Evaluation of Perception Latencies in a Human-Robot Collaborative Environment

Atle Aalerud[1] and Geir Hovland[1]

[1]Mechatronics Group, Faculty of Engineering and Science, Department of Engineering Science, University of Agder, N-4898 Grimstad, Norway. atle.aalerud@uia.no

Abstract **– The latency in vision-based sensor systems used in human-robot collaborative environments is an important safety parameter which in most cases has been neglected by researchers. The main reason for this neglect is the lack of an accurate ground-truth sensor system with a minimal delay to benchmark the vision-sensors against. In this paper the latencies of 3D vision-based sensors are experimentally benchmarked and analyzed using an accurate laser-tracker system which communicates on a dedicated EtherCAT channel with minimal delay. The experimental results in the paper demonstrate that the latency in the vision-based sensor system is many orders higher than the latency in the control and actuation system.**

D.1 Introduction

During the past few years human-robot interaction (HRI) and human-robot collaboration (HRC) have gained a vast amount of interest among researchers and industry world wide, [1]. The review paper [1] considers several vision-based safety systems for human-robot collaboration, but the important parameter of sensor latency is not addressed in any detail. Analysis of delays in collision avoidance and real-time motion planning is identified as an area which needs more work. However, sensor latency is not mentioned, ie. *More specifically, this implies latencies between control systems and delays in motion due to heavy robots with large inertia.*

In [2] it was stated: A key requirement for the robustness of safety systems for human-robot collaboration is low latency. This means that the amount of time necessary to perform a complete cycle of the safety system, from sensing to robot speed adjustment, must be very low. Figure D.1 shows the total latency in such a system. The sensors must be taken into account in addition to the control and actuation latency.

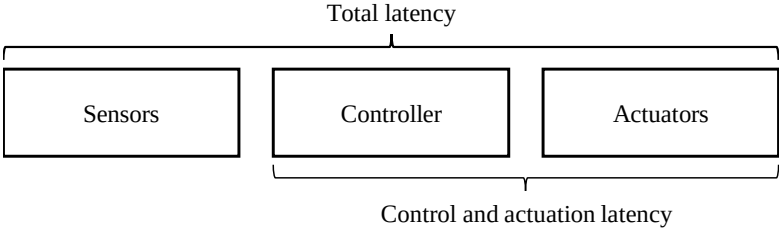


Figure D.1: Total latency in a control system including sensors, controller and actuators.

In [3] a comprehensive report on performance metrics used in collaboration between humans and industrial robots was presented. Typical standard metrics used are such as false negatives (FN), true positives (TP), false positive rate (FPR), etc. In addition to the traditional performance metrics for human detection, a new set was defined: Detection Area vs Safety Coverage Area, Ground-Truth Radius Estimate, False Clear Area/ False Occupied Area, Occlusions and Time Projection for Assumed Reaction Time. Normally, the latency of the sensors has not been evaluated as being part of the "Time Projection for Assumed Reaction Time". The main reason for this exclusion is, in most cases, the lack of an accurate ground-truth system for evaluating the sensors performance. In this paper, the sensor latency is experimentally evaluated in a multiple-sensor human-robot collaborative environment. To the authors' knowledge, this paper presents for the first time an evaluation, benchmark and detailed analysis of vision-sensor latency in a HRC environment.

In [2] the latency of the control system, similar to Time Projection for Assumed Reaction Time defined in [3] was evaluated. A PhaseSpace motion capture system was utilized, but the latency of the sensor system was not included in the evaluation.

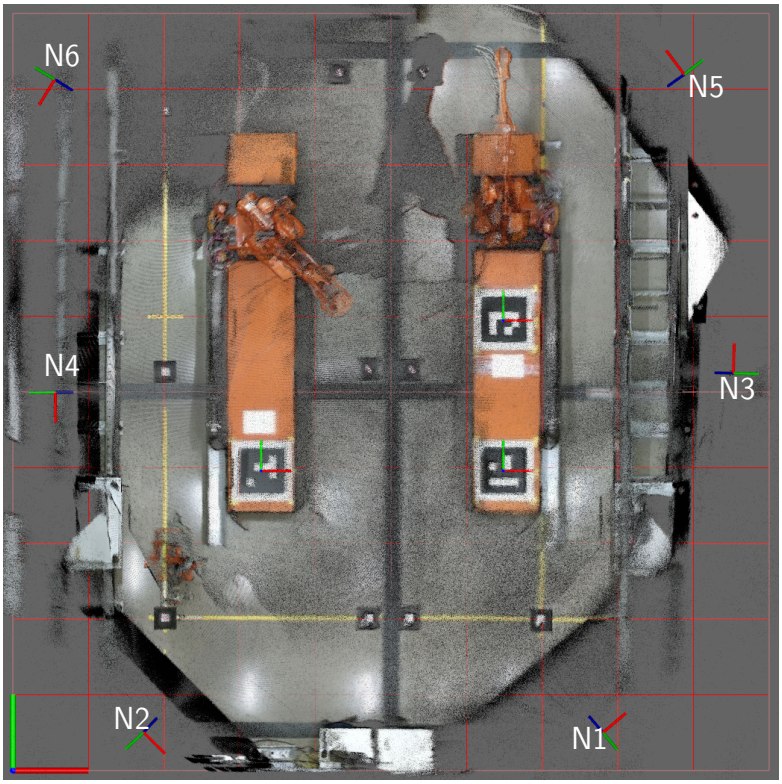


Figure D.2: Overhead view of the combined accumulated RGB point clouds of all six sensors after calibration. Sensor node poses are annotated N1-N6. The sensors are located 4.3 m above the floor and are angled approx. 60° from horizontal. The origin of the world frame is located at the bottom-left, where the red and green lines represent the X and Y-directions, respectively.

