SCIENCE AND TECHNOLOGY ORGANIZATION
CENTRE FOR MARITIME RESEARCH AND EXPERIMENTATION

Reprint Series **CMRE-PR-2019-021**

Scoring robotic competitions: balancing judging promptness and meaningful performance evaluation

Fausto Ferreira, Gabriele Ferri, Yvan Petillot, Xingkun Liu,
Marta Palau Franco, Matteo Matteucci,
Francisco Javier Pérez Grau, Alan Ft Winfield

May 2019

Originally presented at:

18th IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), April 25-27 2018, Torres Vedras, Portugal, pp. 179-185, doi: 10.1109/ICARSC.2018.8374180

About CMRE

The Centre for Maritime Research and Experimentation (CMRE) is a world-class NATO scientific research and experimentation facility located in La Spezia, Italy.

The CMRE was established by the North Atlantic Council on 1 July 2012 as part of the NATO Science & Technology Organization. The CMRE and its predecessors have served NATO for over 50 years as the SACLANT Anti-Submarine Warfare Centre, SACLANT Undersea Research Centre, NATO Undersea Research Centre (NURC) and now as part of the Science & Technology Organization.

CMRE conducts state-of-the-art scientific research and experimentation ranging from concept development to prototype demonstration in an operational environment and has produced leaders in ocean science, modelling and simulation, acoustics and other disciplines, as well as producing critical results and understanding that have been built into the operational concepts of NATO and the nations.

CMRE conducts hands-on scientific and engineering research for the direct benefit of its NATO Customers. It operates two research vessels that enable science and technology solutions to be explored and exploited at sea. The largest of these vessels, the NRV Alliance, is a global class vessel that is acoustically extremely quiet.

CMRE is a leading example of enabling nations to work more effectively and efficiently together by prioritizing national needs, focusing on research and technology challenges, both in and out of the maritime environment, through the collective Power of its world-class scientists, engineers, and specialized laboratories in collaboration with the many partners in and out of the scientific domain.

NOTE: The CMRE Reprint series reprints papers and articles published by CMRE authors in the open literature as an effort to widely disseminate CMRE products. Users are encouraged to cite the original article where possible.

Scoring robotic competitions: balancing judging promptness and meaningful performance evaluation

Fausto Ferreira[1], Gabriele Ferri[1], Yvan Petillot[2], Xingkun Liu[2], Marta Palau Franco[3], Matteo Matteucci[4], Francisco Javier Pérez Grau[5] and Alan Ft Winfield[3]

Abstract— **To have a significant and fair competition an adequate scoring system is necessary. Different scoring systems exist, each one directly related to the nature and goals of the competition, e.g. student/educational, focused on benchmarking, Grand Challenge style, etc. However, the design of such systems is not an easy task. It is mandatory to design approaches which enable the judges to score teams in a reasonable time after the end of their performance. Scoring systems cannot be therefore extremely complex. On the other hand, it is crucial to have a judging system for teams which provides a meaningful and fair performance evaluation of what the teams achieved in the field. In this paper, several approaches to scoring are presented and compared. Our focus is on search and rescue competitions. Each approach is critically analysed and its features comparatively discussed. This analysis is beneficial to provide an overview of scoring systems tailored to different kinds of competitions. The reported examples can be used as building blocks to improve existing scoring systems or to design new approaches.**

I. Introduction

In the last years, robotics competitions have been becoming more and more popular in the robotics community. This can be seen looking at the number of competitions currently going on including million dollars prize challenges and the widening of the scope of early competitions such as RoboCup[1]. More and more competitions are being created or extended. Competitions range from purely student-oriented (e.g., training and education focussed) to Grand Challenges, where the aim is to push the state of the art [1]. Nonetheless, there are many types of competitions that can have multiple goals such as state of the art advancement and standardisation or benchmarking. The usefulness of robotics competitions to form the new generations of engineers/scientists and/or to push the state of the art in particular fields is widely recognised [2], [3]. One of the most famous competition is RoboCup. It started as a soccer challenge and was then extended both in participant targets (RoboCup Junior) as well as scope (Rescue, @Home, Industrial). Another very successful string of competitions was sponsored by the Defense Advanced Research Projects Agency (DARPA). The DARPA Grand Challenge[2] started in 2004, with another edition in 2005 and in 2007 transformed itself in the DARPA Urban Challenge[3] with autonomous ground vehicles operating in a mock urban scenario. This series of competitions had a fundamental role in jump-starting research that is now applied to self-driving cars. The Multi Autonomous Ground-robotic International Challenge (MAGIC) 2010 was a prize competition for multi-vehicle robotic teams that can execute an intelligence, surveillance and reconnaissance mission in a dynamic urban environment [4]. From 2013 to 2015, DARPA ran the DARPA Robotics Challenge[4] dedicated to humanoids. We have been witnessing an extension and maturing of several competitions which evolved in multi-domain competitions making the events more complex from the scoring point of view.

