




Digital Robot Judge

Building a Task-centric Performance Database of Real-World Manipulation With Electronic Task Boards

By Peter So , Andriy Sarabakha , Fan Wu ,
Utku Culha , Fares J. Abu-Dakka ,
and Sami Haddadin 

Robotics aims to develop manipulation skills approaching human performance. However, skill complexity is often over- or underestimated based on individual experience, and the real-world performance gap is difficult or expensive to measure through in-person competitions. To bridge this gap, we propose a compact, Internet-connected, electronic task board to measure manipulation performance remotely; we call it the digital robot judge, or “DR.J.” By detecting key events on the board through performance circuitry, DR.J provides an alternative to transporting equipment to in-person competitions and serves as a portable test and data-generation system that captures and grades performances, making comparisons less expensive. Data collected are automatically published on a web dashboard (WD) that provides a living performance benchmark that can visualize improvements in real-world manipulation skills of robot platforms over time across the globe.

In this article, we share the results of a proof-of-concept electronic task board with industry-inspired tasks used in an international competition in 2021 and 2022 to benchmark localization, insertion, and disassembly tasks. We present data from 10 DR.J task boards, describe a method for deriving the relative task complexity (RTC) from timing data, and compare robot solutions with a human performer. In the best case, robots performed $9\times$ faster than humans in specialized tasks but achieved only 16% of human speed across the full set of tasks. Finally, we present the modular design, instructions, and software needed to replicate the electronic task board or to adapt it to new use cases to promote task-centric benchmarking.

INTRODUCTION

Real-world robot demonstrations are important to convey a system’s capabilities and advancements to a wide audience. Organized robotics competitions provide an arena in which to focus on the development of new capabilities and serve to critically assess solutions with reproducible, well-defined problem statements [1]. Through carefully designed event rules, competition events motivate the robotics community to benchmark progress in hardware and software [2]. Historically, events like the DARPA Grand Challenge, AWS Picking Challenge, RoboCup, and RoCKIn competitions [3], [4], [5], [6] have occurred during in-person gatherings at conferences and operated under stressful conditions and limited development parameters. Benchmarking protocols, like [7], provide instructions for generating standardized, comparable results, but they can be slow to be published. Event organizers put great effort into preparing a fair competition, e.g., organizing an expert judging committee, drafting competition rules, and competing with paper presentations during large conferences, such as IEEE International Conference on Intelligent Robots and Systems (IROS), IEEE International Conference on Robotics and Automation (ICRA), or the European Robotics Forum (ERF).

While these competitions capture interest from a wide community, they often fail to push the state of the art. Logistical challenges, including high costs for participants to travel, complications with shipping robotic equipment, difficulties in reproducing the test environment, and limited development time at conferences, are major obstacles to overcome during in-person events [8]. Furthermore, conference competitions often generate a static list of winners and a compilation of videos that are difficult to search and compare with future solutions. A systematic

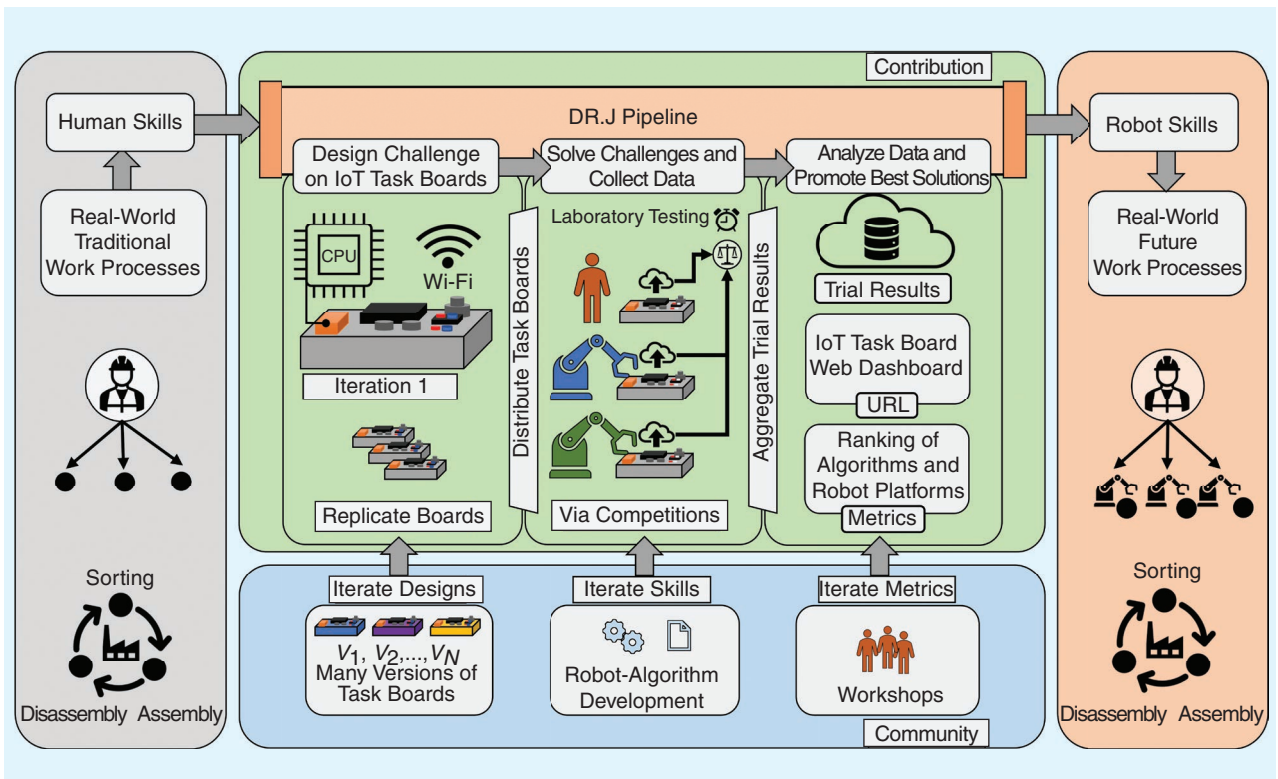


FIGURE 1. DR.J provides a time-based performance measurement platform for roboticists to demonstrate developments in manipulation skills with industry-inspired tasks from their own laboratories. Our proposed pipeline removes the need to collocate equipment to compare results and guides traditional work processes through a community-driven skill development cycle for future work processes powered by robots. The robotics community provides an initial performance benchmark of robot and human skills with a first set of metrics and set of competition task boards and then iterates on future designs to support new work processes through open innovation.v

way to compare outcomes across similar competition events would greatly benefit the robotics community.

To improve upon the stated issues of existing competitions and benchmarking protocols, we present a pipeline for benchmarking industry-inspired manipulation skills using an Internet-of-Things (IoT) electronic task board (see Figure 1), which remotely monitors trial attempts, verifies task completions, and records execution times through performance circuitry. This task board, paired with a WD, serves as a novel remote performance assessment platform for the robotics community.

Our goal with this platform is to provide an objective, task-centric performance data collection tool to simultaneously measure and compare real-world manipulation performance across multiple locations over the Internet. Furthermore, we see the electronic task board and DR.J concept as integral to enabling crowd-sourcing robotic solutions for industry-motivated challenges. The main contributions of our design are

- an electronic task board with a digital controller to automatically record and report performance data to a central server over the Internet
- a web platform for remotely measuring and aggregating decentralized manipulation performances
- a living dataset benchmarking the task execution timing data of 10 different robot platforms contrasted against human performance with a common trial protocol from two consecu-

tive international competitions showcased at automatica in 2021 and 2022.