In [4] the reaction time including delays in the sensor system is addressed. In that paper it is stated: *To measure the reaction time, it is necessary to give a controlled signal*

to the safety system and then measure the latency of the response. For camera-based systems, the stimulus may be a sequence of recorded images showing a person entering a protected space, while laser-based systems may require the use of a controlled physical avatar that can be moved into the work zone using repeatable trajectories. Ultimately, the tester should be able to provide both a ground truth and control triggering of the event.

The approach outlined in [4] is used in this paper. Instead of a control triggering of the event, the ground truth system used has both the desirable properties of high accuracy and minimal delay. The latency of the sensors is found by post-processing and comparison of the measured data with the ground truth system. The paper is organized as follows: Section E.3 presents the experimental setup, section E.4 presents the experiments, while the main results are summarized in F.5. Discussion and conclusions follow in section E.6.

D.2 Framework

D.2.1 Robots

The two robots used in the experiments are of type ABB IRB4400 with 60kg payload capacity each mounted on a 5m linear track motion. The tracks are mounted in parallel such that the robots are 3.88 m apart.

D.2.2 Sensors/Perception

The sensors used are 6 Kinect v2 encapsulated with local processing power using an Nvidia Jetson TX2 in each sensor node. A central computer with Intel i7-7820x 3.6GHz CPU runs the Robot Operating System (ROS) communicating with the sensor nodes over TCP/IP. Previous work using this facility can be found in among others [5], [6]. The sensors N1-N6 are calibrated automatically using the method described in [7].

D.2.3 Ground truth annotation and evaluation

In order to evaluate sensor performance, including both accuracy and latency, the availability of an accurate ground-truth sensor system is essential. Figure D.3 shows a Leica laser tracker AT960 (left) and a Spherically Mounted Retroreflector (SMR, right) used as a ground truth system in this paper. The specified accuracy of the laser tracker is 15 µm + 6 µm/m which is several orders of magnitude more accurate than the N1-N6 sensors to be benchmarked.

Measurements from the laser tracker were recorded with a dedicated network card in the PC using the IgH EtherCAT master for Linux and the laser tracker controller as a single slave. The delay a message would experience over an EtherCAT network is much smaller than the delay in a standard Ethernet network, especially with a single master-slave EtherCAT network. According to [8] the constant node delay in EtherCAT is typically below 0.5 µs, which is many orders of magnitude lower than typical delays in a corresponding Ethernet network. In the experiments with the robots the SMR was located at the tool centre point (TCP) while in the experiments with the human the SMR was

pressed against the xiphoid process. As the human body was facing negative Y-direction during experiments, an offset of 10 cm was applied in Y-direction of the tracker data to approximate the body center position.

(**a**) (**b**)

Figure D.3: Leica tracker AT960 (**a**) and Leica SMR Ball Probe (**b**)

D.3 Experiments

In this paper two different experiments were performed.

- I) Timestamped measurements from the laser tracker compared with the robot controller position to estimate the control and actuation latency. In addition the laser tracker measurements were compared with the tool position to estimate overall perception latency. A sword of length 1.2 m was used as a tool to make it easier to extract and separate the point-cloud data of the tool from the robot.
- II) A human performed a walking test. The ground-truth was measured with the SMR located on the chest and the human position was estimated using depth measurements from a region of interest (ROI) detected using YOLOv3-tiny (You Only Look Once) [9]. The selected human movement velocities were 0.5 m/s, 1.0 m/s and 2.0 m/s which are inside the worst-case maximum operator speed defined in ISO 13855, [10], see for example [4].

D.3.1 Test Case I

Figure D.4 shows the measurements used to evaluate the latencies for both test cases I) and II). The linear track motion was moved a distance of 3 m back and forth, with increasing velocities, while the robot itself was standing still. Measurements from all three sensors systems (ground-truth, robot internal, and N1-N6 sensors) are shown in Figure D.4. Figure D.5 shows a zoomed-in portion of Figure D.4.