For instance, in Europe, competitions that have been running for 10 years or more, like the European Land Robot Trial (ELROB) [5] or the Student Autonomous Underwater Vehicles Competition - Europe (SAUC-E) [3] both since 2006, have joined efforts and together with the organizers of the workshop on Research Development and Education on Unmanned Aerial Systems (RED-UAS), have proposed the first multi-domain robotics competition including the three domains: land, air and sea, the so-called euRathlon Grand Challenge 2015. The European Robotics League (ERL) Emergency Robots 2017 was the follow-up of the euRathlon Grand Challenge [2] and included again the three domains. Also in 2017, the Mohamed Bin Zayed International Robotics Challenge (MBZIRC)[5] had land and aerial vehicles collaborating in different challenges. The trend is currently set as in 2019, the second edition of MBZIRC will have again land and aerial unmanned vehicles and the ERL will have a major event again with the three domains.

Thus, the complexity of worldwide robotics competitions is increasing. Designing an adequate scoring system is not an easy task. It is mandatory to design approaches which enable the judges to score teams in a reasonable time after the end of their performance. At the same time, it is crucial to have

[1]NATO Science and Technology Centre for Maritime Research and Experimentation (CMRE), Viale San Bartolomeo 400, 19126 La Spezia, Italy Fausto.Ferreira,Gabriele.Ferri@cmre.nato.int

[2]Heriot-Watt University, School of EPS, Riccarton Campus, EH144AS, Edinburgh, UK Y.R.Petillot@hw.ac.uk

[3]Bristol Robotics Laboratory University of the West of England, Bristol, UK Marta.PalauFranco, Alan.Winfield@uwe.ac.uk

[4]Politecnico di Milano, DEIB, Via Ponzio 34/5, I-20133, Milan, Italy Matteo.Matteucci@polimi.it

[5]Advanced Center for Aerospace Technologies (CATEC), Parque Tecnológico y Aeronáutico de Andalucía, C/ Wilbur y Orville Wright 19 - 41309 La Rinconada, Sevilla, Spain fjperez@catec.aero

This work was partly supported by the FP7 Coordination and support action EURATHLON project, grant agreement No 601205, the EU Horizon 2020 RockEU2 project under grant agreement No 688441 and EU Horizon 2020 SciRoc project under grant agreement No 780086

[1]http://www.robocup.org

[2]http://archive.darpa.mil/grandchallenge04/

[3]http://archive.darpa.mil/grandchallenge/

[4]http://www.theroboticschallenge.org

[5]http://www.mbzirc.com

a judging system for teams which allows judges to elaborate a meaningful and fair performance evaluation of what the teams achieved in the field. Scoring systems requirements are dictated mainly by the purpose of the competition and what the organizers want to achieve (e.g. if training students, pushing the state of the art or to benchmark several teams on particular functionalities).

The introduction of multi-domain competitions makes the judging process even more complicated, especially when it comes to judge the cooperation between domains and the contribution of each robot to complete the challenge. If one adds the fact that in many of these cases the competitions take place in unstructured outdoor environments, performance evaluation under repeatable conditions becomes challenging.

In such cases, benchmarking is hard to achieve as well as keeping the competition fair since teams do not compete in the same environmental conditions. These are all practical issues that the scoring system designers and the judges face in real-time. Scoring must be clear and easy to understand and use as in large competitions the number of judges can be significant and they should all have the same interpretation of the rules. Judges have different backgrounds and opinions so an objective scoring system with no margins for doubts/interpretations helps to streamline the judging process. A simple system is also easy to understand by the teams and can lead to fewer discussions and dubious situations. At the same time, objectivity, while desirable, can also lead to unfair situations being too reductive.

For instance, if one tries to perform benchmarking based on the competition results, the scoring system might be reduced to simple yes or no questions that could increase judging objectivity. Nonetheless, some tasks might not be well judged by simple yes or no, e.g. mapping of a large area, possibly creating unfair situations that do not reflect the reality. Therefore, it is important to balance fairness and realistic results with objective and easy to use scoring systems. There is no universal solution and the kind of scoring system is as well directly related to each kind of competition (student based or not), the environment where it takes place (unstructured/structured) and the goals of the competition (education, Grand Challenge style, benchmarking, etc).

In this paper, several search and rescue competition scoring systems are presented focussing on their main features and a comparison between them is performed. Characteristics of each system are discussed providing feedback and input to scoring systems designers. Comparing similar (and/or related) competitions helps to distinguish precisely the advantages and disadvantages of each scoring system in a similar set-up and getting informed conclusions. The narrow set of competitions chosen has to do also with the deep insight of the authors. Extending the analysis to unrelated and very different competitions would go out of the scope of this paper and is more fit for a survey journal paper. The remaining of the paper will be organized as follows. Section II describes the SAUC-E scoring system, Section III and IV detail the euRathlon competition scoring for its 2014 and 2015 edition respectively. Section V presents the ERL Emergency Robots 2017 scoring and benchmarking framework. Section VI compares each approach. Finally, Section VII concludes the paper.