The remainder of the article is organized as follows. In the “Related Work” section, we review the existing competition benchmarks in robot manipulation and remote performance assessment tools. In the “DR.J Components and System Design” section, we present the design of DR.J and share our implementation used in decentralized competitions. The “Estimating Task Complexity” section discusses task selection and approaches to estimating task complexity. The “DR.J Results in Decentralized Competition” section presents collected data from two iterations of competitions and discusses performances for multiple robot platforms. The “Discussion” section shares lessons learned and future work. Finally, the “Conclusion” section gives a summary and shares links to instructions to create your own task board.

RELATED WORK

We organize existing benchmarking approaches in robot manipulation into six categories: 1) simulation, 2) article and paper publications with video recordings, 3) Internet-connected public robot platforms, 4) in-person competitions, 5) standardized object sets; and 6) decentralized competitions. Each approach has pros and cons for conveying the performance of the robot system under test. To illustrate, Table 1 compares an example from each approach for the

presence of the following valued real-world benchmarking qualities: the use of real-world tasks, reproducibility in one’s own lab, cross-site compatibility, cycle time analysis, automated grading, and availability of results.

Simulation-based benchmarks are highly reproducible, offer digital submissions that can be automatically graded, and incur zero transportation costs. Numerous simulation environments, like MuJoCo, USARS, Webots, Robosuite, and others, allow robot programs to be tested without the cost of a physical robot platform [9], [10], [11], [12]. However, their utility is limited to the simulated world and often requires tedious adaptation and tuning to transfer results to a robot working in an industrial use case. For this reason, simulated robot solutions alone are discounted for their lack of robustness and the nonnegligible effort necessary to transfer to real-world scenarios [7].

Paper and article publications with video recordings have high integrity and archival value through the review and publication process, but it can be slow for results to be published. Internet-connected public robot platforms, like the Remote Robot Learning Lab [13] and OCRTOC [14], provide high reproducibility through the use of shared physical robot platforms; however, these systems have proven to be expensive to maintain and are often unavailable due to high-demand or maintenance activities. In-person competitions involve real-world tasks and have high grading integrity with onsite judges. However, these events require great effort and suffer from the previously mentioned logistical challenges [3], [4], [5].

Standard object sets, like the YCB Object Set [15], provide a common basis for comparing robot performance results with physical objects. However, these objects are passive and do not provide any information about the quality of the interaction. Additionally, sourcing the exact same items remains challenging, as the manufacturers and branding of objects change over time. Similarly, benchmarking-specific physical task boards, like the NIST Task Board [16], provide a task-centric approach and clearly defined rule set—which are great tools, especially in conjunction with in-person competitions—to compare the performance of robot platforms. However, they lack integrated sensing capabilities, and performance data are still manually reported, leading to a slow availability of results.

Conversely, DR.J builds on the merits of standard objects sets and existing benchmarking task boards with a streamlined reporting workflow (see Figure 2) that can be used both during in-person competitions and in individual laboratories in periods between competitions. Finally, decentralized competitions not based on simulation, like the Cybathlon [17], provide a framework for teams to reproduce exact test conditions independently in their own locations with specific rules and submission instructions to facilitate the cross-comparison of results. While it is difficult to prevent bad actors from circumventing the rules, decentralized competitions can reach wider audiences with the removed burden of travel.

Regardless of benchmarking event format or location, video recordings of robot demonstrations will remain important to document solutions. However, comparing one’s own work to recorded video performances remains difficult without a detailed description of the testing scenario, which is often not sufficiently documented. While video recordings capture a demo’s visual results, they do not capture the number of attempts required to reach the presented performance, nor do they provide a systematic searchable dataset for future comparison. Following the example of the Cybathlon, we have developed our own decentralized competition, the Robothon Grand Challenge, with an electronic task board and automated scoring system using DR.J.

DR.J COMPONENTS AND SYSTEM DESIGN

DR.J enables the construction of a historical task performance database for various robotic solutions to quantitatively assess performance improvements over time. Our data collection pipeline relies on the combination of sensorized task boards with web tools to provide searchable results and automatic reporting. The system architecture is shown in Figure 3. It comprises three interconnected components: 1) a physical electronic task board, 2) a clearly defined trial protocol with start and goal states, and 3) a WD and a cloud database. We used the following design goals for the electronic task board:

- to automatically detect task completions and transitions between start and goal states

TABLE 1. A comparison among state-of-the-art competitions in benchmarking of robotics.

PERFORMANCE BENCHMARK METHOD	REAL-WORLD TASKS	REPRODUCIBLE IN OWN LAB	CROSS-SITE COMPATIBLE	CYCLE TIME ANALYSIS	AUTOMATED GRADING	AVAILABILITY OF RESULTS
Robotbenchmark.net [12]	✗	✓	✓	✓	✓	✓
YCB Object Set [15]	✓	✓	✓	✗	✗	✗
Robot Learning Lab [13]	✓	✓	✗	✗	✗	✗
DARPA Robotics Challenge [3]	✓	✗	✗	✓	✗	✓
NIST Assembly Task Boards [16]	✓	✓	✓	✓	✗	✗
Cybathlon [17]	✓	✓	✓	✓	✗	✓
Internet-Connected Task Board	✓	✓	✓	✓	✓	✓

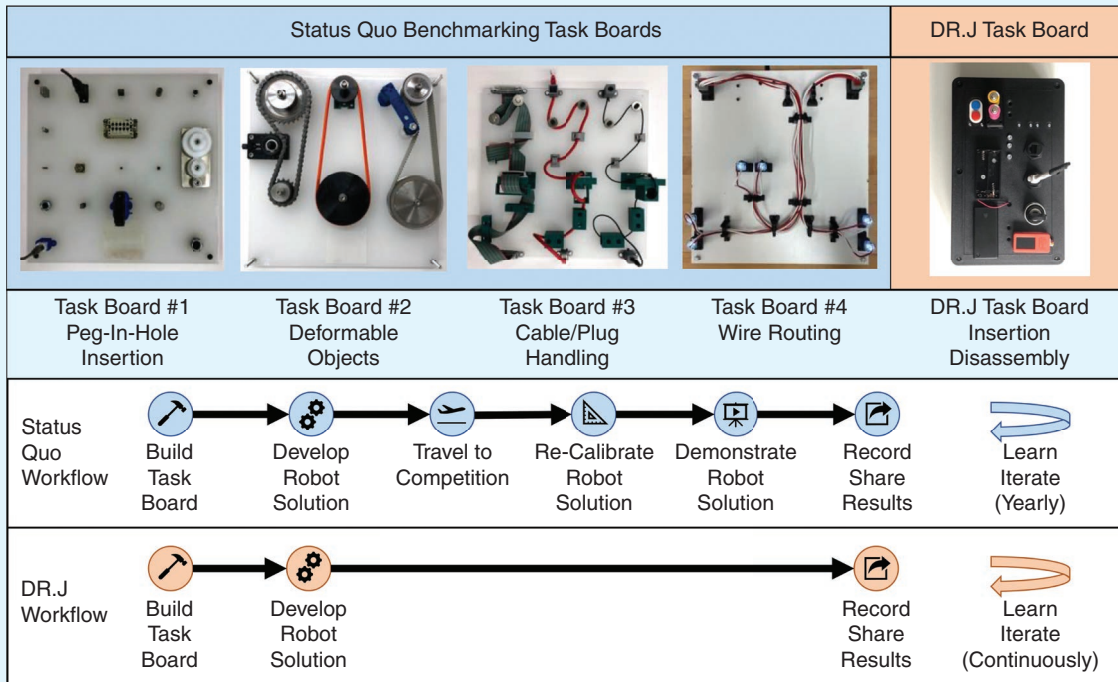


FIGURE 2. DR.J task boards are equipped with an Internet-connected microcontroller to record manipulation experiments with sensors directly on the board and then immediately publish the results online. Recorded scores can be viewed and compared continuously on a WD, making it possible to compete and share results outside of organized conference events without the burden of transporting sensitive equipment to a central location.

