Replacing Motors with Pixels: A Computer Vision-Based System for Hand Rehabilitation

C. A. Brocx (5835925), S. Bourjila (5803438), M. Van Bavel (5798493), and J. S. Zeegelaar (5699517)
Delft University of Technology, Faculty of ME, Mekelweg 2, 2628 CD Delft

This paper presents two computer vision systems for tracking hand motion in a rehabilitation device: one using deep learning (YOLOv11n-pose) and another based on colour segmentation. Both estimate 3D positions using a PnP framework and were evaluated for accuracy, speed, and robustness. The colour-based tracker achieved detection rates between 20.89 and 27.97Hz, approaching the 10Hz target while maintaining low error. In contrast, the YOLO system performed significantly worse (mean absolute error of 126.0±219.7mm for distance, 18.2±25.1° for tilt, and 4.76±0.93Hz), due to model noise and limited optimisation. While the colour tracker offers a more promising, low-cost solution, future work should assess its stability under extreme lighting, where the YOLO model may prove more reliable.

This paper presents two vision-based systems for tracking hand motion in a rehabilitation device: a deep learning model (YOLOv11n-pose) and a colour segmentation tracker. Both estimate 3D positions using a PnP framework and were evaluated for accuracy, speed, and robustness. The colour tracker achieved 20.89–27.97 Hz detection with low error, nearing the 10Hz target. YOLO underperformed (126.0±219.7mm distance error, 18.2±25.1° tilt, 4.76±0.93Hz) due to model noise and limited optimisation. While the colour tracker offers a low-cost alternative, its stability under extreme lighting needs further testing, where the YOLO model may excel.

I. Introduction

Stroke remains a leading cause of long-term disability, with one in four individuals expected to experience a stroke during their lifetime [1]. Among stroke survivors, upper limb motor impairment is both common and persistent, often severely limiting daily activities [2, 3]. Hand function, in particular, plays a critical role in quality of life, independence, and reintegration into daily routines [4].

Various technologies have been developed to support upper limb rehabilitation. Robotic systems like exoskeletons and end-effectors offer intensive, task specific training but are costly, complex, and typically restricted to clinical use [5]. Wearable sensors and data gloves provide accurate motion tracking but are often bulky and uncomfortable for extended use [6]. Vision-based systems have emerged as a non-invasive, affordable alternative, using RGB-D cameras to track hand movement in interactive tasks [7]

Rätz et al. developed a home-based rehabilitation device featuring a 3D-printed compliant shell actuated by a DC motor, integrated into a gamified Unity-based environment for motor and somatosensory recovery [8–10]. While promising, the system is limited by its high cost (€1000) and mechanical complexity, which reduce its accessibility and ease of use for unsupervised home training.

This paper proposes a vision-based alternative that eliminates the need for the motor. The proposed Computer Vision Tracking (CVT) system estimates the hand's interaction with the compliant shell using a single RGB camera, substantially lowering cost, complexity and allowing the rehabilitation task to be monitored passively. We implement two CVT pipelines: one based on key point detection using deep learning (DL), and one based on conventional colour segmentation and detection. Both systems are evaluated independently to determine whether vision-based tracking can replace the motor-based system, whilst maintaining sufficient accuracy, speed,

and robustness.

First, we describe the design of both systems and the experimental setup used to evaluate them (Section II). This is followed by the results (Section III), a discussion of their implications (Section IV), and our conclusions (Section V). This work was conducted as part of the TU Delft Final Bachelor Project in Mechanical Engineering.

II. Methodology

A. System Design

1. Design Goals

To ensure practical value of our CVT the following design goals have been set:

- **Accuracy**: Computer-vision systems in clinical hand training report average spatial errors of 3.88 ± 3.86 mm in 3D fingertip distances [11], and automated vision-based hand measurements for post-stroke patients show mean finger-length differences of 4.5 mm from callipers [12]. Since even small improvements in grip can be clinically significant during rehabilitation [13], we target a mean absolute error of 4.5 mm distance estimation. For tilt (pronated/supinated), this corresponds to a 2.25 mm positional error at both the top and bottom of the shell. Given their separation, this results in an angular deviation of approximately 2.6°, which we use as the reference threshold for acceptable tilt error.
- **Latency**: Human perception treats up to 32 ms delay as instantaneous, but delays above 100 ms degrade user experience significantly [14, 15]. We therefore aim for an end-to-end system latency lower than 100 ms, which results in a detection frequency higher than 10 Hz.
- **Robustness**: Given that the intention of the device is to be used in an unconstrained environment (at home), it is essential that the designed system can perform under various conditions, such as over- and underexposure of the camera, and non contrasting background. We define robustness to be violated when a single factor has a statistically significant effect ($p < 0.05$) with $\eta_p^2 \geq 0.06$, indicating at least a moderate practical effect [16, 17].

To guide our choice of detection approach, we first compared traditional, two-stage, and one-stage methods (see Appendix A). One-stage deep learning detectors stood out for their balance of speed, accuracy, and robustness on benchmark datasets [18], making them ideal for real-time applications. Within this category, YOLO and SSD were particularly promising due to their consistent performance and extensive documentation available online[19–21]. YOLO was ultimately selected for its overall superior performance. Particularly the latest YOLOv11n model, the smallest and fastest version to date based on benchmarks [22], though not yet accompanied by a formal publication. (see Appendix A for a comparison with SSD).

In parallel, initial experiments using OpenCV's Python library [23] revealed that colour based detection offered competitive accuracy and processing speed, with the added benefit of ease of implementation, since there is no need to train a DL model. For this reason, a second tracker based on colour segmentation was developed alongside the YOLO based approach.

2. 3D-Pose estimation from 2D image points

Before designing the full detection systems, a mathematical framework was developed to determine the 3D distance between the two shell ends, regardless of the alignment of the shell relative to the camera. The framework operates under the key assumption that an accurate 3D model of both Shell Ends is known. For each shell end, we define n points of interest in the model space $\mathcal{M}$:

$$\mathbf{p}_i^{\mathcal{M}} = \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix}, \quad i \in \{1, 2, \ldots, n\}, \quad n \in \mathbb{N}. \tag{1}$$

From the detection pipeline (whether via machine learning or colour detection), we obtain the 2D image coordinates of these points of interest:

$$\mathbf{p}_i^{\mathcal{I}} = \begin{bmatrix} u_i \\ v_i \end{bmatrix}. \tag{2}$$

In practice, these measurements include error, so the observed image points are represented as $\bar{\mathbf{p}}_i^{\mathcal{I}}$. To solve the pose estimation problem, we apply Point-n-Perspective (PnP) [24], which maps the 2D image

points to the 3D model points using:

$$s \begin{bmatrix} \mathbf{p}_i^{\mathcal{I}} \\ 1 \end{bmatrix} = \mathbf{K} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} & \mathbf{R} & & \mathbf{T} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{p}_i^{\mathcal{M}} \\ 1 \end{bmatrix}, \tag{3}$$

where:

- s is the scaling factor converting pixels to millimetres,
- $\mathbf{K}$ is the camera intrinsic matrix,
- $\mathbf{R}$ is the rotation matrix in $\mathbb{R}^3$,
- $\mathbf{T}$ is the translation vector in $\mathbb{R}^3$.

With $n \geq 5$ points and a known intrinsic matrix $\mathbf{K}$, the system becomes solvable for $\mathbf{R}$ and $\mathbf{T}$. Importantly, assuming Gaussian noise, increasing the number of points n makes the system more redundant and reduces the solution error. The matrix $\mathbf{K}$ can be determined via online calibration at the start of the program, using Zhangs method [25]. Applying this method to both Shell Ends, gives the ability to map the location of each point in camera space:

$$\mathbf{p}_{i,j}^{C} = \mathbf{R}_j \mathbf{p}_{i,j}^{\mathcal{M}} + \mathbf{T}_j, \quad j \in \{1, 2\} \tag{4}$$

To determine the distance between them simple Euclidean distance in C space will be applied to the centre points of the shell ends. Conveniently, we have determined that these centre points are also the respective origins for each Shell End in their respective $\mathcal{M}$ space , hence the distance between them is simplified to:

$$D = |\mathbf{T}_1 - \mathbf{T}_2| \tag{5}$$

The tilt was estimated using the third Euler angle of each $\mathbf{R}$ matrix, and taking the mean.

Implementing the framework was done using OpenCV's Python library [23].

3. YOLO System Development

To ensure compatibility with the mathematical framework, which requires keypoint input, the YOLOv11n-pose model was chosen for its ability to predict 2D keypoints $\bar{\mathbf{p}}_i^{\mathcal{I}}$. The design process comprised three phases: (1) developing custom markers with distinct outlines to define trackable keypoints, (2) generating a synthetic dataset for training, and (3) training the YOLOv11n-pose model. Each phase is detailed below.

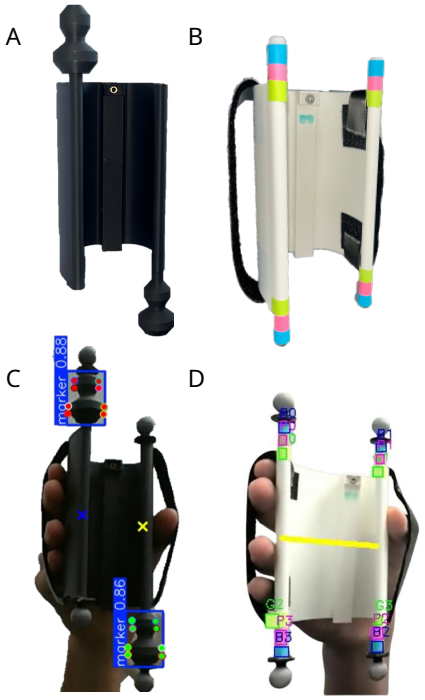

Figure 1. An overview of the two different shell for the two systems developed. A: the shell used for the YOLO system; B: the shell used for the colour tracking system; C: The YOLO System in action, on the Shell with the OptiTrack markers attached, bounding boxes (blue) key points (red and green), and the two midpoints of the shell ends (crosses) are visualised; D: The colour tracker in action with OptiTrack markers attached, each found colour has a bounding box (coloured square outlines) and a label (0-3), the yellow line is the 2D representation of the 3D distance between the shell ends.

a) Shell Design and Key Point Definition. Attachable markers, henceforth referred to as knobs, were developed for the shell ends. The design of the shape and size of these knobs was guided by conventional 3D printing manufacturing knowledge and logical reasoning, particularly in light of creating contrast using edges and corners. Figure 1 A, shows the design. The markers were kept cylindrically symmetric, allowing the key points to be at the outline of them.

b) Data Gathering. Obtaining large amounts of data for training deep learning models is often labor-intensive. To address this, synthetic data was generated using BlenderProc [26], a procedural Blender pipeline for photorealistic rendering and automatic annotation. A versatile custom dataset was created using a 3D model of the knob, with randomized oc-

clusions, lighting, and backgrounds. An overview of the randomization parameters and example images is provided in Appendix B. In total, 5,757 annotated images were generated and used for training.

c) Training the Model. With the dataset prepared and structured, a YOLOv11n-pose model was trained using the Ultralytics Python library [27]. To increase variability, the parameters `hsv_h`, `hsv_v`, `perspective`, and `shear` were set to 0.5, 0.75, 0.0001, and 5, respectively, introducing variation in hue, brightness, and minor geometric distortions. Additional augmentations, such as blurring and greyscaling, were applied using the Albumentations library [28]. All other training parameters were kept at default. The model was trained on 640 x 640 images for 300 epochs. Training details are provided in Appendix B.

4. Colour Tracker Development

As for the Colour Tracker, implementation was done using OpenCV's Library, with the resultant pipeline explained below. Firstly, three distinct colours were chosen: blue, green, and pink (see figure 1B), as these bright colours, especially in combination, are rarely encountered in everyday life. For the detection of these colours, the RGB image is converted to HSV space, as this space is more robust [29]. A binary mask is created for each colour range to isolate each colour. Then, Noise is eliminated from the maps, using morphological opening with a 5x5 kernel as is standard [30], followed by the contours being filtered on aspect ratios to further reduce false positives. Then contours were grouped based on proximity, at most 50 pixels centroid distance, and if at least 2 out of 3 colours were present, this increases robustness with respect to occlusion. Finlay, a bounding box is created around each validated colour, which are at the top and bottom of each shell end. The corner of these boxes function as the input for the 2D to 3D pipeline.