II. SAUC-E

The first competition analysed is SAUC-E [3]. It is an underwater robots competition dedicated to student teams. It started in 2006 in the UK at Pinewood Studios and after several editions in the UK and France, has been hosted by the NATO Science and Technology Organization (STO) Centre for Maritime Research and Experimentation (CMRE) in La Spezia, Italy, without interruption since 2010 (except in 2015 and 2017 in Piombino, Italy). Each year SAUC-E challenges multidisciplinary University teams (at least of 75% student members) to design and build Autonomous Underwater Vehicles (AUVs) capable of performing realistic missions, such as autonomous navigation, structure inspection and "black box" search. SAUC-E is recognized as the most realistic underwater robotics student competition [6] as it takes place in an open water sea basin. The changing conditions of low visibility, turbulence, salty water, tides, and currents have provided a perfect training ground for the young engineers over the years. Since 2013, teams could participate with Autonomous Surface Vehicles (ASVs) for a cooperative task. This task could be performed with an AUV from one team and an ASV from another team bringing points for both teams. This fostered cooperation among different teams and a fair play spirit.

The scoring system is a combination of performance measures (in-water) and subjective measures (static judging and others). Teams get points not only for their performance during the trials but also for their preparation, academic level, etc. As subjective measures, judges evaluate a team description paper and a video (submitted in the application process), and then Technical Merit, Craftsmanship, Safety of Design and Innovation. Subjective measures are judged by observation during the trials and by Static Judging, an assigned timeslot where students present their team. This aims to improve their presentation skills and other soft skills. Other subjective measures include Impress the judges points (for special performance, etc) and Discretionary Points (awarded after the last run) that judges can assign to a team at their discretion. Finally, a team could get bonus or penalty points depending on the weight of their vehicle. This was included in order to stimulate innovation and simple yet effective solutions.

These subjective measures sum around one-third of the total points that the team could get in the Finals. Another characteristic of this competition is that points could be scored throughout the week in practice and qualification sessions preceding the Finals. The difference in the Finals is that teams could get bonus points for performing several tasks sequentially while during practice and qualification, teams could attempt to perform each task individually. Partial scoring of tasks was also assigned in case of uncompleted missions adding flexibility to the scoring system. The final

scoring is the sum of what the team accomplished in the different sessions.

III. EURATHLON 2014

euRathlon[6] was a European project funded by the Framework Program 7 (FP7) running from 2013 to 2015. It was an outdoor robotics competition that invited teams to test the intelligence and autonomy of their robots in challenging scenarios. Specifically, the project conducted three different competitions: a land robotics competition carried out in Berchtesgaden, Germany, in September 2013; a marine robotics competition held in La Spezia, Italy in 2014 and finally, the euRathlon Grand Challenge involving the cooperation of robots from the sea, land and air domains held in Piombino, Italy in 2015. euRathlon 2014 took place in the same place as SAUC-E 2014 although the tasks and philosophy of the competition were different which was also reflected in the scoring system (detailed in [7]).

In euRathlon 2014, each task was considered a competition (scenario) in itself and thus every day a new task was judged separately. These tasks were high-level missions composed of smaller sub-tasks (e.g. map a simulated plume composed of buoys or inspect an underwater structure). The last day, similar to SAUC-E, a combined scenario with all the tasks was proposed. This system allows for a precise evaluation of each task in itself and to understand which team is better at what. The tasks were mostly similar to SAUC-E but in some cases more advanced (e.g. underwater manipulation). This was due mainly to two reasons. euRathlon 2014 was not a competition for students only and thus mixed teams academia-industry were allowed to participate raising the competition bar. The other important reason was that euRathlon 2014 (as well as euRathlon 2013) served as preparatory competition to the euRathlon 2015 multi-domain Grand Challenge and thus the competition scenarios had increased difficulty.