The Kinect N1-N6 sensors generate point-cloud data which is transformed into an Octree with cuboid size of 4 cm and a compression/decompression scheme. For more details about this implementation see [6]. The discrete jumps in the data of 4 cm in

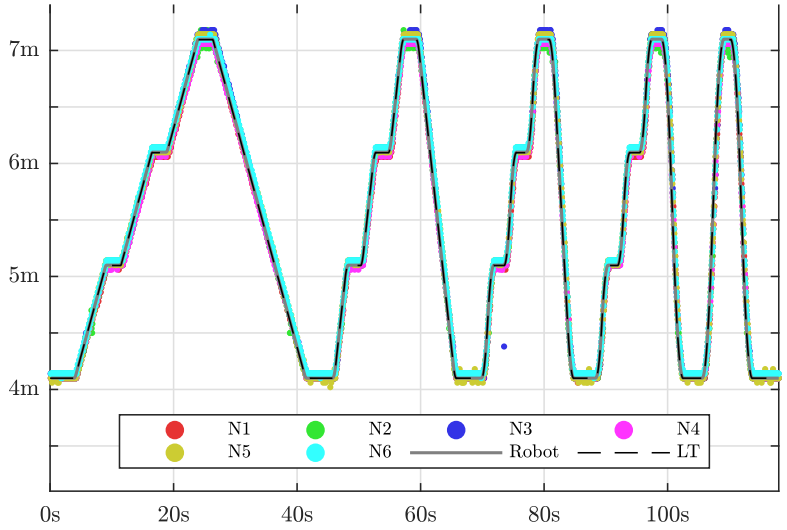
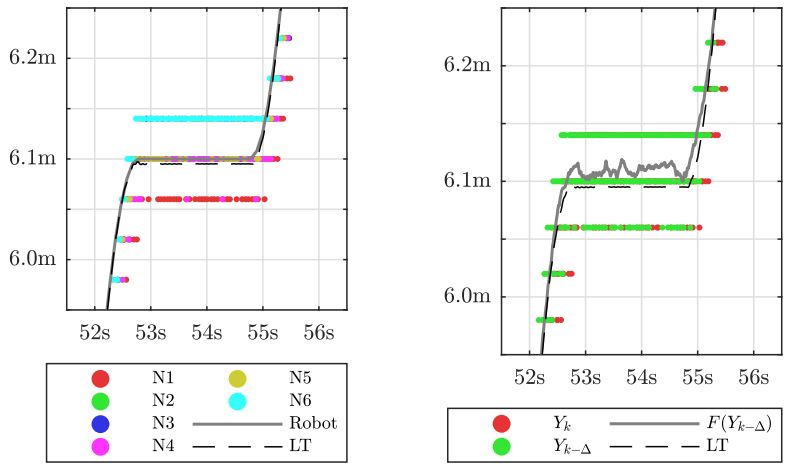


Figure D.4: Robot trajectory in Y-direction measured by the six sensors N1-N6, trajectory via robot and track resolvers and laser tracker.



(**a**) (**b**)

Figure D.5: (**a**) Selected section of Y-Position measured by 6 sensor nodes, robot position via resolvers and laser tracker. (**b**) Concatenated N1-N6 (red), the same with latency offset Δ (green), concatenated and offset values with moving average filter (grey) and laser tracker (dashed).

Figure D.5 correspond to the size of the smallest cuboids in the Octree. In this experiment the average delay of the robot controller and actuation system was found to be 2.8 ms while the average delay of the Kinect sensors and the Octree compression/decompression scheme was found to be 161 ms. The curve representing the average sensor position, grey in Figure D.5(**b**), has a best-case resolution of 4 cm / 6 sensors = 0.67 cm when all the 6 sensors have a non-obstructed line-of-sight to the tool. The delay Δ was found by interpolation of the measured data to a common time-vector with resolution of 0.1 ms and then optimizing for the smallest mean-square-error (MSE) value between the time-shifted sensor data and the ground truth system, ie.

$$\Delta = \arg\min_{\delta} \frac{1}{N} \sum_{k=1}^{N} (Y_{k-\delta} - LT_k)^2 \tag{D.1}$$

where $Y_{k-\delta}$ is the time-shifted data and LT_k is the corresponding time-sample from the laser tracker.

Figure D.6 shows the voxel maps from the 6 sensors overlaid on the robot model at zero velocity, while Figure D.7 shows the same at a velocity of 1.5 m/s. It is clearly evident from these two figures that the voxel maps from sensors N1-N6 lag behind the ground-truth at higher velocities. The delay measured in distance depends on the robot's velocity, while the delay in time was constant.

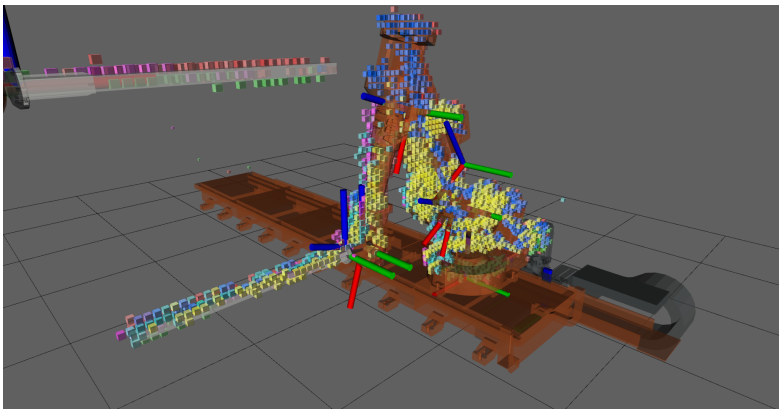


Figure D.6: 6 voxel maps overlaid on robot model at zero velocity.