B. The Experiment

1. Materials

OptiTrack System. To capture the motion of the shell ends an OptiTrack system was employed. In total 6 cameras were used, 5 to capture the motion, and 1 set to greyscale to see the clap, used for time calibration between the OptiTrack and either the YOLO System or the Colour Tracking. All cameras where of the Prime^x variant, having a 1.3 MP resolution, an accuracy of ±0.20 mm, and a frame rate of 180 FPS. On top of each shell end OptiTrack Motion Capture Markers were placed as can be seen in figure 1 C and D.

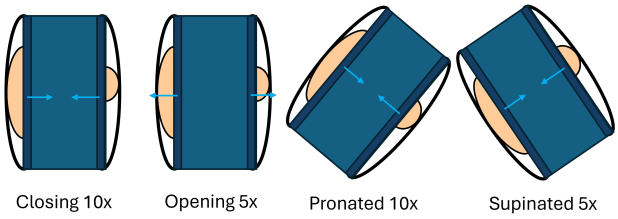

Figure 2. The exercises performed during the experiment in chronological order.

Laptop. A Dell XPS 15 9560 was used to run both the YOLO System and the Colour Tracker on. The laptop has the following specifications: 16 GB ram, Intel CPU i7-7700, Nvidia Geforce GTX 1050 (4 GB) GPU, Windows 10 OS.

IPhone 13. An IPhone 13 with ios 18.5 and the Light Meter LM-3000 application version 1.9.5 was used to measure the lux in the room in conjunction with a self made diffuser (a stroke of paper wrapped around the top of the phone, over the selfie camera). This leads to a uncertainty of ±12% and a cosine error of up to 10%, as per the application's own information menu.

Miscellanea. The 3D printed shell, colourful magnets and paper, whiteboard and black cloth.

2. Experimental Design

a)The Movement Sequence. As the shell's primary function is to mimic a grasping motion, the experiment was designed to reflect this use. Figure 2 illustrates the selected movements. First, the shell ends are brought together by closing the hand, this represents the core motion. Second, the ends are moved apart by opening the hand, extending the range of motion tested. Third and fourth, the closing motion is repeated with approximately 45 degree pronation and supination to assess system performance under varied orientations and evaluate angular tracking. The latter was performed only five times due to discomfort.

b) Environmental Variables. To assess robustness, two variables were examined: background and lighting.

Table 1 outlines their variations. Each combination of background and lighting was tested in three independent trials. The three backgrounds encompass both a simulated whitish or black wall, and a more detailed background with magnets and colours, as house generally have some extend of decoration on their walls. The backgrounds can bee seen in Appendix E.

Variable	Variation 1	Variation 2	Variation 3
Background	Plain whiteboard	Black cloth	Whiteboard with coloured paper and magnets.
Lighting	Daylight, lights off	Daylight, lights on	–

Table 1. Environmental variables and their variations, the movement sequence is performed 3 times independently for each combination of the variables.

c) Experimental Setup Figure 3 demonstrates the trial set up. The OptiTrack was calibrated to the excelent level at the beginning of each experiment session, i.e. the different days the experiment was performed for each system. Time calibration between the tested system and the OptiTrack was done using a clapping gesture visible for both systems.

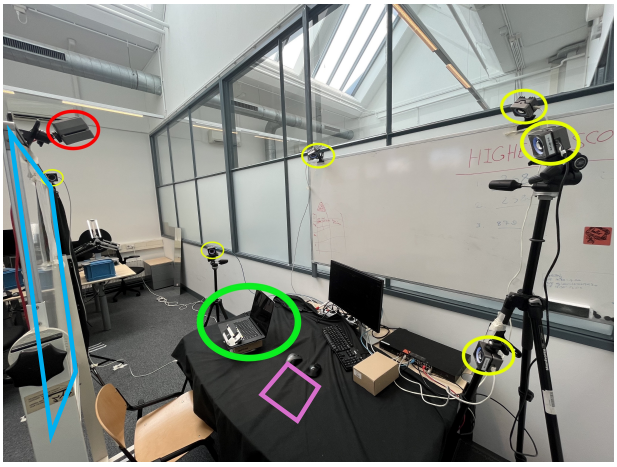

Figure 3. The Experimental Setup. 5 OptiTrack cameras for motion tracking (yellow), 1 greyscale camera for time calibration between OptiTrack and either the YOLO system or the Colour Tracker (red), the laptop which was running our system (green), the whiteboard used to create the various backgrounds (blue), and the location where the Iphone was placed for light measurements (purple).

3. Data Analysis

Raw OptiTrack data and pre-processed distance and angle data from both vision-based systems were obtained. To enable comparison, OptiTrack data were augmented to compute equivalent distances and angles by calculating shell midpoints from paired markers, using known real-world distances.

As OptiTrack recorded at 180 Hz and the vision-based systems at lower, inconsistent frame rates, all data were linearly interpolated and resampled at a fixed frequency. Prior to this, outliers were removed based on the z-score of absolute error ($|z| > 3$), reducing the impact of transient artefacts [31].

Due to occasional missing OptiTrack data in the YOLO-based system, likely from tracking or labelling errors, repetitions, defined as full motion cycles, were semi-automatically segmented using peak detection, followed by manual correction. Repetitions with over 20% missing OptiTrack data were excluded.

RMSE and absolute error (mean ± standard deviation) were calculated per movement and trial, and aggregated per environmental condition. Detection frequency was computed per trial as the number of frames divided by duration, and averaged per condition. To assess effects of Background, Lighting, and Movement, Welch's ANOVA was applied due to violations of standard ANOVA assumptions (see Appendix C).

III. Results

A. YOLO System Results

Tables 2 and 3 summarise the performance of the YOLOv11n-pose tracking system in terms of detection rate and estimated distance and tilt, under three backgrounds (Whiteboard, Black Cloth, Colourful) and two lighting conditions (Lights Off, Lights On). For each combination, the average sampling rate (Hz), the overall root mean squared error (RMSE), and the mean ± standard deviation of the absolute error are reported, both across the entire session and within each movement phase (Closing, Opening, Pronated, Supinated). Appendix D contains the same information split up per session, while Appendix E showcases photos of each experimental scenario.

Average detection rate was lowest with a Black Cloth background (3.91 Hz) compared to the White-

board (4.98 Hz) or Colourful (5.38 Hz), and the lower end of the overall distribution (4.76 ±0.93 Hz) , with minimal difference between the Lights-Off (4.72 Hz) and Lights-On (4.79 Hz) conditions. A two-way Welch ANOVA confirmed the background to be the primary influencing factor, $F(2, 9.32) = 5.09$, $p = 0.032$, $\eta_p^2 = 0.475$, and Lighting to have no significant effect, $F(1, 15.74) = 0.025$, $p = 0.877$, $\eta_p^2 = 0.0015$.

The Overall RMSE and mean AE error ranged from 92.9 mm and 60.6mm (Colourful, lights Off) to 530.9 mm and 308.7 mm (Colourful, lights On), with an overall AE of 126.0±219.7 mm. A three way Welch ANOVA showed that all factors: Background, $F(2, 3320) = 104.58$, $p \ll 0.001$, $\eta_p^2 = 0.0499$; Lighting, $F(1, 3497) = 299.66$, $p \ll 0.001$, $\eta_p^2 = 0.0480$; and Movement, $F(3, 1500) = 17.96$, $p \ll 0.001$, $\eta_p^2 = 0.0071$, had a small but significant effect.

As for the tilt, the RMSE varied between 26.5° (Whiteboard, lights On) and 36.2° (Colourful, lights On) and an overall AE of 18.2±25.1 mm. The three-way Welch ANOVA revealed a strong effect of Movement Phase, $F(3, 3315) = 439.02$, $p \ll 0.001$, $\eta_p^2 = 0.1629$, a small effect of Background, $F(2, 5068) = 26.65$, $p \ll 0.001$, $\eta_p^2 = 0.0069$, and no effect of Lighting, $F(1, 7633) = 0.04$, $p = 0.839$, $\eta_p^2 = 0.0000054$. All Welch ANOVA tables are presented in Appendix C.

B. Colour Tracking

Tables 4 and 5 summarise the colour-based tracking system's performance in estimating the distance between shell ends and the shell's tilt, respectively. The system was evaluated under the same background and lighting conditions as the machine learning pipeline.

Unlike the YOLO-based system, all traditional tracking trials yielded complete datasets. Performance was evaluated for each type of movement: closing, opening, pronation, and supination of the shell. This structure is reflected in the separate error for each of these movements.

The average detection rate was lowest for the Black Cloth background (23.05 Hz), compared to the Whiteboard (26.8 Hz) and Colourful (24.67 Hz). Differences between lighting conditions, Lights-Off (25.82 Hz)and Lights-On (23.86 Hz), were minor. A two-way Welch ANOVA confirmed that neither Background, $F(2, 7.84) = 3.66$, $p = 0.0754$, $\eta_p^2 = 0.311$, nor Lighting, $F(1, 14.974) = 2.727$, $p = 0.119$, $\eta_p^2 = 0.151$, had a statistically significant effect.

The overall RMSE and mean absolute error ranged from 5.40 mm and 4.16 mm (Whiteboard, Lights On) to 12.61 mm and 7.17 mm (Whiteboard, Lights Off). A three-way Welch ANOVA showed a strong effect of Movement Phase, $F(3, 12564.60) = 806.68$, $p \ll 0.001$, $\eta_p^2 = 0.0845$, and small but significant effects of background, $F(2, 19142.46) = 42.04$, $p \ll 0.001$, $\eta_p^2 = 0.0028$, and lighting, $F(1, 21223.24) = 118.70$, $p \ll 0.001$, $\eta_p^2 = 0.0044$.

For tilt estimation, the RMSE ranged from 5.05° (Black Cloth, Lights Off) to 11.81° (Whiteboard, Lights On). A three-way Welch ANOVA revealed a strong effect of Movement Phase, $F(3, 13110.80) = 1137.13$, $p \ll 0.001$, $\eta_p^2 = 0.106$, a small effect of background, $F(2, 17080.67) = 72.48$, $p \ll 0.001$, $\eta_p^2 = 0.0044$, and no significant effect of lighting, $F(1, 29388.78) = 2.23$, $p = 0.135$, $\eta_p^2 = 0.0000662$. Al Welch ANOVA tables can be found in appendix, C.

IV. Discussion

A. Interpretation of Findings

1. Performance of the YOLO System

Ultimately, the YOLOv11n-pose tracking system does not meet the predefined accuracy and speed criteria of a maximum mean absolute error (AE) of 4.5 mm for distance and 2.6° for tilt, and a minimum detection frequency of 10 Hz. Instead, it shows an overall AE of 126.0±219.7 mm for distance, 18.2±25.1° for tilt, and an average detection frequency of 4.76±0.93 Hz. None of the tested environmental conditions satisfy the performance thresholds individually either: the best results were an AE of 60.6±70.4 mm (Colourful, Lights Off) for distance, 4.5±23.7° (Whiteboard, Lights On) for tilt, and a maximum detection rate of 5.45±0.12 Hz (Colourful, Lights On).