- to be portable, i.e., easily mailed directly to participants (“about the size of a laptop and weigh under 1 kg”)
- to be assembled from low-cost parts, with a total value under US\$300
- to be robust to repeated use by robots
- to be reset from the goal state to the start state in under 30 s
- to be populated with a set of progressively complex manipulation tasks.

This design enables teams from different locations to participate in decentralized competitions and opens opportunities for asynchronous demonstrations with comparable performance data and common hardware. To promote transparency and collaboration among DR.J users, a program on the Internet-connected task board regularly sends telemetry data and publishes execution times within seconds of completing trial attempts to a public WD.

INTERNET-CONNECTED TASK BOARD

The task board equipped with an Internet-connected microcontroller

serves as the evaluation network device component of DR.J. The microcontroller, an ESP32-pico-d4 provided by M5Stack, monitors the performance circuits of objects mounted to the task board and provides a timing clock to

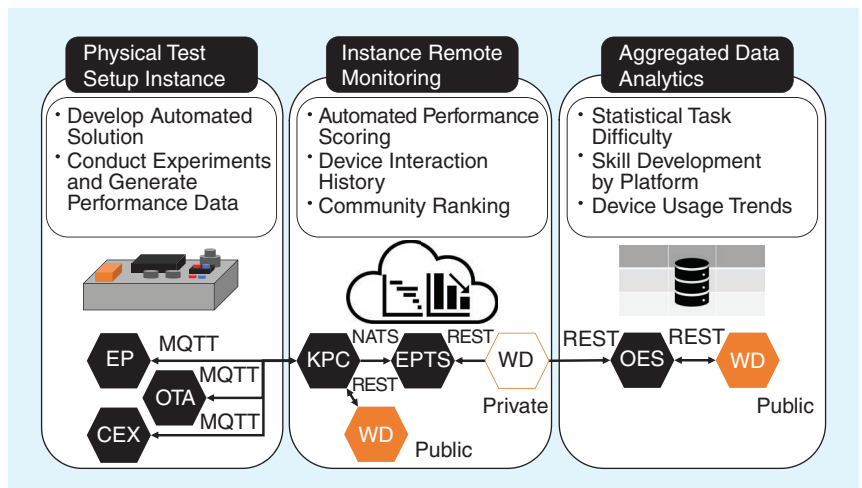


FIGURE 3. The task board devices are IoT endpoints (EPs), which emit telemetry information about experiments that can be viewed remotely as an endpoint time series (EPTS). The data from individual EPs is rendered as a public WD over the Internet. Multiple EPs are managed from a private WD that can send remote commands (CEX) and over-the-air (OTA) updates using the Kaa communication protocol (KPC). Historical data from each EP are aggregated into a cloud database, where data analytics tools, such as open distro elastic search (OES), can be used to get insights in device usage on another WD. REST: Representational State Transfer.

measure task execution times. For our competition, we selected six task objects triggering electrical circuits upon task completion. Specifically, we used a combination of four buttons, a key switch, an Ethernet plug and ports, and a $2 \times$ AA battery case mounted on the task board top plate. Figure 4 shows an exploded view of the task board components. All parts needed to complete the trial protocol are included with and reside on the electronic task board. The DR.J controller monitors the starting state of all parts using the performance circuits and prevents users from beginning a timed trial until this requirement is met. In this way, consistent start conditions can be maintained across subsequent trial attempts. Table 2 shows the specifications for the task board stopwatch resolution, telemetry rate, and peripherals.

An acrylonitrile butadiene styrene (ABS) project box provides the structural base and housing for the electronics. We manufactured the task boards for our competition ourselves to provide participants with consistent test objects. A computer numerical control mill was used to cut mounting features for task objects with ± 0.5 -mm repeatability. The electronics are assembled using off-the-shelf (OTS), plug-and-play components with an extendable Inter-Integrated Circuit (I2C) hub device for ease of assembly and reconfiguration. These design choices enable quick iteration of new task board designs as

well as provide the possibility to support future expansions with additional sensors.

With the selected task objects, we aim to assess competitors' ability to localize objects, perform precise insertions, and disassemble multipart components with a single arm manipulator robot using a single two-finger pinch gripper and camera sensor in under 10 min. Prior to launching our competition, we verified this with a Franka Emika Panda robot arm and arm-mounted Intel Realsense D435 camera internally.

The task board assembly was designed with Onshape, a browser-based CAD program, to be easily viewed, replicated, and extended by anyone with a web browser. The microcontroller was programmed in C using Microsoft Visual Studio Code. Both the mechanical design files and microcontroller source code are available on the GitHub repository shared in the "Conclusion" section.

TRIAL RECORDS, PROTOCOLS, AND PERFORMANCE EVALUATION

A trial record is the information related to a specific trial attempt of a registered user with a DR.J task board. It is created each time a user presses the "start trial" button. Minimally, each trial record contains a timestamp, a unique endpoint ID, a trial protocol ID, and an array of individual task execution times. A trial protocol specifies operational requirements, including individual task descriptions, trial execution rules, start and goal states for the task objects, relevant performance evaluation metrics, and how they are calculated. Competition organizers define trial protocols to accommodate their unique requirements. Custom programs deployed on the microcontroller ensure users follow the prescribed trial protocol, for example, having a specific starting configuration or requiring subtasks (STs) to be completed in a specific order.

In the Robothon Grand Challenge competition, the trial protocol required users to place the task board in a new random location on a flat surface fixed with Velcro strips before each attempt, and teams could complete tasks in any order. To have a coordinated start across the distributed teams, we first mailed teams their task boards and then shared the trial protocol afterward. The primary evaluation metric of DR.J for task performance is execution time. It is measured as an offset from the moment the user presses the trial start button to when an ST is completed, as detected by the performance circuits. The user may only begin a trial attempt if all task objects are detected in their starting states as specified by the trial protocol and enforced by the program on the microcontroller. Figure 5 shows example completions of the competition tasks by a human actor and various robot platforms.

For our competition, teams self-reported the details of their robot platform during the competition application phase. During the competition, teams adapted their robot platform and tested new algorithms with the task board. At the end of a 30-day development phase, teams submitted a documentation package along with a video recording of

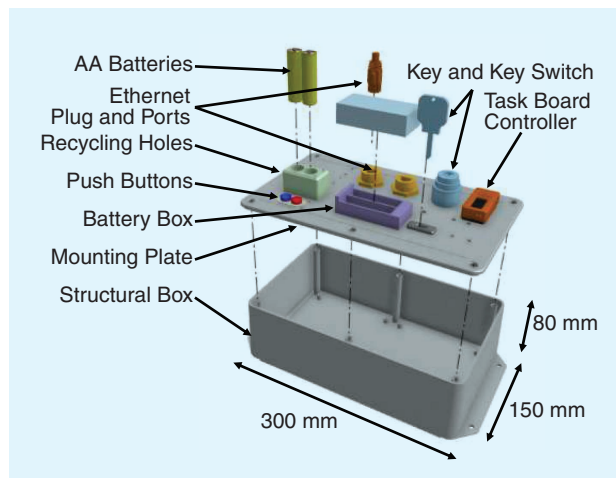


FIGURE 4. An exploded view of the prototype task board assembly. The top plate is interchangeable to support new task objects added by the community in future iterations. The base houses the electronics and provides a structure to mount the device on a flat surface.