Similarly, the scoring system was designed with this in mind. At the same time, the scoring system took into account the need of benchmarking as this was one of the goals of the euRathlon project. Subjective measures are used in a similar way as SAUC-E and accounted for less than 20% of the maximum performance measures points in each scenario. For what concerns performance measures, a timing bonus was introduced as timing was one of the benchmark parameters that the judges wished to evaluate. For some sub-tasks a binary evaluation was performed and either the team would get zero points or the maximum for that particular sub-task. In other sub-tasks instead, a graduation between zero and 100% was used (multiplied by the number of maximum points). This applied mainly to complex tasks like mapping that cannot be broken down in simple binary sub-tasks (see Figure 1) without losing evaluation accuracy. Binary evaluation can be too reductive and does not allow an accurate evaluation of the team. For instance, for evaluating robot navigation, one can be interested in the error with respect to the ground truth. An absolute error of 5 meters or 20 meters can be evaluated differently using a real value between 0 and 100% to distinguish a better performing team. If the evaluation is binary (e.g. error smaller/bigger than 4 meters), 5 or 20 meters error would be evaluated the same way penalising the most accurate navigation. Indeed in the navigation scenario, the distance to waypoints, the total path length and time were used to define precise metrics (not binary) [7]. This quantitative analysis makes the judging more complex, burdening the judges but allows a more precise evaluation (with respect to SAUC-E).

[6]http://www.eurathlon.eu

Leak localisation and structure inspection

Performance Measures	Parameters for task	Value	Max Score
Path Finding	Quality of the Map	0-100%	600
	Ability to use map to find plume	0/1	400
Structure Localisation and inspection	Structure found using plume tracking	0/1	500
	Structure found using other approach	0/1	200
	Time Bonus (m)	10	300
	Full structure tracked and inspected	[0-100]	500
	Stopcock detected	0/1	250
	Stopcock inspected	[0-100]	250
	Quality of map of a structure produced	[0-100]	1000
	Quality of Overall Map	0-100%	500
AUV/USV Collaboration	Bonus	0/1	500

Fig. 1. Scoring table for Scenario 3 of euRathlon 2014

Benchmarks were extracted in a post-competition analysis as detailed in [7] and were independent of the scores in this case. As benchmark metrics, the performance was measured by the Completeness Grade, Timing Grade, Accuracy Grade, and Repetitiveness Grade. In this case, a qualitative analysis was performed by aggregating different metrics (as in the navigation example) and grading the teams in 5 different grades for each metric. This provides feedback to the teams in a graphic way that shows what should be the priority aspects for improvement (accuracy, timing, completeness, etc.).

IV. EURATHLON 2015

The euRathlon 2015 Grand Challenge [2] was the final competition of the euRathlon project and it was the first world competition with land, sea and air robots cooperating in a search and rescue scenario inspired by the 2011 Fukushima accident. This multi-domain competition had different scenarios in increasing order of complexity: single trials, double-domain scenarios (land+sea, land+air, sea+air) and finally the Grand Challenge (three domains). When it comes to double or three domains competitions, one has not only to judge each domain individually but also robot cooperation. Points were given to each cooperative task when direct robot-to-robot communication was used. Another important issue is the unstructured environment of this outdoor competition that took place in the surroundings of a power plant (using an open water harbour, a beach and a ruined building) in Piombino, Italy (see Fig. 2). The

environment changes easily throughout the day so any exact repeatability is hard to achieve (e.g. light changes, wind changes, wave changes). This complicates the benchmarking (which was one of the goals of euRathlon project). Another difficulty is to have a reasonable number and complexity of achievements per domain and a balance among the different domains so a team should perform well in all of them to win the competition.

Fig. 2. Aerial view of the euRathlon 2015 environment

The solution was to design the scoring system having the benchmarking needs in mind. The scoring and benchmarking was thus inspired by the framework developed within the RoCKIn [8] EU FP7 project and includes task benchmarks (system-level) and functionality benchmarks (module-level). Each scenario was considered a task benchmark (TBM). Each task can implement multiple functionalities and each functionality can be evaluated across multiple tasks and domains (see Figure 3). Task benchmarking (TBM) focuses on task achievements. Tasks are decomposed into a set of subgoals or subtasks that are used to evaluate the performance of teams. TBMs were also used to score immediately the teams in a ranking focussed on task accomplishment. For instance, the winner of the Grand Challenge was the team ranked first in the Grand Challenge scenario (or TBM). Thus, scoring and task benchmarking become the same in this case (different from euRathlon 2014). Instead, Functionality benchmarking (FBM) dealt with how well does a team perform on a particular functionality (e.g. object detection). FBMs were based on the data collected during the competition but were evaluated in post-processing. Not all Functionalities were evaluated in all domains and tasks as Figure 3 shows. For instance, 2D Mapping was evaluated for land and air domains (but not for sea due to the difficulty of ground truthing underwater). Instead, Object Recognition could be evaluated across all tasks and domains.