These shortcomings may partially be attributed to the high noise levels in the data (see Figure 4). This suggests that the YOLOv11n-pose model was not fully optimised, a conclusion supported by visual inspection during early deployment, particularly in scenarios where the OptiTrack markers were present. The model

Background	Lighting	Total			Closing	Opening	Pronated	Supinated
		Sampling rate [Hz] ($\mu \pm \sigma$)	RMSE [mm]	AE [mm] ($\mu \pm \sigma$)	AE [mm] ($\mu \pm \sigma$)	AE [mm] ($\mu \pm \sigma$)	AE [mm] ($\mu \pm \sigma$)	AE [mm] ($\mu \pm \sigma$)
Whiteboard	Off	5.15 ± 0.67	133.9	92.3 ± 97.1	139.7 ± 167.3	–	50.0 ± 54.1	135.8 ± 83.1
	On	4.81 ± 0.47	143.7	111.9 ± 90.3	86.1 ± 71.7	81.0 ± 49.3	96.9 ± 77.4	164.4 ± 107.1
Black Cloth	Off	3.70 ± 0.94	122.7	80.8 ± 92.4	85.3 ± 105.2	31.4 ± 60.9	80.3 ± 60.8	108.2 ± 120.5
	On	4.11 ± 1.27	125.9	86.4 ± 91.6	113.6 ± 110.4	93.2 ± 153.8	71.1 ± 62.7	83.2 ± 91.9
Colourful	Off	5.31 ± 0.68	92.9	60.6 ± 70.4	39.5 ± 52.9	26.5 ± 42.9	63.5 ± 73.5	84.1 ± 75.7
	On	5.45 ± 0.12	530.9	308.7 ± 432.2	307.5 ± 570.0	275.7 ± 441.3	296.1 ± 304.7	353.5 ± 401.8

Table 2. Performance of the YOLOv11n-pose tracking system for distance estimation under varying background (plain whiteboard, black cloth, colourful) and lighting (off/on) conditions. For each setup, the table reports the average detection rate (Sampling rate) in Hz, the overall root mean square error (RMSE) in mm, and absolute error (AE) in mm, as well as the absolute error of the estimated distance during the closing, opening, pronated, and supinated phases.

Background	Lighting	Total			Closing	Opening	Pronated	Supinated
		Sampling rate [Hz] ($\mu \pm \sigma$)	RMSE [°]	AE [°] ($\mu \pm \sigma$)	AE [°] ($\mu \pm \sigma$)	AE [°] ($\mu \pm \sigma$)	AE [°] ($\mu \pm \sigma$)	AE [°] ($\mu \pm \sigma$)
Whiteboard	Off	5.15 ± 0.67	33.9	20.4 ± 27.1	36.8 ± 33.8	23.7 ± 31.1	8.1 ± 3.1	6.9 ± 5.5
	On	4.81 ± 0.47	27.8	14.5 ± 23.7	31.3 ± 32.2	14.3 ± 24.5	4.8 ± 4.5	4.6 ± 1.9
Black Cloth	Off	3.70 ± 0.94	30.8	16.8 ± 25.9	32.6 ± 33.1	18.9 ± 30.4	6.3 ± 8.0	5.9 ± 5.9
	On	4.11 ±1.27	26.5	15.6 ± 21.4	15.7 ± 27.2	29.9 ± 32.2	12.5 ± 8.7	9.4 ± 7.1
Colourful	Off	5.31 ± 0.68	30.3	17.8 ± 24.6	34.8 ± 31.7	17.1 ± 24.1	7.5 ± 7.1	5.9 ± 5.8
	On	5.45 ± 0.12	36.2	24.6 ± 26.5	37.1 ± 31.3	24.2 ± 30.5	17.2 ± 17.6	15.9 ± 17.0

Table 3. Performance of the YOLOv11n-pose tracking system for tilt estimation under varying background (plain whiteboard, black cloth, colourful) and lighting (off/on) conditions. For each setup, the table reports the average detection rate (sampling rate) in Hz, the overall root mean square error (RMSE) in mm, and absolute error (AE) in mm, as well as the absolute error of the estimated tilt during the closing & opening and pronated & supinated phases, after outlier removal

Background	Lighting	Total			Closing	Opening	Pronated	Supinated
		Sampling rate [Hz] ($\mu \pm \sigma$)	RMSE [mm]	AE [mm] ($\mu \pm \sigma$)	AE [mm] ($\mu \pm \sigma$)	AE [mm] ($\mu \pm \sigma$)	AE [mm] ($\mu \pm \sigma$)	AE [mm] ($\mu \pm \sigma$)
Whiteboard	Off	27.97 ± 0.34	12.6	7.2 ± 10.4	4.2 ± 3.7	3.7 ± 2.7	9.8 ± 8.9	16.4 ± 19.2
	On	25.63 ± 1.48	5.4	4.2 ± 3.4	3.2 ± 2.7	3.2 ± 2.5	5.8 ± 4.1	3.7 ± 2.8
Black Cloth	Off	25.21 ± 0.70	7.4	5.2 ± 5.3	3.5 ± 3.7	5.4 ± 5.1	7.9 ± 6.6	5.0 ± 4.5
	On	20.89 ± 2.74	10.8	7.0 ± 8.2	5.0 ± 4.8	5.0 ± 3.4	11.0 ± 12.1	6.9 ± 4.9
Colourful	Off	24.27 ± 5.52	10.3	6.4 ± 8.0	3.5 ± 3.5	3.4 ± 3.7	12.6 ± 10.6	7.4 ± 9.0
	On	25.07 ± 3.21	7.5	4.9 ± 5.8	3.3 ± 4.2	3.8 ± 4.2	6.7 ± 7.4	5.8 ± 5.8

Table 4. Performance of the colour-based tracking system for distance estimation under varying background (whiteboard, black cloth, colourful) and lighting (off/on) conditions. For each setup, the table reports the average sampling rate in Hz, the root mean square error (RMSE) and absolute error (AE) in mm for the total trial, as well as during the closing & opening and pronated & supinated phases, after outlier removal.

frequently confused these with the shell knobs. However, due to limited time and computational resources, no extensive optimisation was carried out—neither on the synthetic training dataset, nor during training

Background	Lighting	Total			Closing	Opening	Pronated	Supinated
		Sampling rate [Hz] ($\mu \pm \sigma$)	RMSE [°]	AE [°] ($\mu \pm \sigma$)	AE [°] ($\mu \pm \sigma$)	AE [°] ($\mu \pm \sigma$)	AE [°] ($\mu \pm \sigma$)	AE [°] ($\mu \pm \sigma$)
Whiteboard	Off	27.97 ± 0.34	7.6	4.9 ± 5.8	5.1 ± 7.5	4.3 ± 2.2	5.1 ± 2.8	3.8 ± 3.2
	On	25.63 ± 1.48	11.8	5.6 ± 10.4	2.3 ± 1.2	2.8 ± 3.5	7.6 ± 13.1	9.6 ± 14.4
Black Cloth	Off	25.21 ± 0.70	5.0	4.4 ± 2.5	4.1 ± 1.1	3.2 ± 1.4	5.6 ± 3.1	3.9 ± 3.6
	On	20.89 ± 2.74	7.3	5.2 ± 5.1	2.9 ± 2.0	2.4 ± 1.8	10.3 ± 6.1	3.3 ± 2.0
Colourful	Off	24.27 ± 5.52	6.5	4.9 ± 4.3	2.5 ± 2.1	2.3 ± 2.8	9.3 ± 3.5	6.4 ± 4.5
	On	25.07 ± 3.21	5.1	4.0 ± 3.2	3.2 ± 2.1	2.0 ± 1.9	6.2 ± 3.1	4.0 ± 3.8

Table 5. Performance of the colour-based tracking system for tilt estimation under varying background (whiteboard, black cloth, colourful) and lighting (off/on) conditions. For each setup, the table reports the average sampling rate in Hz, the root mean square error (RMSE) and absolute error (AE) in degrees for the total trial, as well as during static and angled positions, after outlier removal

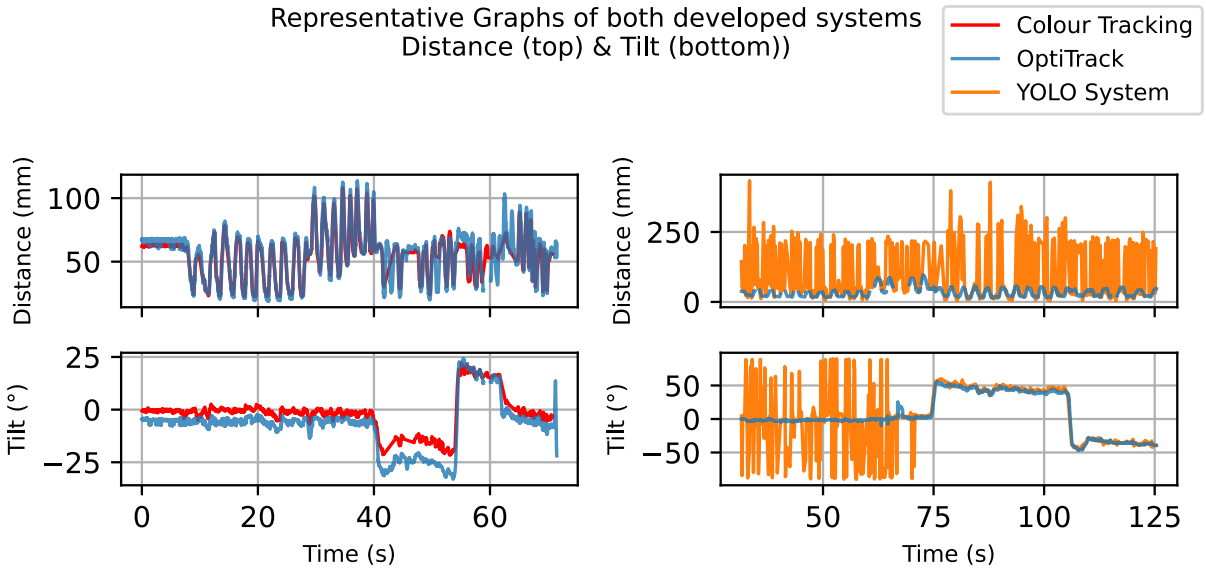

Figure 4. A representative overview of the data obtained from the YOLO System (orange) and Colour Tracking (green) versus the OptiTrack data (blue). No interpolation has been done here, the results are as obtained during the trial.

(e.g. hyperparameter tuning), nor in the post-training phase (e.g. pruning or quantisation). Applying these standard optimisation techniques could significantly improve both inference speed and accuracy [32].

In addition to model refinement, further performance gains could be achieved by addressing the inherent noise in YOLO's probabilistic keypoint estimates. Implementing a Kalman filter in conjunction with model optimisation may prove effective. Kalman filters are well-suited for noise reduction and state prediction in dynamic systems and are efficient under real-time constraints due to their low computational load [33]. In this context, a Kalman filter could not only smooth noisy keypoint estimates but potentially increase the effective detection frequency by treating predicted positions as valid detections, though this would require thorough validation.

Regarding robustness, the system performs consistently across most factors ($p < 0.05$, $\eta_p^2 < 0.06$), with the exception of a pronounced effect of background on detection frequency. Here, detection dropped notably on the Black Cloth background (3.91 Hz) compared to Whiteboard (4.98 Hz) and Colourful (5.38 Hz) settings. A two-way Welch ANOVA confirmed this

effect to be significant, $p = 0.032$, $\eta_p^2 = 0.475$. The likely cause is the low visual contrast between the grey shell and the dark background, which impairs feature detection.

Additionally, robustness criteria are not met with respect to the effect of Movement on the tilt AE, with a highly significant result of $p \ll 0.001$ and $\eta_p^2 = 0.1629$. However, the reliability of this result is uncertain, as tilt ground truth was calculated based on the assumption that the OptiTrack coordinate system's base plane was perpendicular to the camera's viewing axis, an estimation rather than an exact measurement. More accurate OptiTrack-based tilt data could be obtained by placing markers along the shell's spine. To improve the model's own tilt estimation, the knobs could be made asymmetric, or an additional knob could be added along the spine. Both approaches would enrich the 3D spatial information available to the model and allow for more accurate reconstruction of the shell's orientation matrix **R**.

2. Performance of the Colour Tracking System

The colour tracking system delivered strong baseline performance, especially considering its simplicity and low computational cost.