TABLE 2. The prototype DR.J task board specifications.

PARAMETER	SPECIFICATION
Telemetry publishing rate	0.2 Hz (idle)/20 Hz (active trial)
Stopwatch resolution	1 ms
Electrical inputs for objects	10 digital/5 analog
User inputs for controller	3 push buttons
Controller interfaces for user	Organic LED screen/USB/Wi-Fi

their work. Teams with valid solutions were invited to present their automated solution to the expert jury live over video conference. Teams were awarded points, pass/fail, for completing each ST of the trial protocol and were ultimately distinguished by their overall trial completion time. A competition-specific program on the microcontroller monitored and displayed the task board state on a screen, guided the user through trial attempts, and reported trial results and device telemetry to a cloud database. Individual ST completion times, overall trial completion time, and six-axis accelerometer readings were reported in the device telemetry. A score $S_{P,A}$ for a team with platform P , running algorithm A , can be expressed by the following equation:

$$S_{P,A} = \sum_{i=1}^N t_{ST_i} \quad (1)$$

where t_{ST_i} is the execution time of the i th ST in a trial protocol of N STs as measured by the electronic task board in seconds.

AUTOMATIC REPORTING OF RESULTS TO THE WD

Each DR.J task board is registered with a unique endpoint token to distinguish its data on the WD. When a user powers on the task board, it automatically connects to a central web server over Wi-Fi and begins sending telemetry information over Message Queuing Telemetry Transport (MQTT), an OASIS standard messaging protocol for IoT devices (<https://mqtt.org/>). Once connected, the statuses of the task board and its performance circuits are regularly updated on the public WD and saved with a service provided by the IoT platform Kaa. The KaaIoT framework is available at <https://kaaproject.github.io/kaa/docs/v0.10.0/>. To get started, users only need to specify their preferred Wi-Fi network and credentials on a configuration screen once, and then it is saved on the device for future connections. The WD lists all registered task boards and provides a clickable view to visualize data by the team as a time series chart that can be downloaded as a .csv file for further analysis. Data from the competition can be viewed on the WD at the following URL: <https://bit.ly/robothonTaskBoardWebDashboard>. A ranking report can be generated using Kaa's user tools with Kibana for a desired time window.

In our competitions, DR.J awarded trial points to teams for successfully completing each ST according to a trial protocol scorecard. The number of points awarded for each ST corresponded with its difficulty as estimated by the task board designer. This assignment of points to each ST based on its complexity proved to be challenging, as discussed in the next section.

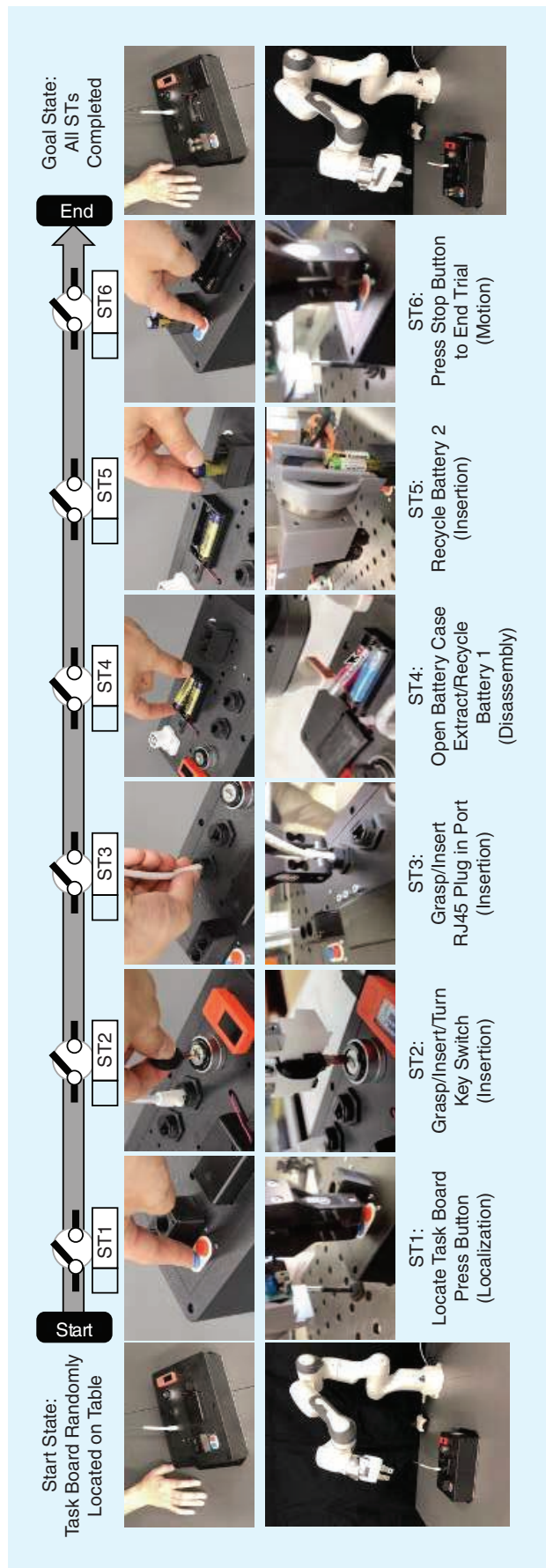


FIGURE 5. In the Robothon Grand Challenge, the task board trial protocol contained six STs ($N = 6$) as illustrated from the execution sequence at the top. The top row illustrates completion by a human using their dominant hand, and the bottom row shows completion by the different robot teams. DR.J remotely recorded the individual ST execution times for more than 400 trial attempts by 29 teams working independently over a 30-day development period in competition years 2021 and 2022.

ESTIMATING TASK COMPLEXITY

A challenge to designing progressively difficult tasks on the task board is finding agreement on their perceived complexity. Task difficulty and perceived complexity are based on the actor’s (human’s or robot’s) experience and capabilities in performing the work. To demonstrate the variance in perceived task complexity, we asked 10 individuals to complete the task board by hand and then surveyed them to describe their perceived task complexity in two ways: 1) to rate the task difficulty on a scale from one to 10 and 2) to assign values for each of the four task complexity dimensions as defined by [18] and stated as follows.

Respondents gave ratings for two execution scenarios: one for a human actor and one for a given robot actor. We chose to incorporate the task complexity dimensions from [18] to use a systematic and coherent method for determining task complexity. However, we found that respondents did not always agree in assigning values to the four dimensions: 1) the number of required actions, 2) the number of required information cues, 3) the number of dependencies with intermediate tasks, and 4) the number of dynamics to consider within the task. We simplified the analysis by asking respondents to assign a discrete value (from zero to 10) for each dimension and then took the sum to arrive at a total task complexity. We then used the totals to rank the STs by their calculated perceived difficulty. We found that respondents disagreed within individual task complexity dimensions, which confirms the ambiguity of the task’s complexity. However, ultimately, their summed total complexity values agreed with the simple rating method in terms of relative task difficulty. The mean averages and standard deviations of the survey results are shown in Table 3 alongside the task board designer’s ratings, indicated by x_d for reference.