The Task Benchmarking in the euRathlon 2015 competition is based on the concept of Performance Class (PC) used as the main element for the ranking and scoring. The Task Benchmarks consist of following elements: PC, AC (Autonomy Class), P (Penalties) and T (Time). The Performance Class (PC) measures how well a robot performs in a specific task and is determined by the number of

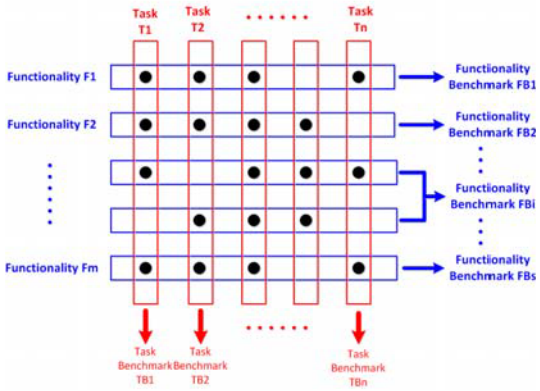

Fig. 3. Task and Functionality Benchmark as in [8]

Achievements (or sub-goals) and Optional Achievements (OA) that the robot achieves during the execution of the task. These achievements are similar to the subtasks used in euRathlon 2014 scoring and depending on the type of achievement can be either binary or a quantized index from 0 to 1 multiplied by a weight. For instance, an achievement such as reaching a given location was binarily evaluated while the quality of mapping was index-based (including factors such as area coverage and the number of objects found). To give an example, a robot that achieves 5 binary Achievements, 2 Optional Achievements and one achievement with a quantized index of 0.5 (weighted by 4, e.g. mapping), has a PC of 9.

The Autonomy Class (AC) measures how well a robot performs in terms of Autonomy. A robot gains an extra autonomy point if a subtask is performed autonomously, or an extra half point if the task is performed semi-autonomously. Penalties are behaviours that robots should not do and the most important ones (Key Penalties) decrease the score. Penalties are important to avoid teams to take advantage by circumventing rules, put safety in peril or overuse human intervention (thus promoting autonomy too). Time is taken into account in case two teams have the same score. Contrarily to previously described systems, the research paper was not included in the scoring although it was needed to enter the competition. There was no static judging either (by observation). A small number of points was given to the teams that provided a video of the vehicle operating before the competition. Finally, given the nature of the different achievements in different domains, each achievement or optional achievement and penalties have an associated weight. This is important as it allows giving more importance to harder subtasks and to weigh less simple achievements.

The final scoring formula is $S = round(V + \sum_{i=1}^{n} A_i * W_i + \sum_{j=1}^{m} O_j * W_j + 0.5 * \sum_{k=1}^{u} AC_k * W_k - \sum_{l=1}^{v} KP_l * W_l)$ Where V is the fixed points for the provided video describing the vehicle(s), n is the number of the Achievements, m is the number of the Optional Achievements, u

is the number of achievements (including Achievements and Optional Achievements) done autonomously (one point) or semi- autonomously (half point); KP is the number of Key Penalising behaviours. W is its achievement weight. This scoring system balances benchmarking with a relatively easy to use subtasks division and fairness of results (by weighting differently each achievement and decreasing the score with penalties).

V. ERL Emergency 2017

The European Robotics League (ERL)[7] Emergency Robots [9] is an outdoor robotics competition funded by the Horizon 2020 Program from EU with a focus on realistic cooperative search and rescue response scenarios for land, sea and air robots. It is the follow-up of the euRathlon competition and it is as well inspired by the 2011 Fukushima disaster. The ERL Emergency Robots competition was part of the RockEU2 Coordination Action from 2016 to 2018, led by euRobotics and supported by SPARC. Since February 2018 it will be part of the SciRoc EU project and will be integrated with the other ERL leagues (Service and Industrial Robots) in a Smart City context. The ERL Emergency Robots Major Tournament took place in Piombino, Italy in September 2017 in the same scenario of euRathlon 2015 Grand Challenge. As euRathlon, ERL Emergency Robots was not a student-based competition and mixed teams or industry teams were allowed.

Besides new tasks for each domain and for inter-domain cooperation, a novelty with respect to the euRathlon 2015 edition was the introduction of a new scoring and benchmarking methodology harmonized with the other ERL competitions and inspired by the work carried out in the RoCKIn project [8]. The division between Task Benchmarks and Functionality Benchmarks was deepened but there were important differences.

Benchmarking and scoring are the same in ERL Emergency. This means the scoring system needs to reflect an ideal benchmarking system that can be repeated across different trials and even competitions. This has to do with the fact that ERL Emergency Robots is moving towards portability. To have a portable competition, one needs to have not only a similar setup/environment but also simply defined achievements, replicable and easy to evaluate in different physical locations. With such a system, a yearly ranking composed of results obtained in similar competitions (same scenarios) can be established as the competitions become comparable.

Therefore, with respect to euRathlon 2015, for TBMs, there were no video points and only achievements, penalties and time count. The autonomy class was not included either *per se*, as achievements were divided in the ones that had to be autonomous and the ones that did not have. The performance class was similarly defined as the number of achievements that the robot achieves. There was no difference between achievements and optional achievements and the biggest change was that all achievements were binary and no weights were used. This means that the scoring system becomes a series of yes or no (1 or 0) achievements which can be easier to understand and use by the judges and teams. Not only that, but a simple binary based scoring can be easily replicated in other competitions fulfilling the goal of portability and repeatability.