Firstly, the distance accuracy of the colour tracking system showed promise based on set conditions. Table 4 demonstrates that especially in the opening and closing phase, trials reached mean accuracies below the 4.5 mm threshold with an AE of 5.4 ± 5.1 mm (Black Cloth, Off) to even 3.2 ± 2.5 mm (Whiteboard, On). In the supination and pronation, these accuracies were lower with a maximum of 16.4 ± 19.2 mm (Whiteboard, Off), reflecting increased calculating difficulty with the rotated shell ends. From this, we conclude that the colour tracking system, overall, did not meet the design goals. The tilt estimation follows the same trend, with an AE in between 3.9° and 5.9°, the error is almost always bigger during pronation and supination than during the opening and closing of the shell.

Moreover, the sampling rate of the colour-based vision tracking system reached the set conditions of a frequency higher than 10 Hz. This sampling ranged from 20.89 Hz up to 27.97 Hz, resulting in real-time feedback for the user. This sampling rate is due to its low computational cost, making it easily implementable and home-user friendly.

At last, regarding robustness, the colour-based tracking system can be considered robust over each factor ($p < 0.05$ and $\eta_p^2 < 0.06$). Background and lighting were both significant for the distance error ($p < 0.05$), but only on a small scale based on the η_p^2 being 0.0028 and 0.0044. For the tilt error, only the background would be considered significant with $p = 4.54 \times 10^{-32}$, but also considering the $\eta_p^2 = 0.0044$, this was only on a small scale, resulting in the colour-based tracking system being considered robust, fulfilling our design goals. For the sampling rate the system can be considered robust over each factor as all of these can be considered insignificant ($p > 0.05$).

In the end, performance stayed stable even under the coloured background, which used the same colours as the markers. This is likely due to the algorithm's grouping and noise filtering. Lighting still had some effect, errors increased with glare, shadows or HSV distortion, but not enough to break robustness.

A key limitation came when the hand rotated over 45°. Markers occasionally grouped incorrectly, switching shell end positions and consequently not performing to the standards. Future versions could improve marker placement or algorithm logic. Also, while pronation or supination angles still worked, upward tilt made the markers invisible which stopped tracking entirely.

Still, the system reached up to 28 Hz, kept AE mostly between 4-7 mm and met the robustness criteria. It tracks opening or closing accurately and while it is not optimal for tracking difficult rotational motions, it handles basic rotation. This makes the colour tracking system a realistic low-cost option for home rehabilitation due to its transparency, ease of calibration and minimal resource requirement.

B. Limitations

Besides the aforementioned model specific limitations the experimental results should in general be interpreted with some caveats. Firstly, direct comparison between the two systems should be done with caution. Since each system was tested separately on separately days, and although OptiTrack should be a viable ground truth in both metrics, differences in software and hardware set ups, such as the calibration and camera location, might introduce variability into

the ground truth data. The missing data in both results, suggest as much. Moreover, lighting conditions can not be proven to be same for both experiments, as the light measurements are inaccurate given the equipment used (±12% and a max. 10% cosine error), and the Colour Tracking has no valid lighting measurement results. However, one can reasonably assume that the effects of the lights being switched on and off will have been similar in each experiments, as the lux measurements during the YOLO system testing, suggest an increase of a couple hundred lux with the lights on, which happened for each trial and under different daylight conditions (due to time passing by on experimentation day). Daylight conditions on both days were similar according to weather reports (29 May and 4 June) [34, 35]. Additionally, for each system tested a different participant performed the movements with the shell. Leading to inevitable variations in movement executions, which could influence the results. The difference, for example, is a measured resultant difference. And finally, the placement of the OptiTrack markers themselves, are not guaranteed to be perfectly centroid to the shell ends axis, giving an error for the exact location of the shell ends centre points that is different for both systems.

V. Conclusion

Ultimately, neither approach meets initial design goals of maximum mean absolute error (AE) of 4.5 mm for distance and 2.6° for tilt, and a minimum detection frequency of 10 Hz and the robustness being invalid for $p < 0.052$, $\eta_p^2 \geq 0.06$. With the YOLO System performing the worst (overall AE of 126.0±219.7 mm for distance, 18.2±25.1° for tilt, and an average detection frequency of 4.76±0.93 Hz and robustness was only invalid for the Background effect on the detection frequency: $p = 0.032$, $\eta_p^2 = 0.475$). This can be attributed to the inherent noise of the probabilist DL model, which was accentuated due to limited optimisation being applied to the model. A Kalman Filter and the optimisation of the data generation, training, and post-training phases, could substantially improve performance. The Colour Tracker, however, performed close to the desired metrics, having a detection frequency ranged from 20.89 Hz up to 27.97 Hz, well above the 10 Hz minimum. However especially with regards to the Pronated and Supinated phases the accuracy was not met with a maximum of 16.4 ± 19.2 mm (Whiteboard, Off). Yet, Robustness is validaded given the insignificant effects of all factor $p < 0.05$. However light was not tested for true over or under exposure, which is commonly the downfall for colour trackers, hence further testing is necessary, especially given that light has shown to be problematic during eraly testing of the Colour Tracker. If improvements, cannot be made to create robust lighting results, the YOLO-system might be a better alternative, if its optimised.

Acknowledgments

We would like to thank, Dr. Ir. L Marchal Crespo, Dr. Ir. A. Bellitto, Ir. I. Beck, and Ir. B. Horstman for their support and guidance throughout the project. Author C.A. Brocx would also like to thank Mr. M.G. Brocx for his generosity in extending acces to much needed computing power.

References

[1] Feigin, V. L., Brainin, M., Norrving, B., Martins, S. O., Pandian, J., Lindsay, P., Grupper, M. F., and Rautalin, I., "World Stroke Organization: Global Stroke Fact Sheet 2025," *International Journal of Stroke*, Vol. 20, No. 2, 2025, pp. 132–144. doi: 10.1177/17474930241308142, URL https://doi.org/10.1177/17474930241308142, pMID: 39635884.

[2] Raghavan, P., "Upper Limb Motor Impairment After Stroke," *Physical Medicine and Rehabilitation Clinics of North America*, Vol. 26, No. 4, 2015, pp. 599–610. doi: 10.1016/j.pmr.2015.06.008, URL https://doi.org/10.1016/j.pmr.2015.06.008.

[3] Zbytniewska-Mégret, M., Salzmann, C., Kanzler, C. M., Hassa, T., Gassert, R., Lambercy, O., and Liepert, J., "The Evolution of Hand Proprioceptive and Motor Impairments in the Sub-Acute Phase After Stroke," *Neurorehabilitation and Neural Repair*, Vol. 37, No. 11-12, 2023, pp. 823–836. doi: 10.1177/15459683231207355, URL https://doi.org/10.1177/15459683231207355, pMID: 37953595.

[4] Langer, D., Horwitz, A., Melchior, H., Atoun, E., and Mazor-Karsenty, T., "Understanding the implications of hand impairments in light of the International Classification of Function model," *Journal of Hand Therapy*, Vol. 38, No. 1, 2025, pp. 122–128. doi: 10.1016/j.jht.2024.05.004, URL

https://www.sciencedirect.com/science/article/pii/S0894113024000474.

[5] Kabir, R., Sunny, M. S. H., Ahmed, H. U., and Rahman, M. H., "Hand Rehabilitation Devices: A Comprehensive Systematic Review," *Micromachines*, Vol. 13, No. 7, 2022. doi: 10.3390/mi13071033, URL https://www.mdpi.com/2072-666X/13/7/1033.

[6] Polygerinos, P., Galloway, K. C., Savage, E., Herman, M., Donnell, K. O., and Walsh, C. J., "Soft robotic glove for hand rehabilitation and task specific training," *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 2913–2919. doi: 10.1109/ICRA.2015.7139597.

[7] Rho, E., Lee, H., Lee, Y., Lee, K.-D., Mun, J., Kim, M., Kim, D., Park, H.-S., and Jo, S., "Multiple Hand Posture Rehabilitation System Using Vision-Based Intention Detection and Soft-Robotic Glove," *IEEE Transactions on Industrial Informatics*, Vol. 20, No. 4, 2024, pp. 6499–6509. doi: 10.1109/TII.2023.3348826.

[8] Rätz, R., Ratschat, A. L., Cividanes-Garcia, N., Ribbers, G. M., and Marchal-Crespo, L., "Designing for usability: development and evaluation of a portable minimally-actuated haptic hand and forearm trainer for unsupervised stroke rehabilitation," *Frontiers in Neurorobotics*, Vol. Volume 18 - 2024, 2024. doi: 10.3389/fnbot.2024.1351700, URL https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2024.1351700.

[9] Van Damme, N., Rätz, R., and Marchal-Crespo, L., "Towards Unsupervised Rehabilitation: Development of a Portable Compliant Device for Sensorimotor Hand Rehabilitation," *2022 International Conference on Rehabilitation Robotics (ICORR)*, 2022, pp. 1–6. doi: 10.1109/ICORR55369.2022.9896556.

[10] Rätz, R., Van Damme, N., Marchal-Crespo, L., and Aksöz, E. A., "Sensomotoric Hand Therapy Device," , 2023. 5th Edn. Churchill Livingstone.

[11] Li, Z., Kanazuka, A., Hojo, A., Nomura, Y., and Nakaguchi, T., "Multi-metric assessment of hand motion with multi-camera for puncture technique training," *Measurement*, Vol. 248, 2025, p. 116865. doi: 10.1016/j.measurement.2025.116865, URL https://www.sciencedirect.com/science/article/pii/S0263224125002246.

[12] Koter, K., Samowicz, M., Redlicka, J., and Zubrycki, I., "Hand Measurement System Based on Haptic and Vision Devices towards Post-Stroke Patients," *Sensors*, Vol. 22, No. 5, 2022. doi: 10.3390/s22052060, URL https://www.mdpi.com/1424-8220/22/5/2060.

[13] Lang, C. E., Edwards, D. F., Birkenmeier, R. L., and Dromerick, A. W., "Estimating minimal clinically important differences of upper-extremity measures early after stroke," *Archives of Physical Medicine and Rehabilitation*, Vol. 89, No. 9, 2008, pp. 1693–1700.

[14] Attig, C., Rauh, N., Franke, T., and Krems, J., "System Latency Guidelines Then and Now – Is Zero Latency Really Considered Necessary?" 2017, pp. 3–14. doi: 10.1007/978-3-319-58475-1-1.

[15] Kaaresoja, T., Brewster, S., and Lantz, V., "Towards the Temporally Perfect Virtual Button: Touch-Feedback Simultaneity and Perceived Quality in Mobile Touchscreen Press Interactions," *ACM Transactions on Applied Perception*, Vol. 11, 2014, pp. 1–25. doi: 10.1145/2611387.

[16] Cohen, J., *Statistical power analysis for the behavioral sciences*, routledge, 2013.

[17] Lakens, D., "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs," *Frontiers in psychology*, Vol. 4, 2013, p. 863.

[18] Zhao, Z.-Q., Zheng, P., Xu, S.-T., and Wu, X., "Object Detection With Deep Learning: A Review," *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 30, No. 11, 2019, pp. 3212–3232. doi: 10.1109/TNNLS.2018.2876865.

[19] Zhang, Y., Li, X., Wang, F., Wei, B., and Li, L., "A comprehensive review of one-stage networks for object detection," *2021 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, IEEE, 2021, pp. 1–6.

[20] Kaur, J., and Singh, W., "Tools, techniques, datasets and application areas for object detection in an image: a review," *Multimedia Tools and Applications*, Vol. 81, No. 27, 2022, pp. 38297–38351.

[21] Shetty, A. K., Saha, I., Sanghvi, R. M., Save, S. A., and Patel, Y. J., "A review: Object detection models," *2021 6th International Conference for Convergence in Technology (I2CT)*, IEEE, 2021, pp. 1–8.

[22] Ultralytics, "YOLOv11 Performance Metrics," , 2024. URL https://docs.ultralytics.com/models/yolo11/#performance-metrics, accessed: 2025-04-06.

[23] Bradski, G., "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[24] Marchand, E., Uchiyama, H., and Spindler, F., "Pose Estimation for Augmented Reality: A Hands-On Survey," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 22, No. 12, 2016, pp. 2633–2651. doi: 10.1109/TVCG.2015.2513408.