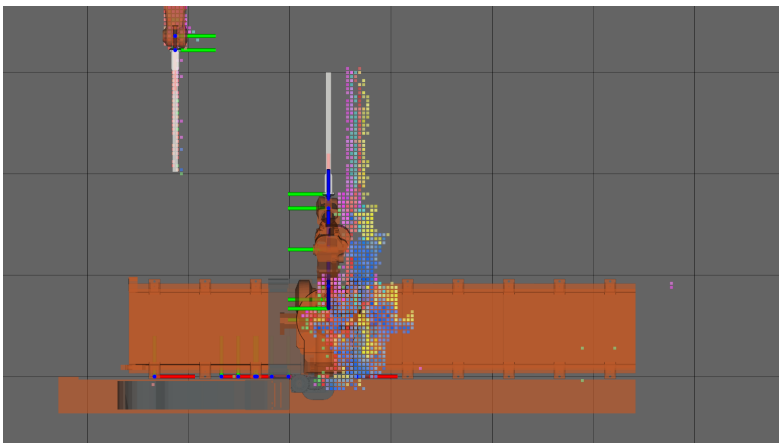


Figure D.7: 6 voxel maps on robot model at velocity 1.5 m/s

D.3.2 Test Case II

In the second test case a human placed the SMR on the chest (xiphoid process) and walked through the cell with varying velocities. The velocity was controlled by stepping on 0.5 m separated markers at the audible beat of a metronome. In this way, local velocity variations due to acceleration could occur, but the average velocity would still be correct. The human's positions were measured with the laser tracker in addition to YOLO-selected depth measurements using the 6 Kinect sensors.

Figure D.8 shows the walked human path in the XY plane at velocity approximately 0.5 m/s. When moving in the negative Y-direction in the figure the human was walking forwards, while he was walking backward in the positive Y-direction, to maintain a non-obstructed line-of-sight with the laser tracker. Figure D.9 shows the same data vs. time in both the X- and Y-directions. Figure D.10 shows a zoomed-in version of Figure D.9. The latency of the fused data from the 6 sensor nodes compared to the ground-truth measurement system was found to be 333 ms.

Figure D.11 shows the results from a second experiment where the human doubled the velocity to 1.0 m/s. Figure D.12 shows the same data vs. time in both the X- and Y-directions. Figure D.13 shows a zoomed-in version of Figure D.12. The latency of the fused data from the 6 sensor nodes compared to the ground-truth measurement system was found to be 333 ms, the same as for the experiment performed at 0.5 m/s. At some locations in Figure D.12 the laser tracker beam was obstructed as shown in the figure by the steps in the measurement between time 10 s and 13 s.

Figure D.15 shows the results from a third experiment where the human again doubled the velocity to 2.0 m/s. Figure D.15 shows the same data vs. time in both the X- and Y-directions. Figure D.16 shows a zoomed-in version of Figure D.15. At this velocity the human was not able to follow the same path as in the previous two experiments when walking backward, so a different path was taken. The laser tracker beam was obstructed, as shown in Figure D.15, just after time 13 s. As in the previous two experiments the latency of the fused data from the 6 sensor nodes compared to the ground-truth measurement system was found to be 333 ms.

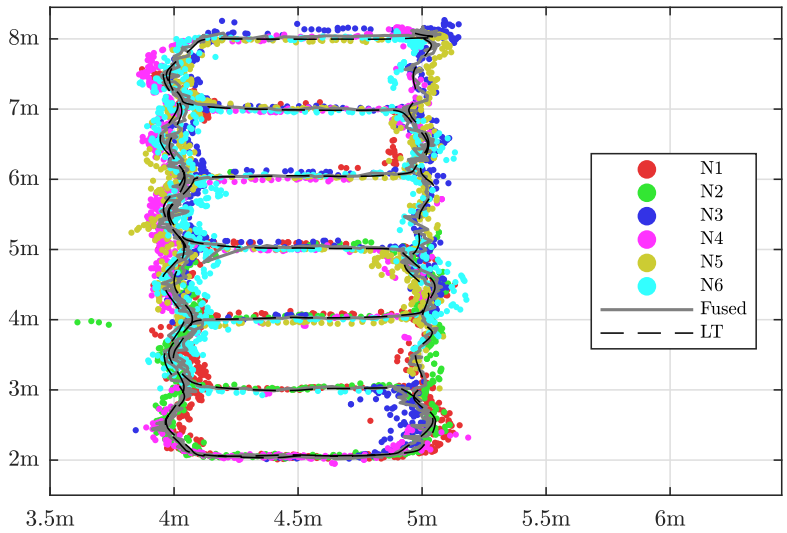


Figure D.8: Walked human path in XY plane at velocity 0.5 m/s using six sensor nodes N1-N6 and depth/YOLO (colored), averaged path (grey) and laser tracker.