The performance class of a team is the sum of achievements accomplished. The ranking of a TBM is based on performance classes (which are teams scores). In case of a tie, penalties are used to untie two teams. If the number of penalties is also the same, then time is used to untie teams. With such a system, achievements need to be very well defined and broken down in order to bring fairness to the competition and highlight the most important/difficult subtasks. For instance, it would not be fair to make the mapping of a large area count as much as passing through a gate/door. Therefore, and to obviate the lack of weights, the solution was to divide even further the subtasks. In the case of mapping, this was accomplished by dividing the mapping area into several smaller areas, each of them counting as one achievement. The impact of the scoring and benchmarking system in tasks design could also be noticed for what regards autonomy. As there was no autonomy class, tasks could not be optionally done autonomously or not be given different weights. Tasks that the organizing committee retained important to be autonomous were mandatory to be executed autonomously and other tasks could be executed as the teams wished (with no extra points for a higher degree of autonomy).

This has to do with the fact that the TBMs judge task accomplishment. How well a team was able to execute those same tasks was evaluated looking at individual functionalities in the FBMs. The FBMs were evaluated post-event through a rigorous analysis of the data collected during the TBMs trials.

VI. Comparison

Looking at the four different scoring systems presented in the previous sections and based on the actual results from those competitions, several conclusions can be drawn. First of all, the goal of the competition impacts strongly the scoring approach one should use. If the goal is education and the competition is dedicated to students, then the scoring system should also include an evaluation of team quality outside the competition field. This includes innovation, safety of design, quality of research paper, etc and is important not only to stimulate students to write good papers and to think in out-of-the-box solutions but also to train their academic skills.

At the same time, these subjective measures can weight significantly in the total score when the team performs badly during the actual competition. That happened for instance in SAUC-E'14 and to a much lesser extent in euRathlon 2014. In SAUC-E'14 one of the winning teams (*ex aequo* had over 45% of points coming from subjective measures and the two teams placed third (*ex aequo* had over 60%

[7] http://robotics-league.eu

of subjective measures points while in euRathlon 2014 for any given task the total amount of subjective measure points was never higher than 40% for the winning team with a minimum of 9% in the Combined Scenario and an average of 27.8% across the five scenarios.. This means that one has to adapt the weight of subjective measures when compared to in-water points to make sure the ranking reflects the actual performance and not just promoting "theoretically good" teams. Another important scoring system feature for student competitions such as SAUC-E is to give plenty of time for teams to score points (during qualification and practice days, not just in the Finals) and accept partial credits for attempts while promoting a more complex mission with sequential execution of several tasks and extra points in the Finals day. This stimulates also the team's creativity.

Of course, SAUC-E scoring system is not perfect. For instance, as disadvantages, one can argue that the discretionary points and impress the judges' points are too subjective and can lead to teams disagreements. Moreover, the intrinsic nature of the competition is education and training and thus it is not possible to perform benchmarking. Instead, in a competition such as euRathlon 2014 where each task was a competition in itself, one can actually compare teams for each task (going on the direction of benchmarking). Another advantage of euRathlon 2014 scoring system is that the fact that some achievements are then binary and no partial credits are allowed provides more objectivity (while not helping lower-performance teams). Furthermore, having a combined scenario on the final day permits to benchmark the overall performance of the team. Extracting Completeness Grade, Timing Grade, Accuracy Grade, and Repetitiveness Grade across different metrics gives a qualitative analysis of the level of the teams in the post-event analysis. The feedback provided to the teams through such a scoring system can also help them prepare for a more complex competition as they will know exactly on what they are good or not. By using subjective measures, the organizers can also have an idea of how ready teams are to prepare for a harder competition. This is important to help select teams for a Grand Challenge and to match teams across domains with similar characteristics. The issue with this system is that as each calendar day there is a different task to be done, if a team has some hardware issue on that particular day, it might not be able to prove that it was the best on that task and there is no second chance (opposite to scoring throughout the week). Another disadvantage is that by giving a series of tasks a team has to accomplish in different days it also means that the team needs to spread its testing and practice over several tasks, which can result in failing some of them. Also, adding the time factor to the scoring puts more pressure on teams. However, this raises the competition level and as euRathlon 2014 served as preparation for an even more complex multi-domain competition, it aligns with the goal of pushing the state of the art.