[25] Zhang, Z., "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 11, 2000, pp. 1330–1334. doi: 10.1109/34.888718.

[26] Denninger, M., Winkelbauer, D., Sundermeyer, M., Boerdijk, W., Knauer, M., Strobl, K. H., Humt, M., and Triebel, R., "BlenderProc2: A Procedural Pipeline for Photorealistic Rendering," *Journal of Open Source Software*, Vol. 8, No. 82, 2023, p. 4901. doi: 10.21105/joss.04901, URL https://doi.org/10.21105/joss.04901.

[27] Jocher, G., Qiu, J., and Chaurasia, A., "Ultralytics YOLO," , 2023. URL https://ultralytics.com, if you use this software, please cite it using this metadata.

[28] A. Buslaev, E. K. V. I. I., A. Parinov, and Kalinin, A. A., "Albumentations: fast and flexible image augmentations," *ArXiv e-prints*, 2018.

[29] Kang, H.-C., Han, H.-N., Bae, H.-C., Kim, M.-G., Son, J.-Y., and Kim, Y.-K., "HSV Color-Space-Based Automated Object Localization for Robot Grasping without Prior Knowledge," *Applied Sciences*, Vol. 11, No. 16, 2021. doi: 10.3390/app11167593, URL https://www.mdpi.com/2076-3417/11/16/7593.

[30] Gonzalez, R. C., and Woods, R. E., *Digital Image Processing*, 3rd ed., Pearson, Upper Saddle River, NJ, 2008.

[31] Barnett, V., and Lewis, T., *Outliers in Statistical Data*, 3rd ed., John Wiley & Sons, Chichester, UK, 1994.

[32] Ultralytics, "What is Model Optimization? A Quick Guide," , 2024. URL https://www.ultralytics.com/blog/what-is-model-optimization-a-quick-guide, accessed: 2025-06-13.

[33] Khodarahmi, M., and Maihami, V., "A review on Kalman filter models," *Archives of Computational Methods in Engineering*, Vol. 30, No. 1, 2023, pp. 727–747.

[34] Time and Date, "Historic Weather: Amsterdam, Netherlands – May 2025," https://www.timeanddate.com/weather/netherlands/amsterdam/historic?month=5&year=2025, May 2025. Accessed: 2025-06-13.

[35] Time and Date, "Historic Weather: Amsterdam, Netherlands – June 2025," https://www.timeanddate.com/weather/netherlands/amsterdam/historic?month=6&year=2025, June 2025. Accessed: 2025-06-13.

[36] Viola, P., and Jones, M., "Rapid object detection using a boosted cascade of simple features," *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, Vol. 1, Ieee, 2001, pp. I–I.

[37] Friedman, J. H., "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, Vol. 29, No. 5, 2001, pp. 1189–1232.

[38] Lowe, D. G., "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, Vol. 60, No. 2, 2004, pp. 91–110.

[39] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A., "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

[40] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A., "The pascal visual object classes (voc) challenge," *International journal of computer vision*, Vol. 88, 2010, pp. 303–338.

[41] Zou, Z., Chen, K., Shi, Z., Guo, Y., and Ye, J., "Object detection in 20 years: A survey," *Proceedings of the IEEE*, Vol. 111, No. 3, 2023, pp. 257–276.

[42] Neha, F., Bhati, D., Shukla, D. K., and Amiruzzaman, M., "From classical techniques to convolution-based models: A review of object detection algorithms," *2025 IEEE 6th International Conference on Image Processing, Applications and Systems (IPAS)*, IEEE, 2025, pp. 1–6.

[43] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C., "SSD: Single Shot MultiBox Detector," *European Conference on Computer Vision*, Springer, 2016, pp. 21–37.

[44] Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A., and Berg, A. C., "DSSD: Deconvolutional Single Shot Detector," *arXiv preprint arXiv:1701.06659*, 2017.

[45] Li, Z., Zhou, C., Xu, Y., Liang, B., Yu, G., and Feng, Z., "FSSD: Feature Fusion Single Shot Multibox Detector," *arXiv preprint arXiv:1712.00960*, 2017.

[46] Chiu, Y.-C., Tsai, C.-Y., Ruan, M.-D., Shen, G.-Y., and Lee, T.-T., "Mobilenet-SSDv2: An Improved Object Detection Model for Embedded Systems," *2020 International Conference on System Science and Engineering (ICSSE)*, 2020, pp. 1–5. doi: 10.1109/ICSSE50014.2020.9219319.

[47] Magalhães, S. A., Castro, L., Moreira, G., dos Santos, F. N., Cunha, M., Dias, J., and Moreira, A. P., "Evaluating the Single-Shot MultiBox Detector and YOLO Deep Learning Models for the Detection of Tomatoes in a Greenhouse," *Sensors*, Vol. 21, No. 10, 2021. doi: 10.3390/s21103569, URL https://www.mdpi.com/1424-8220/21/10/3569.

[48] Jeong, J., Park, H., and Kwak, N., "Enhancement of SSD by concatenating feature maps for object detection," *arXiv preprint arXiv:1705.09587*, 2017.

Appendix

Nomenclature

C = Camera space (coordinate system of the camera)

D = Estimated Euclidean distance between shell ends

$\mathcal{I}$ = Image space (2D pixel coordinates)

$\mathbf{K}$ = Camera intrinsic matrix

$\mathcal{M}$ = Model space (3D reference frame of each shell end)

$\mathbf{p}_{i,j}^{C}$ = 3D position of point i on shell end j in camera space

$\bar{\mathbf{p}}_{i}^{\mathcal{I}}$ = Measured 2D image coordinates (with noise)

$\mathbf{p}_{i}^{\mathcal{I}}$ = 2D image coordinates of point i

$\mathbf{p}_{i}^{\mathcal{M}}$ = 3D coordinates of point i in the model space

$\mathbf{R}$ = Rotation matrix from model to camera frame

s = Scaling factor from pixel space to camera space

$\mathbf{T}$ = Translation vector from model to camera frame

A. Comparison of Detection Algorithms

Criterion	Traditional		DL Two-Stage		DL One-Stage	
	Score	Reasoning	Score	Reasoning	Score	Reasoning
Accuracy/ Robustness	2	33.7 mAP on VOC07 (DPMv5); manual features limit adaptability.	5	66–76 mAP on Pascal VOC.	5	72.5–75 mAP on Pascal VOC.
Speed	3	15 FPS (Haar Cascades).	2	5–10 FPS.	5	19–60 FPS.
Example Architectures	Haar Cascades [36], Gradient boosting [37], SIFT [38]		R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN, R-FCN, FPN, G-RCNN		SSD family, YOLO family	

Table 6. Comparison of traditional and deep-learning (two-stage and one-stage) object detection methods. The combined "Accuracy/Robustness" are grouped, given that the Visual Object Classes (VOC) benchmark dataset (2007 [39] and 2010 [40]) is multi class and has varied conditions, hence the resultant mean Average Precision (mAP) can be seen as a measure for both accuracy and robustness. The scores and backing thereof is intended as a preliminary guide rather than a set of absolute metrics, and are based on a results from different algorithms within their category. These values helped orient our choice of approach: DL one stage. Data from [36, 41, 42]

.

Attribute	**YOLO**		**SSD**	
	Score	Reasoning	Score	Reasoning
Speed	4	ONYX CPU: 56.1–183 ms; T4 TensorRT: 1.5–4.7 ms	4	V100 GPU: 39 ms
Accuracy	5	COCO mAP (640×640, 50–95): 39.5–51.7	3	COCO mAP (640×640, 50–95): 28.2
Reliability	4	Performs well across varied environments due to robust training and design	2	Degrades under lighting or background changes
Implementation	5	Python-based, modular, easy integration	3	Less support, more fragmented implementations
Computing Power	4	BFLOPS: 6.5–68 depending on model	5	very light weight
Key Point	5	Dedicated pose models available	1	No native support, and no literature found where a custom extension is added
Segmentation	5	Dedicated segmentation models exist	1	No native support, and no literature found where a custom extension is added
Extra remarks	Based on YOLOv11-n, s, m. Pose and segmentation models influence both speed and accuracy.		SSD performance depends on backbone. MobileNetv2 used benchmark. because it is lightweight and has shown to be good in single-object detection. See appendix for variants.	
Sources	[22]		For detailed references see table A2	

Table A1. Compact comparison between YOLO and SSD

Attribute	Model A: SSD		Model B: DSSD	
	Score	Reasoning	Score	Reasoning
Framework	4	Python supported	4	Python supported
Speed	3	Titan X: 19 FPS on VOC 2007	1	Titan X: 5.5–6.6 FPS on VOC
Accuracy	4	VOC: 79.5, COCO: 28.8	4	VOC: 81.5, COCO: 33.2
Implementation	4	VGG16-based with 8,732 default boxes	2	ResNet-101 backbone with deconv layers
Computing Power	5	Low requirements	2	High compute due to context layers
Key Point	0	No integrated models	0	No integrated models
Segmentation	0	No integrated models	0	No integrated models
Extra remarks	Based on SSD512 (most accurate standard SSD variant)		Based on DSSD513. Improves accuracy over SSD512 at cost of speed.	
Sources	[43, 44]		[44]	

Attribute	Model C: FSSD		Model D: MobileNet-SSD	
	Score	Reasoning	Score	Reasoning
Framework	5	Python supported	3	Python supported
Speed	5	1080Ti: 35.7 FPS on VOC 2007	3	Jetson Xavier: 21 FPS on VOC 2007
Accuracy	5	VOC: 84.5	3	VOC: 75.6
Implementation	4	Feature fusion module added for accuracy	5	MobileNet backbone optimized for embedded use
Computing Power	3	Moderate compute cost	5	Very efficient for low-power devices
Key Point	0	No integrated models	0	No integrated models
Segmentation	0	No integrated models	0	No integrated models
Extra remarks	Based on FSSD512, improves accuracy over SSD512 with moderate overhead		Based on MobileNet-SSDv2 + FPN. Excellent for single-class detection.	
Sources	[45]		[46, 47]	

Attribute	Model E: R-SSD		—	
	Score	Reasoning	Score	Reasoning
Framework	2	Python supported	—	
Speed	3	Titan X: 16.6 FPS on VOC 2007	—	
Accuracy	4	VOC: 80.8	—	
Implementation	3	Increased input size and complexity	—	
Computing Power	3	Moderate–high compute category	—	
Key Point	0	No integrated models	—	
Segmentation	0	No integrated models	—	
Extra remarks	Based on RSSD512, with improved accuracy over SSD512		—	
Sources	[48]		—	

Table A2. SSD-Based Model Comparison: Speed, Accuracy, and Deployment Tradeoffs

B. Synthetic Data Generation and Model Training

Component	Randomised Elements	Purpose
Marker Instances	Instances, position, orientation	Scene structure
Shell & Hand	Conditional presence based on setup	Object variation
Materials	Color, reflectance, transparency, texture	Visual diversity
Camera	Position, orientation, intrinsics	Viewpoint variation
Background	HDRI map, brightness, rotation	Enviromental diversity
Scene Clutter	Quantity, shape, size, placement	Occlusion and realism
Marker Occlusion	Optional stacks of primitive shapes (cubes, cylinders, etc) ontop specific markers	Simulate OptiTrack Marker like Occlusion

Table B1. Randomised components in the BlenderProc data pipeline. Marker instances define keypoint annotations; shell and hand are conditionally added.