The trial protocol, depicted in Figure 6, was designed to include popular manipulation tasks in the literature [16], [19], [20]. The ST designed to be the least complex, “ST1: Press Button,” requires the actor to locate the button and press it

down. The ST designed to be the most complex, “ST4: Recycle Battery,” requires the actor to locate the battery box, slide open the lid, extract a spring-loaded battery, pick it up, orient it, and place it into a recycling receptacle over another button and press down. Ordering the STs by the survey difficulty ratings, respondents agreed that the button pressing tasks (ST1 and ST6) were the least difficult, while opening the battery case and recycling a battery (ST4) was the most difficult. From the literature, the term *task complexity* appears in several different contexts, making it difficult to consistently apply a method. To remove the dependence on individual interpretation, we propose using the timing data provided by DR.J. as a data-driven method to determine task complexity. The more time required to complete a task in aggregate, the more complex it is for a given actor class. We refer to the best score of any robot platform and any algorithm as $S_{anyRobot/anyAlgorithm}$.

By using DR.J to collect performance data through our competitions, we can begin to build a historical database of ST completion times across the network of electronic task boards for different combinations of platforms and algorithms. As users continue to use the DR.J task boards, we will have more comprehensive data on robot performance to better understand the evolution of robot manipulation skill development. Performance improvements for identical platforms can be attributed to improved algorithms up to an assumed theoretical limit based on its physics. Following the estimated task complexity values from Table 3, we expect teams to spend more effort and time to complete tasks of higher complexity.

As the design of robot platforms evolves quickly and often emulates humans, we normalize the robot platform performance from our competitions against an expert human performance captured by the task board. Using the timing data collected with DR.J, we can define a new term, *RTC*, to compare performances between two platform-task system pairs. RTC, denoted by $\theta_{A/A_{ref}}$ of two different platforms solving the

TABLE 3. Survey results for the estimated difficulty and task complexity for each ST in the Robothon trial protocol ($N = 10$). *

TRIAL PROTOCOL STs	ST1			ST2			ST3			ST4			ST5			ST6		
	x_d	\bar{x}	σ	x_d	\bar{x}	σ	x_d	\bar{x}	σ	x_d	\bar{x}	σ	x_d	\bar{x}	σ	x_d	\bar{x}	σ
Difficulty rating (1 =low to 10 = high)																		
Human actor	1	1.2	0.4	3	4.2	2.3	2	2.9	1.4	4	4.9	2.3	3	4.3	1.9	1	1.6	1.6
Robot actor	3	1.8	0.6	7	6.6	2.1	5	5.1	2.2	8	7.8	1.4	6	6.7	1.9	1	1.8	0.6
Task complexity dimensions from [18]																		
Number of required actions (D1)	2	1.4	0.7	4	4.1	1.4	2	3	1.2	6	5.6	2	4	4.1	1.7	1	1.1	0.3
Number of information cues (D2)	1	1.5	0.5	5	3.1	1.9	2	2.6	1.2	4	4.4	2.3	2	3.1	2.1	2	1.5	0.5
Number of dependencies (D3)	1	0.8	1.4	2	3.7	1.8	1	2.6	1.8	3	4.8	2.4	1	3.8	2.5	0	0.8	1.4
Number of dynamics (D4)	0	0.6	0.7	0	1.6	2.1	0	1.7	1.6	0	2.7	2.3	0	2.6	2.3	0	0.5	0.7
Estimated task complexity (ΣD)	4	4.3	2.5	11	12.5	5.8	5	9.9	4.5	13	17.5	6.8	7	13.6	7.7	3	3.9	2.2

D: dimension.

The BOLD values indicate the subtask (ST) perceived to be the most difficult to complete.

same trial protocol, can be expressed as a percentage with the following notation:

$$\theta_{A/A_{ref}} = \frac{S_{P_{ref},A_{ref}}}{S_{P,A}} * 100 \quad (2)$$

where $S_{P,A}$ refers to the best score of the platform P and algorithm A , and $S_{P_{ref},A_{ref}}$ refers to the best score of the reference platform and algorithm.

RTC provides a numerical method and an alternative to user survey ratings to describe task complexity with respect to a reference platform that is independent of individual task interpretation and opinion.

DR.J RESULTS IN DECENTRALIZED COMPETITION

The Internet-connected task board and WD platform were validated in an international competition, the Robothon Grand Challenge, in 2021 and 2022. Each year, selected teams were mailed an electronic task board and given the grading rubric and trial protocol during a kickoff meeting over a conference call for a coordinated start. From then on, each team had 30 days to develop an automated solution to the trial protocol using the task board with their own robot platform. At the end of the development period, each team sent a documentation package comprising three components: 1) a list of hardware and software used, along with quick start instructions to run their demo as a repository; 2) a 5-min video recording of their team presenting their strategy and best solution; and 3) an uncut video recording of their robot solving the task board five times in a row. Teams with a valid solution were invited to present their solution live to our expert jury over a video conference call.

In 2021, 20 team applications were received; nine were selected, and each was mailed a task board to participate. Only four of the nine successfully finished all tasks with their robot platform. In 2022, 27 team applications were received; 20 were selected, and each was mailed a task board to participate. Six of the 20 successfully finished all tasks with their robot platform. In total, 10 of the 29 teams successfully completed the trial protocol. Their platforms are shown in Figure 7. Their strategies and robot platform components are listed in Table 4. The winning teams for each year are denoted with an asterisk. The winners were determined with the timing data from DR.J during the supervised demonstrations with the expert jury since only then could they certify the solution and that no one circumvented the rules through the use of fixtures or teleoperation. The distribution of task execution times from the finishing teams overlaid with the estimated task complexities from the survey responses

is shown in Figure 8. The required task execution time from the competition follows the estimated task complexity curve for all STs except for ST1.

COMPARING PERFORMANCE ACROSS ROBOT PLATFORMS

Across the range of robot platforms, patterns emerged in how teams solved each ST. Using each team’s competition application, final documentation submission, and DR.J, we compared robot platform performance data across the several automation components used during the competition. We grouped teams’ strategies by ST and listed each team’s OTS components (Table 4) to show the most popular approaches and equipment used. We noticed that, while some teams used similar or exactly the same hardware, they achieved different results based on their strategy and implementation. With DR.J and the historical timing data collected, we can show how each robot platform performs relative to the others.

Looking across all robot platforms to see which STs required the most time, we can rank tasks in order of their RTC with respect to a reference platform. Examining the most time-consuming tasks and strategies can prioritize topic areas for robotic skill improvement with the greatest impact on performance. For example, ST1 took the majority of the time for all teams in both years despite being estimated to be one of the least complex tasks. As a result, more attention should be given to developing robot techniques for identifying and locating objects in space. Interestingly, robot teams spent more time solving ST1 than ST4, which was estimated to be the most complex task. This mismatch highlights the bias of humans in predicting task complexity for robot platforms.

We analyzed the telemetry data collected from each team’s task board to understand their behavior during the 30-day development period, summarized in Table 5. The number of days to a team’s first task board connection can indicate the

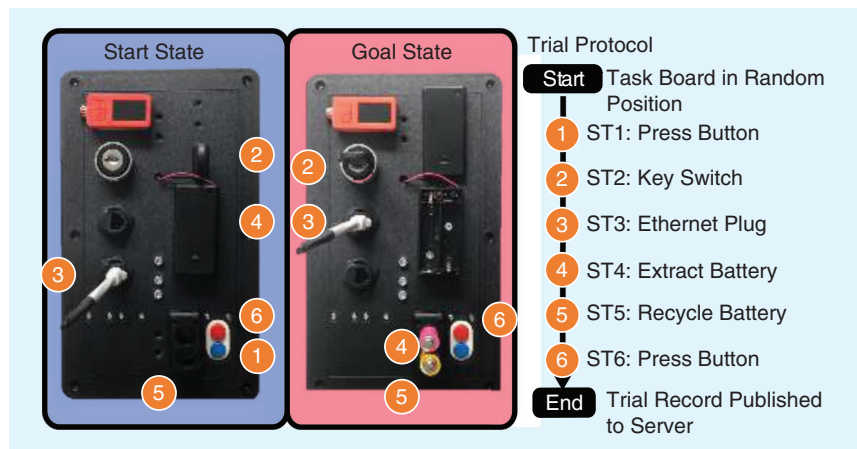


FIGURE 6. The electronic task board with the initial state and end state side by side with task classifications. The DR.J microcontroller verifies that task objects are in their starting state through the performance circuits prior to allowing the user to start a trial to promote consistency between tests. In the Robothon Grand Challenge, teams were allowed to complete the STs in any order.

team's eagerness to join the network. The number of days to a team's first successful trial completion can indicate the team's competency in applying its robot platform to solve the trial protocol. The number of successful runs can approximate the team's commitment to finding an optimal solution. The mean averages, standard deviations, and best trial times are shown to give insight into the range of performances by each team.

