.

However, due to the short pause in between each step, a large number of data points are collected. Here, the peak detection takes the maximum values of each step, but all other points close in the surrounding region should be analyzed too.

4.3.2 Analysis of Kernel Density Estimation

Examining the same representative measurement as before, it Is common to observe data cluster touching or extending below the 500 mm mark. Actually most of the data points gather around this mark. A different representation of the data can be achieved utilizing the mode calculation as seen in Figure 12.

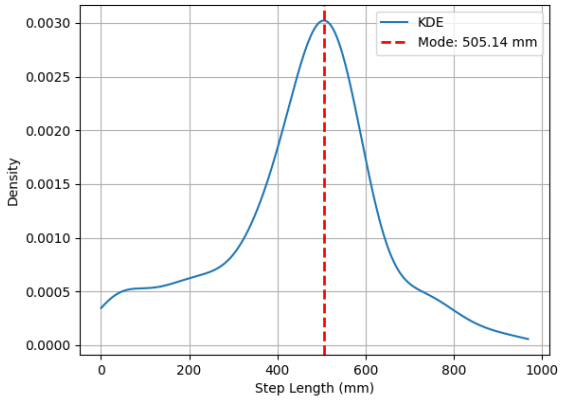


Figure 12: Kernel Density Estimation with Mode (only thresholding)

Utilizing the Kernel Density Estimation an underlying value of 505.1 mm is being measured. Extending the analysis from a single measurement to include the other nine measurements provides a comprehensive overview. Here the calculated mode is even closer to the established Ground Truth (Table 2).

Table 2: Comparison of Ground Truth vs. Mode

Ground Truth	**Data points recorded**	**Mode**	**Absolute Error**
500 mm (avg.)	5105	504.7 mm	4.7 mm
600 mm (avg.)	4957	579.6 mm	21.4 mm
700 mm (avg.)	5434	695.7 mm	4.3 mm