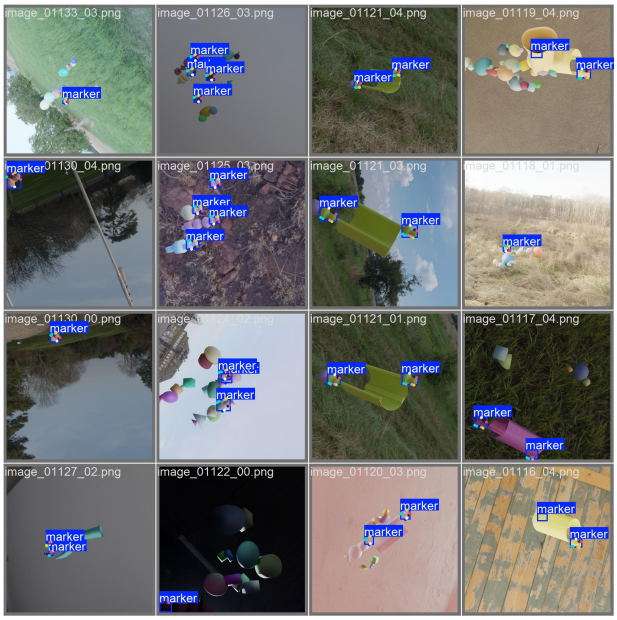

Figure B1. A batch of the synthetic data generated. One can see a variety of poses and angles, and the different randomisations, with or without shell and or hand, with our without stacks of primitive shapes. The bounding box is blue, whilst the key points are labelled as different coloured dots.

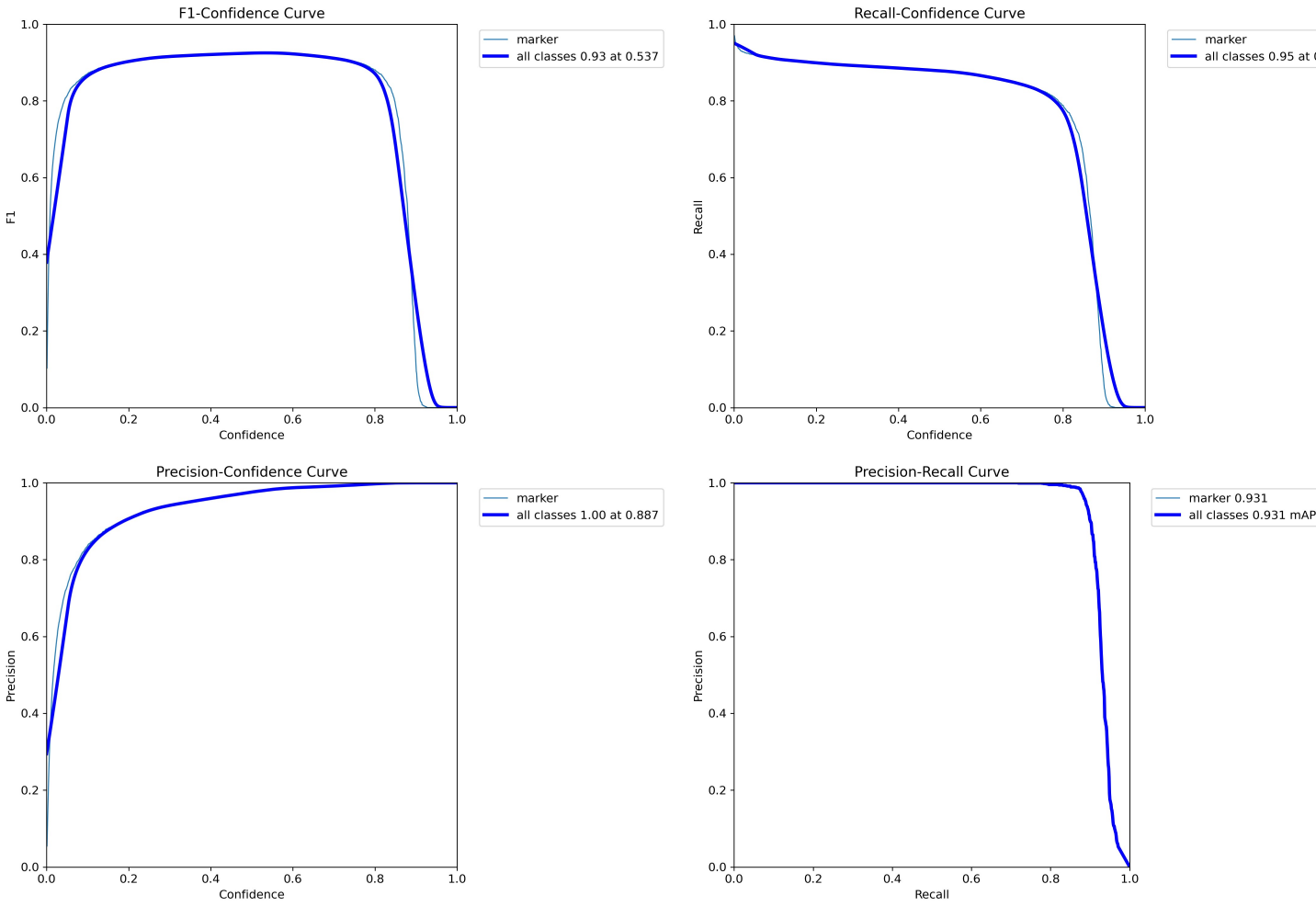

Figure B2. Training performance of the YOLOv11n-pose model on bounding box detection. The plots show: (1) Precision and Recall as functions of detection confidence, (2) F1 score across confidence thresholds, and (3) the Precision–Recall curve summarizing overall detection quality. High F1 values and a Precision–Recall curve approaching the top-right corner indicate that the model achieved strong bounding box localization performance over the training dataset.

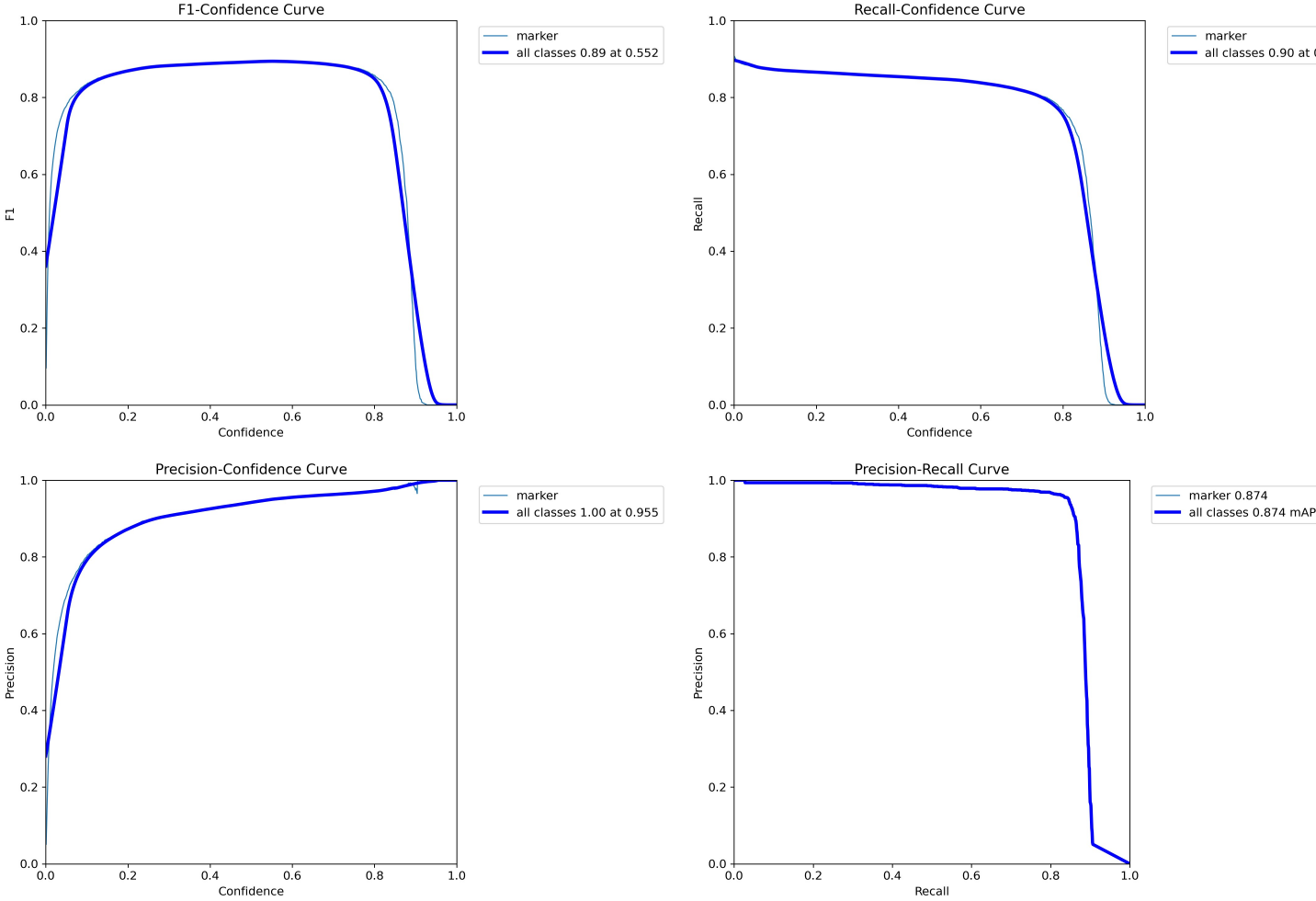

Figure B3. Training performance of the YOLOv11n-pose model on key point (pose) estimation. The plots show: (1) Precision and Recall as functions of detection confidence, (2) F1 score across confidence thresholds, and (3) the Precision–Recall curve for key point predictions. Compared to bounding box detection, pose estimation typically presents a greater challenge, and performance is reflected by lower curve sharpness. Nonetheless, the model shows stable key point learning across the training dataset.

C. Assumption Checks for ANOVA

To determine whether standard ANOVA could be used, we evaluated the assumptions of homogeneity of variance and normality of residuals for each analysis. Welch's ANOVA was applied when these assumptions were violated.

1. Deep Learnign ANOVA Checks

Detection Frequency (Hz) Levene's Test for Homogeneity of Variance (center = median)

- Background: $W = 5.215$, $p = 0.019$
- Lighting: $W = 0.413$, $p = 0.530$

Normality of Residuals

- Shapiro–Wilk: $W = 0.975$, $p = 0.878$
- D'Agostino–Pearson: $\chi^2 = 0.264$, $p = 0.876$
- Jarque–Bera: $JB = 0.281$, $p = 0.869$

Conclusion: Normality assumed, but unequal variances for Background → Welch's ANOVA applied.

Tilt Error(°) Levene's Test (center = median)

- Background: $W = 26.946$, $p < 0.001$
- Lighting: $W = 0.815$, $p = 0.367$
- Movement: $W = 772.270$, $p < 0.001$

Normality of Residuals

- D'Agostino–Pearson: $\chi^2 = 1153.024$, $p < 0.001$
- Jarque–Bera: $JB = 1819.120$, $p < 0.001$

Conclusion: Assumptions violated → Welch's ANOVA applied.

Distance Error (mm) Levene's Test (center = median)

- Background: $W = 209.472$, $p < 0.001$
- Lighting: $W = 222.721$, $p < 0.001$
- Movement: $W = 7.335$, $p < 0.001$

Normality of Residuals

- D'Agostino–Pearson: $\chi^2 = 6121.478$, $p < 0.001$
- Jarque–Bera: $JB = 599105.187$, $p < 0.001$

Conclusion: Assumptions violated → Welch's ANOVA applied.

2. Colour Tracking Assumption Checks

Detection Frequency (Hz)

Tilt Error (°) Levene's Test for Homogeneity of Variance (center = median)

- Background: $W = 90.748$, $p \ll 0.001$
- Lighting: $W = 117.246$, $p \ll 0.001$
- Movement: $W = 622.040$, $p \ll 0.001$

Normality of Residuals

- D'Agostino–Pearson: $\chi^2 = 26875.687$, $p \ll 0.001$
- Jarque–Bera: $JB = 1226554.192$, $p \ll 0.001$

Conclusion: Assumptions violated → Welch's ANOVA applied.

Distance Error (mm) Levene's Test for Homogeneity of Variance (center = median)

- Background: $W = 19.688$, $p \ll 0.001$
- Lighting: $W = 178.540$, $p \ll 0.001$
- Movement: $W = 733.101$, $p \ll 0.001$

Normality of Residuals

- D'Agostino–Pearson: $\chi^2 = 12283.025$, $p \ll 0.001$
- Jarque–Bera: $JB = 69270.560$, $p \ll 0.001$

Conclusion: Assumptions violated → Welch's ANOVA applied.