Moving to multi-domain competitions, judging and scoring become more complex. Having multi-domain teams in many cases formed by different institutions would prevent the heavy usage of subjective measures also taking into account that each domain has its own specificity. Nonetheless, interoperability can be pushed through multi-domain competitions [10]. Therefore, in euRathlon 2015 the only subjective measure was the video quality and the number of points for this was very small. As a first world competition of its type, the goals were very ambitious and a balance between immediate scoring (for a Grand Challenge style) and precise benchmarking was tried to be attained. The breakdown into achievements had to be more carefully designed with many subtasks and different weights for each achievement depending on their importance. There was also a need to balance the number and importance of achievements among the different domains. Autonomy was duly rewarded as well as cooperation as the two secondary goals (after task achievement) and the most important penalties would decrease the score in order to prevent teams gaining advantage from incorrect behaviour. Functionalities benchmarks were extracted *a posteriori* from the data collected. The mixture of binary and index-value achievements with different weights is an advantage as it allows to represent the reality in a better way and improves the fairness of the competition. Moreover, the advantage of being able to do scoring and benchmarking at the same time is appealing. The disadvantage of this system is perhaps its complexity given the formula based on many factors and weights and the difficulty of setting up the correct weights, penalties, and achievements.

Finally, in ERL Emergency Robots 2017, a complete integration between benchmarking and scoring was used. This choice was done taking into account harmonization of European competitions and portability of the competition for similar scenarios. The developed approach can now be used easily in future competitions even if the location changes and adapting/integrating it with a widening of the scope and/or other competitions will be easier. An important advantage is that this scoring system allowed for an easier judging even if at the same time increased the number of achievements available to teams (which might decrease the effective rate of accomplishments in a binary framework). For judges, it is much easier evaluating achievements by yes or no instead of using index values and evaluating map quality with metrics that require fiducial artificial markers [11]. For teams, the system might be more transparent. However, it is disadvantageous for competition designers, as the task of deciding each achievement becomes much harder, especially taking into account the fact that all achievements weight the same. Same applies to defining which achievements should be autonomous and which do not need to be. Another disadvantage is the risk of not being able to fully represent the performance of a team as having a good map or a lower quality one would both give one point to the team in the task benchmarks. Also, there are no weights in these TBMs as the goal is to evaluate the number of tasks a team can achieve. This is compensated by evaluating the efficiency while performing a given task looking at the different FBMs.

Nonetheless, even if the designers of the scoring system establish a similar number of achievements across domains,

the result can be very unbalanced as only the total number of achievements counts for the TBM ranking. For instance, if a team performs very poorly in a domain but achieves most of the achievements in another domain, it might be ranked better than a team that performed averagely in all domains. This happened for example in the TBM 1 - Grand Challenge as the team ranked second scored a much lower number of points in the marine domain than the third team (but ranked second due to their good performance in the other two domains). Alternative methods to balance correctly the weight of each domain must be studied. Another disadvantage intrinsic to the outdoor environment characteristic is that analysing benchmarks obtained under different environmental conditions can be misleading. Nonetheless, it is an effort worth to be taken if one wishes to give a more scientific component to competitions rather than educational or show-style.

VII. Conclusion

As seen in Section VI, there is no perfect generic scoring system. The way to score a team in a competition depends on the purpose of that competition and the target participants. It is very hard to design a system that can achieve all goals ideally intended by the competition designers. Nonetheless, some general conclusions can be taken. The first thing one has to bear in mind when devising the scoring system is what are the goals of that competition. Training, education, technological advancement, pushing scientific state-of-the-art, benchmarking, are all valid goals. Each competition has its own goals and the scoring system should reflect them. It can try to balance among several goals but it cannot fulfil all at the same time.

For instance, in student competitions like SAUC-E where the goal is training and education, it will be hard to do effective benchmarking as in ERL Emergency Robots 2017. Nonetheless, innovative technological advancement can come from student teams with small budgets. If instead the goal is to have precise benchmarks then one can forget subjective measures used in student competitions. If one wishes to push the scientific state-of-the-art, not only the achievement of a task but also the way this task was achieved should be rewarded (e.g. autonomously or not, with dead-reckoning or SLAM navigation etc) as in euRathlon 2014 and especially euRathlon 2015.

Benchmarking through competitions is not a trivial task as repeatability, a key point, is hard to achieve unless the competition is designed with benchmarking in mind. Even then, in unstructured outdoor environments with changing conditions this is hardly possible. Finally, the thrill of competing in Grand Challenges affects the teams making benchmarking purely the robots even harder (instead of evaluating heterogeneous human-robot teams).

In this paper, four different related competitions and their scoring systems highlighting main features and pros and cons of each were presented. None of these is the best scoring system and thus no absolute judgement can be given but the authors believe that these conclusions can inform competition designers and contribute to the field. Future work includes a deep analysis of each of these scoring systems to be published in a journal paper.

Acknowledgment

The authors wish to thank all the participants in all the past competitions for their effort during the events, the organization staff from CMRE and the euRathlon and RockEU2 project partners for their valuable input and collaboration. Furthermore, the authors would also like to thank the support of all the sponsors and supporters in particular IEEE Oceanic Engineering Society.