DEMONSTRATING PERFORMANCE IMPROVEMENTS FOR TEAMS WITH SIMILAR ROBOT PLATFORMS

While teams used similar hardware in the competition, they reported a diverse range in performance. As previously stated, we assume a theoretical performance limit exists for each robot platform based on its mechanical properties. Teams approach this limit differently based on their experience levels and implementations. With DR.J and the historical timing data collected, we can follow the performance evolution of a particular team or robot platform. Classes of robot platforms can be derived by looking at the intersection of similar OTS components used. From Table 4, we can see that teams B, F, and H have nearly identical hardware platforms, with a Universal Robot UR5 manipulator and Robotiq Hand-E Gripper with an on-arm Intel Realsense d435i camera. While none of these teams won a competition, they reported similar best times, as seen in Table 5. This grouping could indicate a performance limit of the hardware used by these teams. Team D and team I had similar hardware; in fact, they were from the same university, and a new algorithm resulted in a large performance improvement—a reduction of 327 s—between the competition years 2021 and 2022. In practice, implementations will vary among robot teams, and more data need to be

collected to make claims about the limits of individual pieces of automation hardware.

COMPARING PERFORMANCE BETWEEN HUMANS AND ROBOTS

After the competitions, a small set of individuals ($N = 10$) was asked to complete the trial protocol with the task board by hand a minimum of 10 times. No improvement was observed beyond 10 attempts. This group's average trial completion time was 18.9 s, with a standard deviation of 9.5 s and a best time of 8.1 s. Figure 9 shows the summary of the performance of the best robot team for each year against the best human performance. Using the individual ST execution times from DR.J, we can analyze their contributions to an actor's overall trial execution time.

In Table 6, we calculate RTC for each ST to show the performance gap between the most interesting robot platforms and algorithms, the winning robot teams, and a human. Furthermore, we can reference the best score of any robot and any algorithm on record as the known historical limit for each ST. We found that the best robot performance is only 16% as fast as the best human performance across the entire set of tasks. Taking a closer look at individual STs, the best team in 2021 was $9\times$ faster (850%) than the best human in the battery recycling task (ST5). To achieve this, the robot team used an electromagnet and fixture attached to the end effector to pick and place the two batteries at once, while the human only picked and placed one battery at a time. This clever technique was a great example of a purpose-built solution excelling in a single task. The fastest robot team in 2022 achieved a better overall score with a more well-rounded solution to all of the STs with

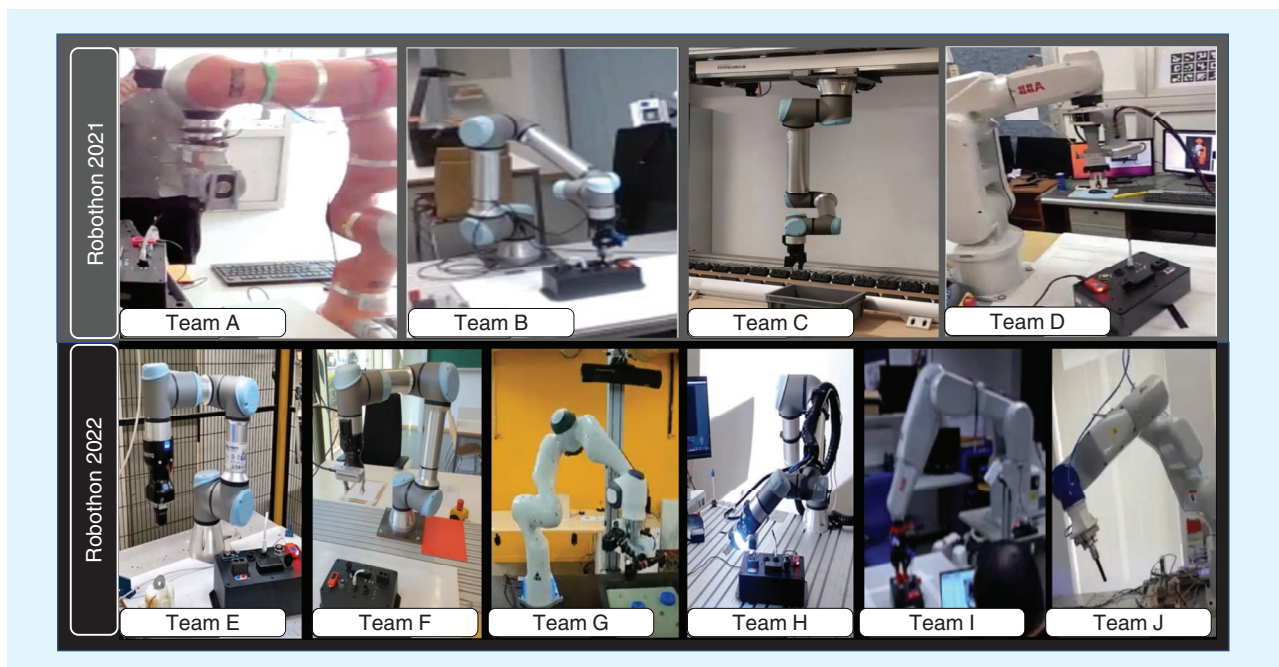


FIGURE 7. Screen captures of the robot teams that fully completed the trial protocol from the set of the different robot platforms evaluated by DR.J during the Robothon Grand Challenge competition. The first row shows finalists from 2021. The second row shows finalists from 2022.

motivate developers to focus on the most challenging aspects to reach human-level performance, like object localization techniques, between competition years. Furthermore, the combination of the DR.J timing data, task difficulty survey data, and documented robot capabilities can inform future competition task selection.

While our approach has its advantages, it also has limitations. DR.J does not prevent users from circumventing the rules or replacing the expert jury, but, rather, it provides a convenient tool for collecting and sharing experimental performance data across multiple locations. Nearly all respondents agreed an expert jury is still necessary to feel fairly judged in a competition setting. We acknowledge that testing conditions in remote locations cannot be fully controlled, and the integrity of the data collected by the Internet-connected task board needs to be considered such that bad actors could transmit data that do not comply with the trial protocol. Furthermore, task elements and task definitions are limited to those that can

physically fit on the task board and whose completion close an electrical circuit. Also, a tradeoff must be made to determine the telemetry publishing rate. Faster data rates provide more data but also consume more computation and network bandwidth as well as consume more device battery power. The current task board design is equipped with a small battery lasting approximately 20 min when broadcasting over Wi-Fi. Therefore, we implemented a 10-min timeout for trial attempts

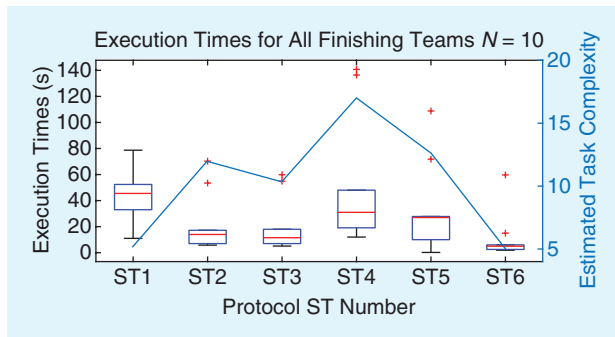


FIGURE 8. A box plot of robot team execution time by ST overlaid with the estimated task complexity from survey results in the Robothon Grand Challenge competition. The average time required for robot teams to complete each ST followed the estimated task complexity except for ST1. We expected more complex tasks to require more time to complete. However, humans can underestimate the complexity of intuitive tasks, like object localization, for robot platforms.