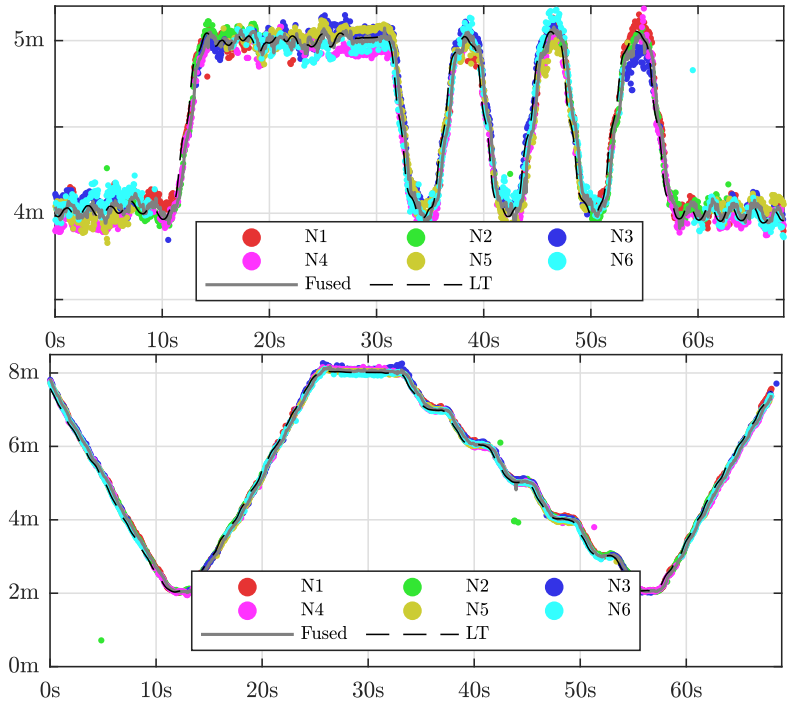
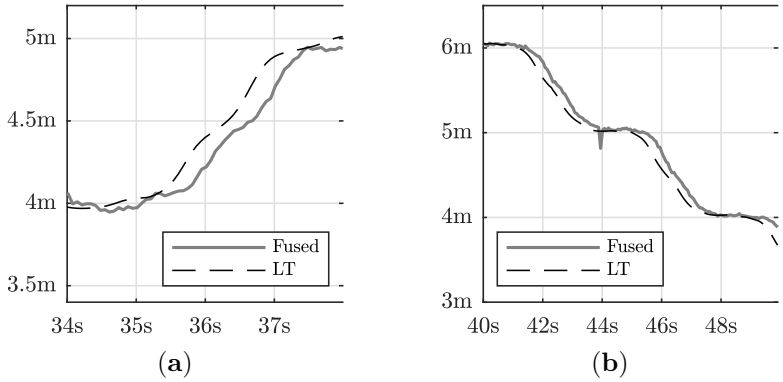


Figure D.9: Walked human path in the X-direction (top) and Y-direction (bottom) vs. time at velocity 0.5 m/s using six sensor nodes N1-N6 and depth/YOLO (colored), averaged path (grey) and laser tracker.



(**a**) (**b**)

Figure D.10: Zoomed in versions of Figure D.9 without N1-N6 sensor nodes. (**a**) X-direction (**b**) Y-direction

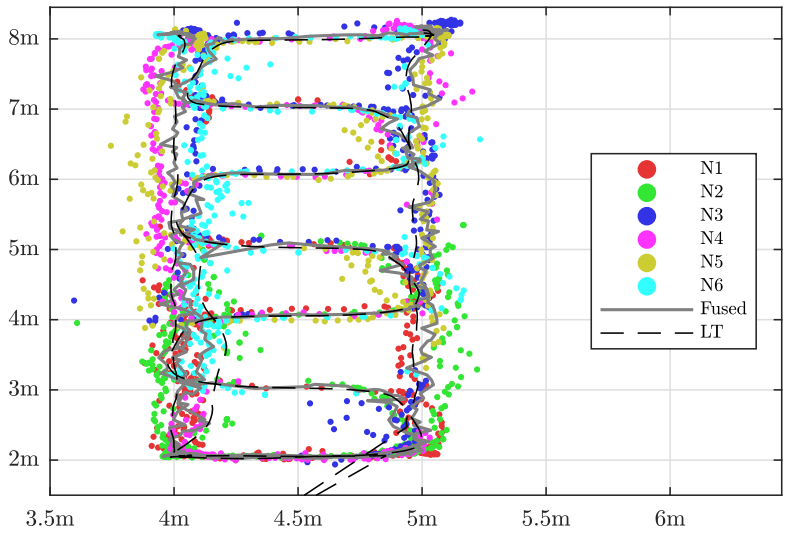


Figure D.11: Walked human path in XY plane at velocity 1.0 m/s using six sensor nodes N1-N6 and depth/YOLO (colored), averaged path (grey) and laser tracker.