Conclusion: Assumptions violated → Welch's ANOVA applied.

D. Results in Detail

Background	**Whiteboard**		**Black Cloth**		**Colourful**	
Lighting	**Off**	**On**	**Off**	**On**	**Off**	**On**
1st trial	14-12	16-20	12	20-19	24-20	23-22
2nd trial	12-13	17-16	12-10	19-18	17-18	22-21
3rd trial	13	17-18	11-12	17	17-19	24-21

Table D1. Overview of the lighting conditions during the testing of the YOLO-based tracker in 1000 lux. Also, it was measured at the start and end of each trial, the first value - last value.

Background	**Whiteboard**								**Black Cloth**								**Colourful**							
Lighting	**Off**				**On**				**Off**				**On**				**Off**				**On**			
	C	O	P	S	C	O	P	S	C	O	P	S	C	O	P	S	C	O	P	S	C	O	P	S
1st trial	1	0	9	5	4	0	10	5	9	5	10	5	6	0	10	5	4	0	9	5	10	4	10	5
2nd trial	5	0	10	5	8	1	9	5	10	5	10	5	10	5	10	5	8	5	10	5	9	5	9	5
3rd trial	2	0	10	5	1	0	10	4	10	2	10	5	6	0	10	5	4	0	10	5	3	1	10	4
Total	**8**	**0**	29	15	**13**	**1**	29	14	29	12	30	15	22	**5**	30	15	16	**5**	29	15	22	10	29	14

Table D2. Overview of the amount of valid repetitions per background, lighting condition, and gesture category (C = Closing, O = Opening, P = Pronated, S = Supinated).

Background	Lighting		Total			Closing	Opening	Pronated	Supinated
			Sampling rate [Hz] ($\mu \pm \sigma$)	RMSE [mm]	AE [mm] ($\mu \pm \sigma$)	AE [mm] ($\mu \pm \sigma$)	AE [mm] ($\mu \pm \sigma$)	AE [mm] ($\mu \pm \sigma$)	AE [mm] ($\mu \pm \sigma$)
Whiteboard	Off	1st	4.49	96.5	68.1 ± 68.6	23.8 ± 32.2	—	51.9 ± 69.4	100.0 ± 58.4
		2nd	5.83	167.8	109.2 ± 127.6	189.5 ± 195.9	—	47.8 ± 41.6	142.8 ± 108.0
		3rd	5.11	122.2	96.4 ± 75.3	83.5 ± 52.2	—	50.1 ± 46.4	160.6 ± 64.2
		total	5.15 ± 0.67	133.9	92.3 ± 97.1	139.7 ± 167.3	—	50.0 ± 54.1	135.7 ± 83.1
Whiteboard	On	1st	4.27	117.8	88.0 ± 78.5	90.5 ± 111.0	81.0 ± 49.3	89.2 ± 74.3	83.9 ± 59.6
		2nd	5.12	156.2	123.7 ± 95.5	87.0 ± 51.2	49.3 ± 110.6	110.6 ± 93.8	201.2 ± 106.5
		3rd	5.04	150.3	120.9 ± 89.6	55.1 ± 41.2	—	92.2 ± 61.8	204.6 ± 100.0
		total	4.81 ± 0.47	143.7	111.9 ± 90.3	86.0 ± 71.7	49.3 ± 81.0	96.9 ± 77.4	164.4 ± 107.1
Black Cloth	Off	1st	3.38	176.5	121.7 ± 128.1	137.3 ± 157.9	40.4 ± 80.3	121.7 ± 65.8	159.9 ± 154.6
		2nd	4.76	72.8	50.6 ± 52.5	61.4 ± 50.1	33.5 ± 49.5	29.9 ± 25.4	100.9 ± 73.8
		3rd	2.97	94.2	70.8 ± 62.3	64.7 ± 67.3	10.2 ± 10.3	100.3 ± 42.0	53.0 ± 68.8
		total	3.70 ± 0.94	122.7	80.8 ± 92.4	85.3 ± 105.2	31.4 ± 60.9	80.3 ± 60.8	108.2 ± 120.5
Black Cloth	On	1st	4.44	114.9	79.8 ± 82.8	111.2 ± 69.9	—	52.8 ± 48.4	111.4 ± 125.2
		2nd	2.71	129.7	66.7 ± 111.4	111.6 ± 145.4	93.2 ± 153.8	41.0 ± 62.2	28.9 ± 46.0
		3rd	5.18	132.0	117.0 ± 61.2	117.8 ± 87.8	—	122.5 ± 43.8	102.9 ± 45.5
		total	4.11 ± 1.27	125.9	86.4 ± 91.6	113.6 ± 110.4	93.2 ± 153.8	71.1 ± 62.7	83.2 ± 91.9
Colourful	Off	1st	4.55	111.9	82.9 ± 75.3	58.2 ± 59.3	—	81.1 ± 81.6	104.9 ± 68.4
		2nd	5.86	63.7	32.8 ± 54.6	29.2 ± 45.5	26.5 ± 42.9	43.1 ± 68.2	23.5 ± 42.1
		3rd	5.53	107.0	79.1 ± 72.2	41.4 ± 55.5	—	68.1 ± 66.3	124.1 ± 71.2
		total	5.31 ± 0.68	92.9	60.6 ± 70.4	39.5 ± 52.9	26.5 ± 42.9	84.1 ± 75.7	84.1 ± 75.7
Colourful	On	1st	5.31	253.6	139.9 ± 211.9	87.8 ± 289.8	73.9 ± 67.5	196.8 ± 109.3	190.6 ± 193.4
		2nd	5.53	805.1	604.4 ± 532.5	629.8 ± 720.8	445.6 ± 542.3	606.0 ± 323.3	698.3 ± 435.3
		3rd	5.52	119.0	83.0 ± 85.5	61.7 ± 84.2	—	86.8 ± 92.9	88.7 ± 72.6
		total	5.45 ± 0.12	530.9	308.7 ± 432.2	307.5 ± 569.9	275.7 ± 441.3	296.1 ± 304.7	353.5 ± 401.8

Table D3. Trial-level and total performance of the YOLOv11n-pose tracking system for distance estimation under varying background (Whiteboard, Black Cloth, Colourful) and lighting (Off, On) conditions. Speed [Hz] is shown for each trial (1st/2nd/3rd) according to the measured SYS sampling rate, and as mean± SD in the total row. RMSE [mm] and mean ± SD of absolute error (AE) are reported both for the entire repetition (Total) and for each movement phase (Closing, Opening, Pronated, Supinated).

Background	Lighting		Total			Closing	Opening	Pronated	Supinated
			Sampling rate [Hz] ($\mu \pm \sigma$)	RMSE [°]	AE [°] ($\mu \pm \sigma$)	AE [°] ($\mu \pm \sigma$)	AE [°] ($\mu \pm \sigma$)	AE [°] ($\mu \pm \sigma$)	AE [°] ($\mu \pm \sigma$)
Whiteboard	Off	1st	4.49	34.84	19.75 ± 28.73	39.39 ± 34.37	19.59 ± 28.15	5.60 ± 3.24	1.83 ± 1.82
		2nd	5.83	38.32	25.28 ± 28.83	41.04 ± 35.27	40.87 ± 36.63	8.98 ± 1.51	13.94 ± 1.32
		3rd	5.11	28.48	16.75 ± 23.07	30.77 ± 31.30	12.41 ± 20.67	9.80 ± 2.29	5.20 ± 3.26
		total	5.15 ± 0.67	33.96	20.37 ± 27.19	36.80 ± 33.80	23.16 ± 30.78	8.08 ± 3.07	6.94 ± 5.53
Whiteboard	On	1st	4.27	34.09	18.92 ± 28.39	47.15 ± 32.76	13.55 ± 24.10	2.91 ± 1.98	5.36 ± 1.73
		2nd	5.12	27.29	16.53 ± 21.75	35.25 ± 30.04	7.07 ± 11.34	10.83 ± 2.28	3.26 ± 1.51
		3rd	5.04	21.09	8.88 ± 19.16	14.33 ± 24.65	19.23 ± 29.06	1.25 ± 1.18	5.30 ± 1.82
		total	4.81 ± 0.47	27.80	14.53 ± 23.71	31.34 ± 32.20	14.33 ± 24.48	4.79 ± 12.48	4.59 ± 1.96
Black Cloth	Off	1st	3.38	31.27	17.59 ± 25.89	36.21 ± 35.06	10.82 ± 23.21	6.81 ± 13.73	12.75 ± 2.62
		2nd	4.76	32.37	18.55 ± 26.55	34.11 ± 33.02	28.70 ± 35.42	8.30 ± 2.54	1.30 ± 0.95
		3rd	2.97	28.86	14.39 ± 25.04	28.66 ± 31.25	18.14 ± 29.84	3.53 ± 2.25	1.35 ± 0.96
		total	3.70 ± 0.94	30.82	16.78 ± 25.86	32.65 ± 33.06	18.90 ± 30.42	6.29 ± 8.01	5.86 ± 5.89
Black Cloth	On	1st	4.44	23.63	12.51 ± 20.07	9.81 ± 23.15	23.15 ± 30.18	11.09 ± 12.50	6.62 ± 2.49
		2nd	2.71	23.54	13.57 ± 19.26	8.87 ± 14.63	48.01 ± 28.18	9.93 ± 5.50	2.98 ± 1.86
		3rd	5.18	31.36	20.63 ± 23.64	27.53 ± 34.19	17.51 ± 30.91	16.55 ± 16.55	19.21 ± 1.81
		total	4.11 ± 1.27	26.47	15.57 ± 21.42	15.74 ± 27.20	29.92 ± 32.19	12.48 ± 8.70	9.41 ± 7.07
Colourful	Off	1st	4.55	35.78	21.19 ± 28.87	44.10 ± 29.32	32.31 ± 32.43	1.79 ± 1.91	3.25 ± 2.24
		2nd	5.86	20.70	8.59 ± 18.85	20.35 ± 29.35	3.72 ± 5.16	3.38 ± 1.79	1.08 ± 0.96
		3rd	5.53	31.85	22.94 ± 22.11	39.95 ± 31.19	15.57 ± 15.84	16.58 ± 2.75	12.76 ± 2.66
		total	5.31 ± 0.68	30.09	17.77 ± 24.29	34.84 ± 31.65	16.88 ± 22.65	7.53 ± 7.05	5.89 ± 5.58
Colourful	On	1st	5.31	41.98	37.39 ± 19.11	30.25 ± 25.88	35.91 ± 25.62	42.99 ± 2.07	43.05 ± 2.23
		2nd	5.53	33.38	19.59 ± 27.06	39.04 ± 36.25	17.64 ± 30.38	7.72 ± 3.17	11.00 ± 1.38
		3rd	5.52	33.25	18.50 ± 27.66	41.96 ± 30.52	22.73 ± 31.46	3.40 ± 2.01	1.48 ± 1.33
		total	5.45 ± 0.12	36.16	24.60 ± 26.52	37.11 ± 31.33	24.16 ± 30.35	17.16 ± 17.63	15.94 ± 16.96

Table D4. Trial-level and total performance of the YOLOv11n-pose tracking system for tilt estimation under varying background (Whiteboard, Black Cloth, Colourful) and lighting (Off, On) conditions. Speed is shown for each trial (1st/2nd/3rd) according to the measured SYS sampling rate, and as mean ± SD in the total row. RMSE and mean ± SD of absolute error (AE) are reported both for the entire repetition (Total) and for each movement phase (Closing, Opening, Pronated, Supinated).