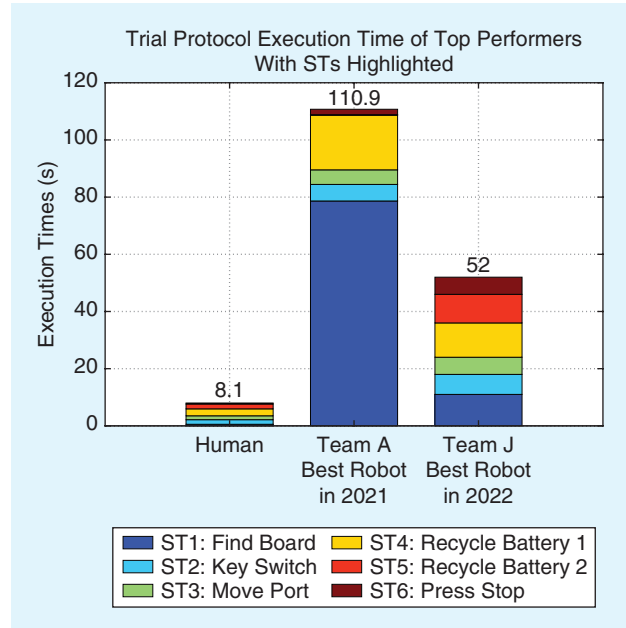


FIGURE 9. A stacked bar plot of timing data collected with DR.J showing the performance gap between the best human performance and the best robot performance captured during the Robothon competition. Individual ST completion times highlight areas for improvement with the greatest impact on total trial execution time. This is most apparent in the robot team improvement in the object localization of ST1 between competition years 2021 and 2022.

TABLE 5. Performance development evaluation metrics.

ACTOR	TIME TO FIRST TASK BOARD CONNECTION (DAYS)	TIME TO FIRST SOLUTION (DAYS)	NUMBER OF SUCCESSFUL RUNS	MEAN TRIAL TIME \bar{s} (S)	TRIAL TIME STANDARD DEVIATION σ (S)	BEST TRIAL TIME s (S)
Team A*	13	13	62	184.6	243.3	110.9
Team B	30	30	59	201.1	30.5	178
Team C	21	29	9	410.2	152	338
Team D	22	27	42	502.1	183.8	437
Team E	28	29	2	223.8	181.9	117
Team F	0	27	58	125.4	270.9	105
Team G	12	27	7	552.2	310.3	127
Team H	15	28	56	163.1	83.3	107
Team I	28	27	39	269	330.9	110.4
Team J*	12	27	35	169.5	180	52
Humans (N = 10)	1	1	120	18.9	9.5	8.1

and recommended keeping the task board plugged in to avoid interruptions in the telemetry report.

Despite the aforementioned limitations, we believe DR.J and the electronic task board can be valuable tools for the robotics community to benchmark, aggregate, and compare decentralized manipulation performances.

FUTURE WORK

Feedback from both the expert jury and the participants in our competitions has been positive, and we want to continue to develop the DR.J platform. We plan to add more sophisticated sensors to increase the diversity of interactions with the task board. Analog sensors, such as potentiometers to measure the linear and angular positions of parts or strain gauges to measure forces, can extend DR.J to go beyond binary pass/fail assessments. New metrics can be introduced to grade not only execution time but also the quality of the actor's work. In addition to the WD, we want to add a task board–robot communication interface so that users can directly access the feedback from the performance circuits to confirm their execution results, which would be helpful for machine learning techniques. We aim to make it easy for community members to design their own permutations of the trial protocol or add new elements to the task board to fit their use cases. To this end, the design files and software for the electronic task board are made publicly available with the associated project repository so that readers can replicate their own task boards, shared in the “Conclusion” section.

To increase adoption and support for performance benchmarking, we will continue organizing the Robothon Grand Challenge competition, which will coincide with the automata trade show. Furthermore, our proof-of-concept platform has been picked up by a funded European Union project called *euROBIN*, which will develop a pan-European network for robotics research and collaboration. (The euROBIN network aims to advance artificial intelligence tools, software, architectures, and hardware components in a reproducible approach: <https://www.eurobin-project.eu/>.) Within this

project, the electronic task board will be used to develop, share, and validate results. We plan to scale the production of the task boards to make competition versions available and provide support for new designs proposed by the community to increase the number of users on the platform.

DR.J AS A TOOL FOR RESEARCH COLLABORATION

Collaborators working on the same or similar challenges can use the DR.J platform to share performance data results with their own task board design across sites. As experimental data are collected and scored automatically, Internet-connected task boards present a new paradigm in benchmarking real-world robot performance. Historical telemetry data can be downloaded and analyzed by anyone directly from the WD for further analysis of how roboticists on the DR.J network are progressing in developing new manipulation skills based on a single data source. Robot solutions from competitions can be referenced for inspiration for new robot platform designs and used as a starting point for teaching robot manipulation techniques in workshops and courses.

CONCLUSION

In this work, we presented the digital robotic judge known as DR.J, a platform for recording and automatically reporting benchmarking manipulation performances remotely with electronic task boards. We introduced RTC for describing the platform-specific difficulty of tasks free of individual bias using only the historical task execution data we collected with DR.J. We validated the proof-of-concept data collection features of DR.J through two instances of an international robot manipulation competition, the Robothon Grand Challenge, in 2021 and 2022 and shared data collected from 10 robot teams around the world.

Our platform provides a streamlined workflow for benchmarking manipulation performances without transporting robot equipment to in-person competition events. The Internet-connected electronic task board is a tool for measuring and comparing manipulation performances across robot platforms and can quantitatively show the performance gap between humans and robots. Through our own decentralized competition, we found that robot platforms can greatly exceed human performance for specialized tasks, like battery handling, but still lack the general dexterity to perform a range of manipulation skills as well as a human. We hope automated reporting with IoT devices will promote more transparency and frequent collaboration across sites in periods between and during conference events. We believe DR.J can be the basis for the robotics community to build a network of physical task board devices on a modular and scalable platform that encourages researchers to generate reproducible benchmarking results, share data, and benefit from one another.

Details about the Robothon Grand Challenge competition are available at the website <https://www.robothon-grand-challenge.com>. A video describing the DR.J platform as

TABLE 6. Score in seconds and RTC for the trial protocol between the best robots and best human performance.