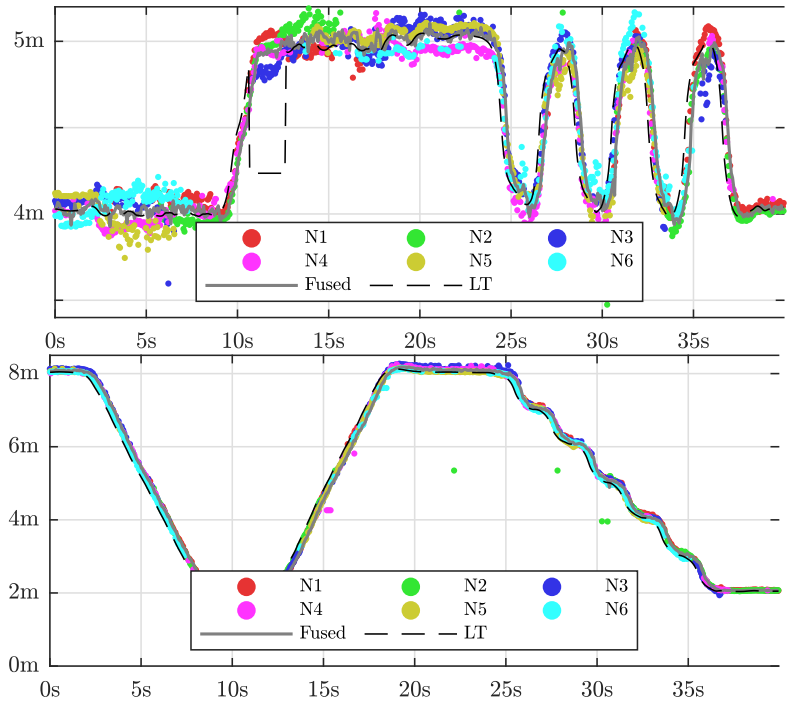
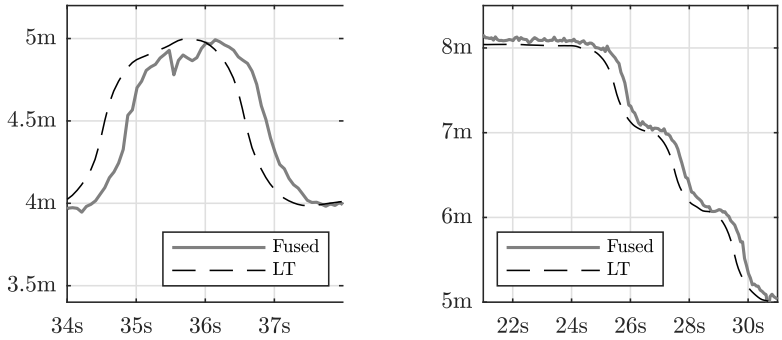


Figure D.12: Walked human path in the X-direction (top) and Y-direction (bottom) vs. time at velocity 1.0 m/s using six sensor nodes N1-N6 and depth/YOLO (colored), averaged path (grey) and laser tracker.



(**a**) (**b**)

Figure D.13: Zoomed in versions of Figure D.12 without N1-N6 sensor nodes. (**a**) X-direction (**b**) Y-direction

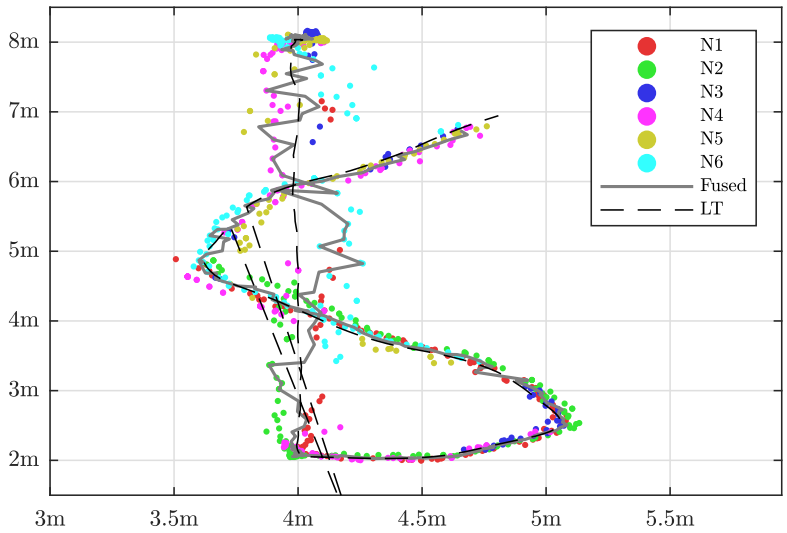


Figure D.14: Walked human path in XY plane at velocity 2.0 m/s using six sensor nodes N1-N6 and depth/YOLO (colored), averaged path (grey) and laser tracker.