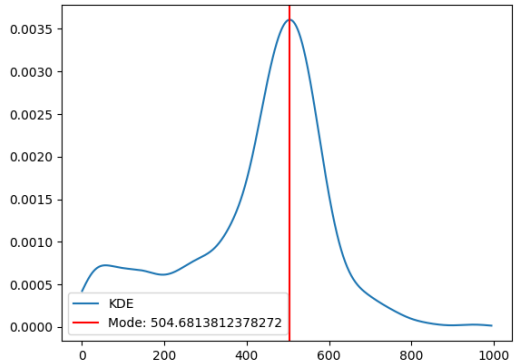


Figure 13: KDE and Mode for all 50 cm measurements

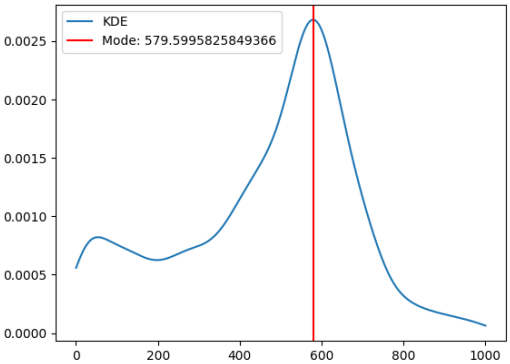


Figure 14: KDE and Mode for all 60 cm measurements

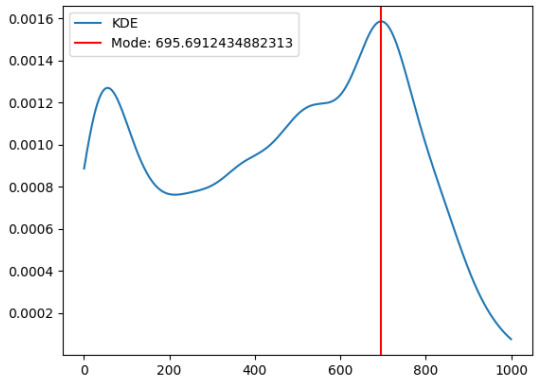


Figure 15: KDE and Mode for all 70 cm measurements

5 Discussion

It becomes clear that the depth value estimation has a strong variance; only through a broader perspective can values similar to the ground truth be observed. This shift away from a qualitative to a more quantitative approach could be used for long-term step length estimations but leads to possible irregularities in measuring sudden changes. The influence of illumination and object variety is also a strong factor in the quality of the depth estimation [19]. The hallway used for the experiments (Figure: 5) featured suboptimal lightning and low object variety, reducing the depth estimation capabilities. Another factor is the position of the camera, the "ideal range" for the OAK-D Pro is between 80 cm to 12 m [9]. With the current mounting configuration, some foot placements are too close for optimal disparity estimation. There is also a limitation because of blind spots, where one leg obstructs the other making it impossible for the mono camera to estimate a disparity value. The ground truth used is too vague, a different method for a more accurate measurements would have established a stronger foundation for later validation.

6 Outlook

To achieve more accurate step length calculation through stereo depth estimation, new methods in filtering and refining need to be included and tested. Other possibilities include, the usage of deep neural networks to improve the quality of depth maps [17, 18]. Implementation in real-time on hardware like the Luxonis OAK Cameras would be a significant achievement for the development of an autonomous system. Furthermore experimenting with different camera positions

References

[1] D. Marr and T. Poggio, "Cooperative Computation of Stereo Disparity," 1976.

[2] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press.

[3] Ultralytics YOLO. https://web.archive.org/web/20251226120223/https://github.com/ultralytics/ultralytics

[4] Ultralytics Website https://web.archive.org/web/20251226120318/https://docs.ultralytics.com/de/#the-evolution-of-object-detection

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *arXiv*, 2016. http://arxiv.org/abs/1506.02640

[6] A. Toshev and C. Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks," in

[7] D. Maji, S. Nagori, M. Mathew, and D. Poddar, "YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss," *arXiv*, 2022. http://arxiv.org/abs/2204.06806

[8] T.Y. Lin et al., "Microsoft COCO: Common Objects in Context," *arXiv*, 2015.

[9] Luxonis OAK-D Pro. Available: https://web.archive.org/web/20251226115340/https://shop.luxonis.com/products/oak-d-pro1

[10] Luxonis Stereo Depth Nodes. https://web.archive.org/web/20251226120354/https://docs.luxonis.com/software/depthai-components/nodes/stereo_depth/#StereoDepth-Disparity

[11] Luxonis Documentation. Configuring Stereo Depth. Available: https://web.archive.org/web/20251226120025/https://docs.luxonis.com/hardware/platform/depth/configuring-stereo-depth#fixing-noisy-depth

[12] Luxonis Documentation. RVC2 NN Performance. Available: https://web.archive.org/web/20251226120058/https://docs.luxonis.com/hardware/platform/rvc/rvc2/#RVC2%20NN%20Performance

[13] Luxonis Documentation. AI Inference Conversion. https://web.archive.org/web/20251226120157/https://docs.luxonis.com/software/ai-inference/conversion

2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[14] Luxonis DepthAI https://web.archive.org/web/20251226180939/https://github.com/luxonis/depthai

[15] Luxonis Spatial location calculator. https://web.archive.org/web/20251226120404/https://docs.luxonis.com/software-v3/depthai/examples/spatial_location_calculator/spatial_location_calculator/

[16] I. Bytyçi and M.Y. Henein, "Stride Length Predicts Adverse Clinical Events in Older Adults: A Systematic Review and Meta-Analysis," *JCM*, vol. 10, no. 12, p. 2670, 2021.

[17] A. Kendall et al., "End-to-End Learning of Geometry and Context for Deep Stereo Regression," *arXiv*, 2017. http://arxiv.org/abs/1703.04309

[18] K. Shankar et al., "A Learned Stereo Depth System for Robotic Manipulation in Homes," *arXiv*, 2021.

[19] Luxonis dyanmic calibrationhttps://web.archive.org/web/20251226141842/https://docs.luxonis.com/software-v3/depthai/depthai-components/host_nodes/dynamic_calibration/

[20] Keypoints illustration https://web.archive.org/web/20250812152731/https://learnopencv.com/wp-content/uploads/2021/05/fix-overlay-issue.jpg