ACTOR SCORE AND RTC	ST1	ST2	ST3	ST4	ST5	ST6	TRIAL RUN
$S_{\text{HumanBody,Human}^*}$	0.6	1.6	1.4	2.4	1.7	0.4	8.1
$S_{\text{Kuka-iiwa,Team A}}$	78.8	5.8	5.1	19.1	0.2	1.9	110.9
$S_{\text{Epson-VT6,Team J}}$	11	7	6	12	10	6	52
$S_{\text{AnyRobot,AnyAlgorithm}}$	11	5.8	5.1	12	0.2	1.9	52
$\theta_{\text{Team A/Human}^*}$	1%	28%	27%	13%	850%	21%	7%
$\theta_{\text{Team J/Human}^*}$	5%	23%	23%	20%	17%	7%	16%
$\theta_{\text{AnyRobot/Human}^*}$	5%	28%	27%	20%	850%	21%	16%

* Indicates the best in class performer. Values in bold are the best reported scores for the subtask.

presented in this article can be viewed at <https://www.tiny.cc/DRJPaperVideo>. Design files, software, and construction details for the replication of the presented task board can be found on the project repository hosted on GitHub: <https://www.github.com/peterso/robotlearningblock>.

ACKNOWLEDGMENT

We thank the Robothon Grand Challenge competition teams, sponsors, and expert jury for their hard work, trust, and support. This work was performed in collaboration with industry partners Messe München and Microsoft. We gratefully acknowledge the funding of the Lighthouse Initiative KI.FABRIK Bayern by StMWi Bayern (KI.FABRIK Bayern Phase 1: Aufbau Infrastruktur und KI.Fabrik Bayern Forschungs- und Entwicklungsprojekt, Grant DIK0249). This work was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy—EXC 2050/1—Project ID 390696704—Cluster of Excellence “Centre for Tactile Internet with Human-in-the-Loop” of Technische Universität Dresden.

This work was supported by the European project euROBIN under Grant Agreement 101070596.

Part of the research presented in this work was conducted when Fares Abu-Dakka was at the Munich Institute of Robotics and Machine Intelligence, Technical University of Munich.

Peter So is the corresponding author for this article.

DISCLAIMER

Sami Haddadin has a potential conflict of interest as a shareholder of Franka Emika GmbH.

AUTHORS

Peter So, Munich Institute of Robotics and Machine Intelligence, Technical University of Munich, 80797 Munich, Germany. E-mail: peter.so@tum.de.

Andriy Sarabakha, School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore. E-mail: andriy001@e.ntu.edu.sg.

Fan Wu, Munich Institute of Robotics and Machine Intelligence, Technical University of Munich, 80797 Munich, Germany. E-mail: frankie.wufan@gmail.com.

Utku Culha, Munich Institute of Robotics and Machine Intelligence, Technical University of Munich, 80797 Munich, Germany. E-mail: utku.culha@tum.de.

Fares J. Abu-Dakka, Electronic and Informatics Department, Faculty of Engineering, Mondragon Unibertsitatea, 20500 Arrasate, Spain. E-mail: fabudakka@mondragon.edu.

Sami Haddadin, Munich Institute of Robotics and Machine Intelligence, Technical University of Munich, 80797 Munich, Germany. E-mail: haddadin@tum.de.

REFERENCES

- [1] F. Amigoni et al., “Competitions for benchmarking: Task and functionality scoring complete performance assessment,” *IEEE Robot. Autom. Mag.*, vol. 22, no. 3, pp. 53–61, Sep. 2015, doi: 10.1109/MRA.2015.2448871.
- [2] Y. Sun, J. Falco, M. A. Roa, and B. Calli, “Research challenges and progress in robotic grasping and manipulation competitions,” *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 874–881, Apr. 2022, doi: 10.1109/LRA.2021.3129134.
- [3] E. Krotkov et al., “The DARPA robotics challenge finals: Results and perspectives,” *J. Field Robot.*, vol. 34, no. 2, pp. 229–240, Mar. 2017, doi: 10.1002/rob.21683.
- [4] N. Correll et al., “Analysis and observations from the first amazon picking challenge,” *IEEE Trans. Automat. Sci. Eng.*, vol. 15, no. 1, pp. 172–188, Jan. 2018, doi: 10.1109/TASE.2016.2600527.
- [5] A. Saffiotti et al., *RoCKIn – Benchmarking Through Robot Competitions*. London, U.K.: InTechOpen, 2017.
- [6] H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda, and E. Osawa, “RoboCup: The robot world cup initiative,” in *Proc. 1st Int. Conf. Auton. Agents (AGENTS)*, New York, NY, USA: ACM, 1997, pp. 340–347, doi: 10.1145/267658.267738.
- [7] B. Calli, A. Dollar, M. A. Roa, S. Srinivasa, and Y. Sun, “Guest editorial: Introduction to the special issue on benchmarking protocols for robotic manipulation,” *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 8678–8680, Oct. 2021, doi: 10.1109/LRA.2021.3114936.
- [8] D. J. Clark and T. Nilssen, “Creating balance in dynamic competitions,” *Int. J. Ind. Org.*, vol. 69, Mar. 2020, Art. no. 102578, doi: 10.1016/j.ijindorg.2019.102578.
- [9] S. Carpin, M. Lewis, J. Wang, S. Balakirsky, and C. Scrapper, “USARSim: A robot simulator for research and education,” in *Proc. IEEE Int. Conf. Robot. Autom.*, 2007, pp. 1400–1405, doi: 10.1109/ROBOT.2007.363180.
- [10] Y. Zhu, J. Wong, A. Mandelkar, and R. Martín-Martín, “robosuite: A modular simulation framework and benchmark for robot learning,” 2020, [arXiv:2009.12293](https://arxiv.org/abs/2009.12293).
- [11] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, “RLBench: The robot learning benchmark and learning environment,” 2019, [arXiv:1909.12271](https://arxiv.org/abs/1909.12271).
- [12] O. Michel, “Online simulation robot benchmark homepage,” robotbenchmark. [Online]. Available: <https://webots.cloud/competition>
- [13] W. Wiedmeyer, M. Mende, D. Hartmann, R. Bischoff, C. Ledermann, and T. Kroger, “Robotics education and research at scale: A remotely accessible robotics development platform,” in *Proc. Int. Conf. Robot. Autom. (ICRA)*, 2019, pp. 3679–3685, doi: 10.1109/ICRA.2019.8793976.
- [14] Z. Liu et al., “OCRTOC: A cloud-based competition and benchmark for robotic grasping and manipulation,” *IEEE Robot. Autom. Lett.*, vol. 7, no. 1, pp. 486–493, Jan. 2022, doi: 10.1109/LRA.2021.3129136.
- [15] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, “Benchmarking in manipulation research: Using the Yale-CMU-Berkeley object and model set,” *IEEE Robot. Autom. Mag.*, vol. 22, no. 3, pp. 36–52, Sep. 2015, doi: 10.1109/MRA.2015.2448951.
- [16] K. Kimble et al., “Benchmarking protocols for evaluating small parts robotic assembly systems,” *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 883–889, Apr. 2020, doi: 10.1109/LRA.2020.2965869.
- [17] E. Strickland, “Cyborgs go for gold,” *IEEE Spectr.*, vol. 53, no. 1, pp. 30–33, Jan. 2016, doi: 10.1109/MSPEC.2016.7367459.
- [18] R. Wood, “Task complexity: Definition of the construct,” *Organizational Behav. Human Decis. Processes*, vol. 37, no. 1, pp. 60–82, Feb. 1986, doi: 10.1016/0749-5978(86)90044-0.
- [19] L. Johannsmeier, M. Gerchow, and S. Haddadin, “A framework for robot manipulation: Skill formalism, meta learning and adaptive control,” in *Proc. Int. Conf. Robot. Autom. (ICRA)*, 2019, pp. 5844–5850, doi: 10.1109/ICRA.2019.8793542.
- [20] R. J. Kirschner et al., “Towards a reference framework for tactile robot performance and safety benchmarking,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2021, pp. 4290–4297, doi: 10.1109/IROS51168.2021.9636329.