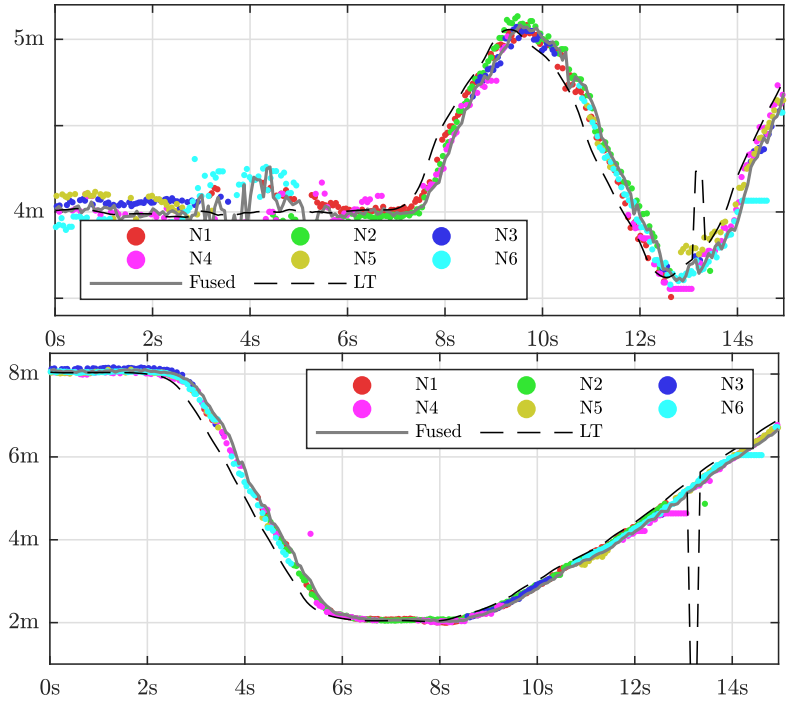
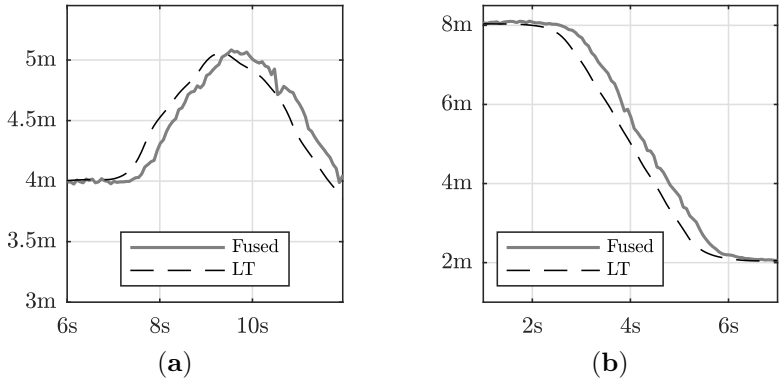


Figure D.15: Walked human path in the X-direction (top) and Y-direction (bottom) vs. time at velocity 2.0 m/s using six sensor nodes N1-N6 and depth/YOLO (colored), averaged path (grey) and laser tracker.



(**a**) (**b**)

Figure D.16: Zoomed in versions of Figure D.15 without N1-N6 sensor nodes. (**a**) X-direction (**b**) Y-direction

D.4 Results

Table D.1 summarizes the latency results obtained in this paper for the different cases. The robot controller and actuation systems show an average latency of 2.8 ms. The Octree implementation which utilizes a compression/decompression scheme when the nodes N1-N6 communicate with the central server shows an average latency of 161 ms, of which approximately 50 ms is required by the compression/decompression and a maximum of 20 ms internally in the Kinect (image to USB). The remaining 91 ms is assumed to be required by the Kinect driver and data transmission. For the experiments with the human detection and depth/YOLO the average latency was found to be 333 ms for all three velocities tested (0.5 m/s, 1.0 m/s and 2.0 m/s). The total depth/YOLO latency can be divided into the internal sensor latency (20 ms), driver, detection and transmission (246 ms) and accumulation and averaging (67 ms). Possible small variations in detection and transmission were compensated by the accumulation and averaging as the latter ran at 15 Hz.

Table D.1: Sensor latencies in the two cases. The compressor/decompressor latency is broken down into sub-categories. * are estimates.

Type	**Latency**
Robot Control and Actuation Latency	2.8ms
Compression/Decompression Case (total)	161 ms
Kinect (image to USB, from spec.)	20 ms
*Driver and transmission	91 ms
Compressor/Decompressor Algorithms	50 ms
depth/YOLO Case (total)	333 ms
Kinect (image to USB, from spec.)	20 ms
*Driver, detection and transmission	246 ms
Accumulation and averaging	67 ms

Table D.2 summarizes the achieved accuracies (mean value and standard deviation). The standard deviation increases with the velocity and the maximum mean error was found to be 54 mm at velocity 2.0 m/s.

Table D.2: Mean position error [mm] and sample standard deviation for the three human test cases, in X-, Y- and distance directions.