Background	Lighting		Total			Closing	Opening	Pronated	Supinated
			Sampling rate [Hz] ($\mu \pm \sigma$)	RMSE [mm]	AE [mm] ($\mu \pm \sigma$)	AE [mm] ($\mu \pm \sigma$)	AE [mm] ($\mu \pm \sigma$)	AE [mm] ($\mu \pm \sigma$)	AE [mm] ($\mu \pm \sigma$)
Whiteboard	Off	1st	27.76	7.79	6.64 ± 4.1	6.6 ± 4.1	–	–	–
		2nd	27.78	18.9	11.0 ± 15.4	2.0 ± 1.5	3.6 ± 2.9	22.3 ± 17.8	–
		3rd	28.35	5.2	3.8 ± 3.5	2.7 ± 2.4	3.8 ± 2.5	7.4 ± 5.0	2.9 ± 2.3
		total	27.96 ± 0.335	12.6	7.2 ± 10.4	4.2 ± 3.7	3.7 ± 2.7	17.7 ± 16.5	2.9 ± 2.3
Whiteboard	On	1st	25.94	5.6	4.5 ± 3.3	5.1 ± 3.2	3.1 ± 2.8	5.7 ± 3.6	2.9 ± 2.1
		2nd	26.92	5.1	4.2 ± 3.0	2.7 ± 2.1	3.8 ± 2.7	5.4 ± 2.9	5.3 ± 3.1
		3rd	24.02	5.4	3.8 ± 3.9	2.3 ± 2.0	2.8 ± 1.9	6.2 ± 5.0	2.6 ± 2.3
		total	25.63 ± 1.48	5.4	4.2 ± 3.4	3.2 ± 2.7	3.2 ± 2.5	5.8 ± 4.1	3.7 ± 2.8
Black Cloth	Off	1st	–	–	–	–	–	–	–
		2nd	25.7	6.4	4.7 ± 4.3	3.2 ± 3.2	4.3 ± 3.7	7.4 ± 4.9	4.6 ± 4.4
		3rd	24.71	8.7	6.0 ± 6.3	3.9 ± 4.3	6.9 ± 6.3	8.6 ± 8.1	5.6 ± 4.7
		total	25.21 ± 0.70	7.4	5.2 ± 5.3	3.5 ± 3.7	5.4 ± 5.1	7.9 ± 6.6	5.0 ± 4.5
Black Cloth	On	1st	17.84	15.7	11.1 ± 11.2	6.8 ± 5.9	5.4 ± 3.5	22.1 ± 13.5	–
		2nd	21.66	5.5	4.3 ± 3.5	3.2 ± 2.4	5.1 ± 3.5	3.5 ± 3.1	3.1 ± 7.2
		3rd	23.15	6.3	4.8 ± 4.1	3.3 ± 2.3	4.2 ± 3.0	6.2 ± 5.1	6.7 ± 5.2
		total	20.88 ± 2.74	10.8	7.0 ± 8.2	5.0 ± 3.4	5.0 ± 3.4	11.0 ± 12.1	6.9 ± 4.9
Colourful	Off	1st	20.37	11.5	7.0 ± 9.2	3.1 ± 3.4	2.2 ± 1.7	15.5 ± 11.5	8.2 ± 10.0
		2nd	28.18	7.3	5.4 ± 4.9	4.2 ± 3.4	5.7 ± 5.3	7.7 ± 6.3	5.1 ± 4.2
		3rd	–	–	–	–	–	–	–
		total	24.28 ± 5.52	10.3	6.4 ± 8.0	3.5 ± 3.5	3.4 ± 3.7	12.6 ± 10.6	7.4 ± 9.0
Colourful	On	1st	21.79	8.3	5.3 ± 6.4	3.4 ± 4.1	1.4 ± 1.0	10.5 ± 8.4	3.8 ± 2.6
		2nd	28.21	7.3	4.6 ± 5.7	1.8 ± 1.4	6.3 ± 4.4	2.1 ± 1.7	9.8 ± 7.9
		3rd	25.21	6.4	4.4 ± 4.6	4.7 ± 5.5	4.9 ± 4.9	3.5 ± 4.0	4.2 ± 3.0
		total	25.07 ± 3.21	7.5	4.9 ± 5.8	3.3 ± 4.2	3.8 ± 3.8	6.7 ± 7.4	5.8 ± 5.8

Table D5. Trial-level and total performance of the Colour-tracker system for distance estimation under varying background (Whiteboard, Black Cloth, Colourful) and lighting (Off, On). RMSE and absolute error (AE) are reported per trial; AE is shown as mean ± standard deviation. Sampling-rate cells are left blank.

Factor	df_1	df_2	F	p	η_p^2
Background	2	3320.02	104.58	9.0×10^{-45}	0.0499
Lighting	1	3497.19	299.66	1.8×10^{-64}	0.0480
Movement	3	1500.07	17.96	1.9×10^{-11}	0.0071

Table D6. Welch's ANOVA on the YOLO System's absolute distance error for the factors Background, Lighting, and Movement.

Factor	df_1	df_2	F	p	η_p^2
Background	2	5068.40	26.65	3.06×10^{-12}	0.0069
Lighting	1	7633.12	0.04	0.8395	0.0000054
Movement	3	3315.48	439.02	3.66×10^{-240}	0.1629

Table D7. Welch's ANOVA on the YOLO System's absolute tilt error for the factors Background, Lighting, and Movement.

Factor	df_1	df_2	F	p	η_p^2
Background	2	9.32	5.09	0.0320	0.4750
Lighting	1	15.74	0.025	0.877	0.0015

Table D8. Welch's ANOVA on the YOLO system's detection frequency for the factors Background and Lighting.

Factor	df_1	df_2	F	p	η_p^2
Background	2	19142.46	42.04	6.07×10^{-19}	0.0028
Lighting	1	21223.24	118.70	1.44×10^{-27}	0.0044
Movement	3	12564.60	806.68	0	0.0845

Table D9. Welch's ANOVA on the Colour Tracker absolute distance error for the factors Background, Lighting, and Movement.

Factor	df_1	df_2	F	p	η_p^2
Background	2	17080.67	72.48	4.54×10^{-32}	0.0044
Lighting	1	29388.78	2.23	0.135	6.62×10^{-5}
Movement	3	13110.80	1137.13	0	0.106

Table D10. Welch's ANOVA on the Colour Tracker absolute tilt error for the factors Background, Lighting, and Movement.

Factor	df_1	df_2	F	p	η_p^2
Background	2	7.843	3.660	0.0754	0.311
Lighting	1	14.974	2.727	0.119	0.151

Table D11. Welch's ANOVA on the Colour Tracker detection frequency for the factors Background and Lighting.

E. Machine Learning Experiment Pictures

Figure E1. White background, light off

Figure E2. White background, light on

Figure E3. Black background, light off

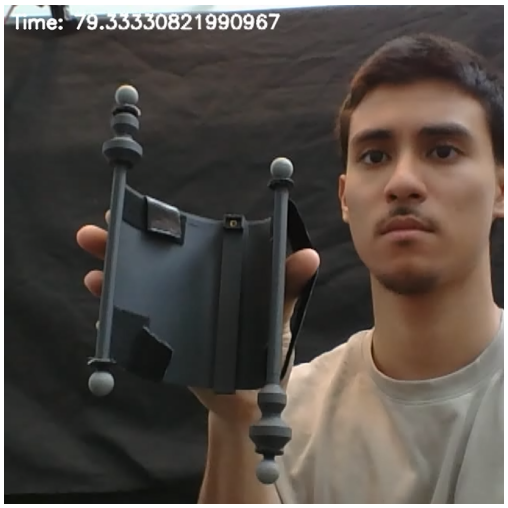

Figure E4. Black background, light on

Figure E5. Coloured background, light off

Figure E6. Coloured background, light on

F. Colour Tracking Experiment Pictures

Figure F1. White background, light off

Figure F2. White background, light on

Figure F3. Black background, light off

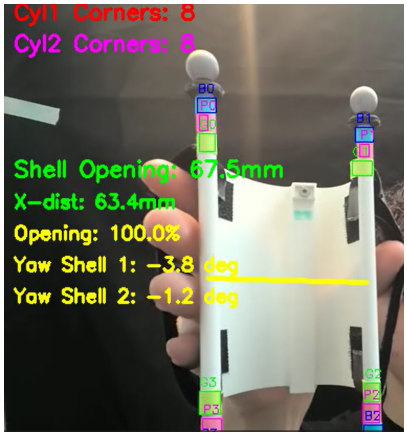

Figure F4. Black background, light on

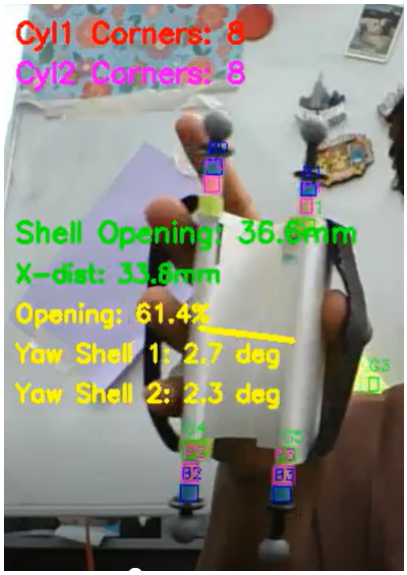

Figure F5. Coloured background, light off

Figure F6. Coloured background, light on

G. Additional graph Colour Detection

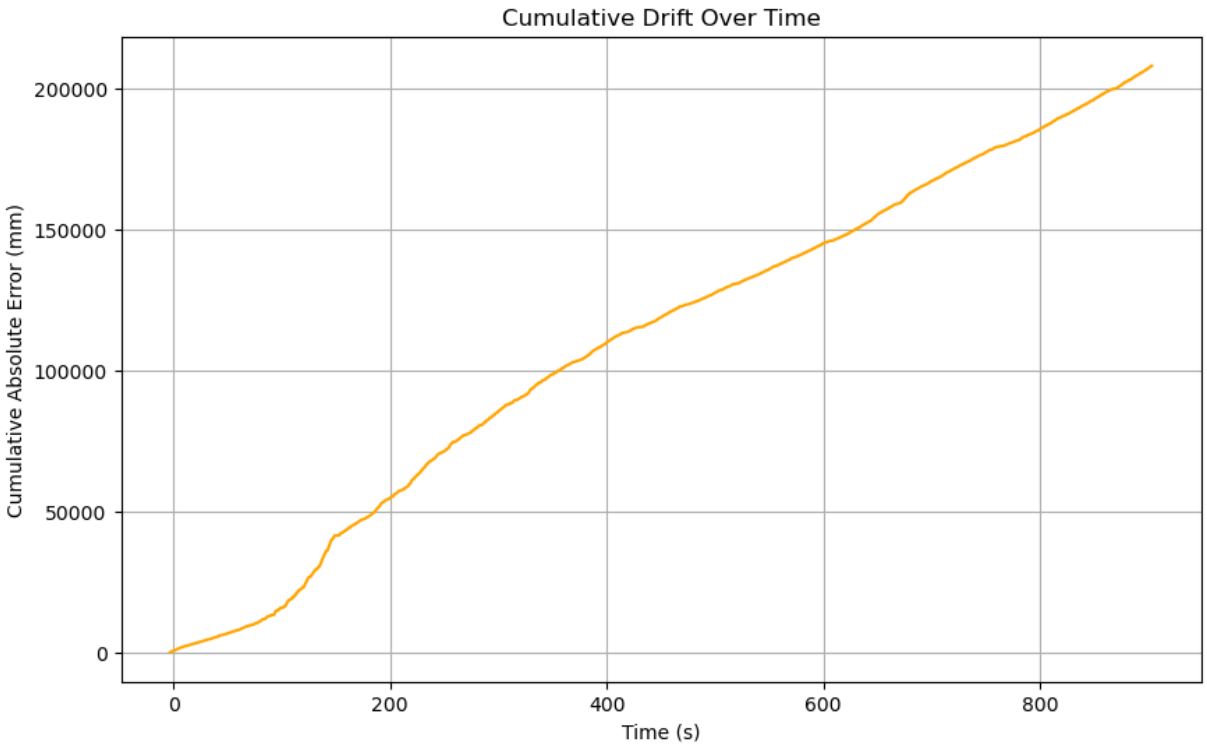

Figure D7. The cumulative absolute error of the distance between the two shell ends during a 15 minute trial with the colour tracking model. The participant held the shell straight in front of the camera and performed an opening and closing movement during the whole take.