	$\overline{e_x}/s_x$	$\overline{e_y}/s_y$	$\overline{e_d}/s_d$
0.5 m/s	1 / 22	10 / 24	29 / 18
1.0 m/s	0 / 39	18 / 29	44 / 28
2.0 m/s	13 / 54	19 / 41	54 / 47

D.5 Discussion and Conclusions

In this paper the latency of 3D sensors used in a human-robot collaborative environment has been benchmarked and extensive experimental results have been presented. It was found that the latency of the sensor system was significantly larger than the latency of the control and actuation system, by more than a factor of 100 in the case of human movement and position detection. In most previous work in the literature the latency of the sensor system has been neglected while the latency of the control and actuation system has been taken into account. The main reason why sensor latency in most cases has not been addressed, is the lack of a ground-truth sensor system with minimal delay to benchmark against. In this paper a sensor system using EtherCAT communication with minimal delay and high accuracy has been used to establish the ground-truth. An alternative approach based on system timestamps cannot measure the full latency from acquisition to position as internal latency of perception sensors cannot be included. The work presented in this paper demonstrates that the sensor latency can be the dominating factor and should, in general, be given more attention in safe and collaborative human-robot environments.

In [4] the following system safety is outlined: *Operator safety is maintained via an external observer system integrated into a robotic workcell. The observer system monitors the regions surrounding the robot, and issues commands to slow or stop the robot as humans ("operators") approach. If the distance between the robot and a detected operator (i.e., the separation distance S) is less than the value of S at time t_0, the safety system initiates a safety-rated, controlled stop. The robot may then resume moving once the separation distance is greater than S. The minimum protective distance for stationary machine tools outlined in ISO 13855 is:*

$$S = vT + C \tag{D.2}$$

In Equation (D.2)*, v is the approach speed of human body parts. The value T is the system stopping performance, and is the combination of the time between sensing and actuation (i.e., T_R) and the response time of the machine (i.e., T_S). C is a constant.*

This paper focused on the latency of a vision system which was shown to be the main contributor to T_R described above. Future work by the authors will handle system safety in a similar fashion as the one outlined in [4], and in addition extending this approach to also handle moving robots, by incorporating the estimated latencies into the collision detection and avoidance control system. When all the latencies in the system are known, from sensors to actuation, adaptive, velocity-dependent primitives like spheres, boxes and cylinders could be used, surrounding the robotic equipment. The adaptive sizing of the primitives would contain both the current velocities of the human and the robots, as well as the robots stopping time, T_S. The avoidance system will be triggered if a human would be detected inside the adaptive primitives.

Acknowledgment

The research presented in this paper has received funding from the Norwegian Research Council, SFI Offshore Mechatronics, project number 237896.

References

[1] R.-J. Halme, M. Lanz, J. Kämäräinen, R. Pieters, J. Latokartano, and A. Hietanen, "Review of vision-based safety systems for human-robot collaboration," *Procedia CIRP*, vol. 72, pp. 111–116, 2018. doi: 10.1016/j.procir.2018.03.043

[2] P. A. Lasota, G. F. Rossano, and J. A. Shah, "Toward safe close-proximity human-robot interaction with standard industrial robots," in *2014 IEEE International Conference on Automation Science and Engineering (CASE)*, vol. 2014-Janua. IEEE, 8 2014. doi: 10.1109/CoASE.2014.6899348. ISBN 978-1-4799-5283-0. ISSN 21618089 pp. 339–344.

[3] M. O. Shneier, T. Hong, G. Cheok, K. Saidi, and W. Shackleford, "Performance Evaluation Methods for Human Detection and Tracking Systems for Robotic Applications," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep., Mar. 2015. doi: 10.6028/NIST.IR.8045

[4] J. A. Marvel and R. Norcross, "Implementing speed and separation monitoring in collaborative robot workcells," *Robotics and Computer-Integrated Manufacturing*, vol. 44, pp. 144–155, 2017. doi: 10.1016/j.rcim.2016.08.001

[5] A. Aalerud, J. Dybedal, E. Ujkani, and G. Hovland, "Industrial Environment Mapping Using Distributed Static 3D Sensor Nodes," in *2018 14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA)*. IEEE, 7 2018. doi: 10.1109/MESA.2018.8449203. ISBN 978-1-5386-4643-4 pp. 1–6.

[6] J. Dybedal, A. Aalerud, and G. Hovland, "Embedded Processing and Compression of 3D Sensor Data for Large Scale Industrial Environments," *Sensors*, vol. 19, no. 3, p. 636, 2 2019. doi: 10.3390/s19030636

[7] A. Aalerud, J. Dybedal, and G. Hovland, "Automatic Calibration of an Industrial RGB-D Camera Network Using Retroreflective Fiducial Markers," *Sensors 2019, Vol. 19, Page 1561*, vol. 19, no. 7, p. 1561, 3 2019. doi: 10.3390/S19071561

[8] G. Prytz, "A performance analysis of EtherCAT and PROFINET IRT," in *2008 IEEE Intl. Conf. Emerging Technologies and Factory Automation*. Hamburg: IEEE, 8 2008. doi: 10.1109/ETFA.2008.4638425 pp. 408–415.

[9] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," 4 2018. arXiv: 1804.02767

[10] *Safety of Machinery – Positioning of Safeguards with Respect to the Approach Speeds of Parts of the Human Body*, International Organization for Standardization Std. ISO 13 855:2010, 2010.