References

[1] G. Ferri, F. Ferreira, V. Djapic, Y. Petillot, M. Palau, and A. Winfield, "The euRathlon 2015 Grand Challenge: The first outdoor multi-domain search and rescue robotics competition - a marine perspective," *Marine Technology Society Journal*, vol. 50, no. 4, pp. 81–97, 2016.

[2] A. F. T. Winfield, M. P. Franco, B. Brueggemann, A. Castro, M. C. Limon, G. Ferri, F. Ferreira, X. Liu, Y. Petillot, J. Roning, F. Schneider, E. Stengler, D. Sosa, and A. Viguria, *euRathlon 2015: A Multi-domain Multi-robot Grand Challenge for Search and Rescue Robots*. Springer International Publishing, 2016, pp. 351–363.

[3] G. Ferri, F. Ferreira, and V. Djapic, "Boosting the talent of new generations of marine engineers through robotics competitions in realistic environments: The SAUC-E and euRathlon experience," in *OCEANS 2015 - Genova*, May 2015, pp. 1–6.

[4] E. Olson, J. Strom, R. Morton, A. Richardson, P. Ranganathan, R. Goeddel, M. Bulic, J. Crossman, and B. Marinier, "Progress toward multi-robot reconnaissance and the magic 2010 competition," *Journal of Field Robotics*, vol. 29, no. 5, pp. 762–792, 2012. [Online]. Available: http://dx.doi.org/10.1002/rob.21426

[5] F. E. Schneider, D. Wildermuth, and H. L. Wolf, "ELROB and EURATHLON: Improving search & rescue robotics through real-world robot competitions," in *2015 10th International Workshop on Robot Motion and Control (RoMoCo)*, July 2015, pp. 118–123.

[6] K. T. Harold Tay and V. Pallayil, "The Singapore AUV Challenge (SAUVC) 2016," in *IEEE OES Beacon Newsletter, May 2016, Volume 5, Number 2,*, 206.

[7] Y. Petillot, F. Ferreira, and G. Ferri, "Performance measures to improve evaluation of teams in the euRathlon 2014 sea robotics competition," *IFAC-PapersOnLine*, vol. 48, no. 2, pp. 224–230, 2015.

[8] F. Amigoni, E. Bastianelli, J. Berghofer, A. Bonarini, G. Fontana, N. Hochgeschwender, L. Iocchi, G. Kraetzschmar, P. Lima, M. Matteucci, P. Miraldo, D. Nardi, and V. Schiaffonati, "Competitions for Benchmarking: Task and Functionality Scoring Complete Performance Assessment," *IEEE Robotics and Automation Magazine*, vol. 22, no. 3, pp. 53–61, Sept. 2015.

[9] G. Ferri, F. Ferreira, and V. Djapic, "Multi-domain robotics competitions: the cmre experience from sauc-e to the european robotics league emergency robots," in *OCEANS 2017 - Aberdeen*, June 2017, pp. 1–7.

[10] D. S. Lopez, G. Moreno, J. Cordero, J. Sanchez, S. Govindaraj, M. M. Marques, V. Lobo, S. Fioravanti, A. Grati, K. Rudin, M. Tosa, A. Matos, A. Dias, A. Martins, J. Bedkowski, H. Balta, and G. D. Cubber, "Interoperability in a heterogeneous team of search and rescue robots," in *Search and Rescue Robotics - From Theory to Practice*. Rijeka: InTech, 2017, ch. 06. [Online]. Available: http://dx.doi.org/10.5772/intechopen.69493

[11] S. Schwertfeger, A. Jacoff, C. Scrapper, J. Pellenz, and A. Kleiner, "Evaluation of maps using fixed shapes: The fiducial map metric," in *Proceedings of PerMIS*, 2010.

Document Data Sheet

Security Classification		Project No.
Document Serial No.	**Date of Issue**	**Total Pages**
CMRE-PR-2019-021	May 2019	7 pp.

Author(s)

Fausto Ferreira, Gabriele Ferri, Yvan Petillot, Xingkun Liu, Marta Palau Franco, Matteo Matteucci, Francisco Javier Pérez Grau, Alan Ft Winfield

Title

Scoring robotic competitions: balancing judging promptness and meaningful performance evaluation

Abstract

To have a significant and fair competition an adequate scoring system is necessary. Different scoring systems exist, each one directly related to the nature and goals of the competition, e.g. student/educational, focused on benchmarking, Grand Challenge style, etc. However, the design of such systems is not an easy task. It is mandatory to design approaches which enable the judges to score teams in a reasonable time after the end of their performance. Scoring systems cannot be therefore extremely complex. On the other hand, it is crucial to have a judging system for teams which provides a meaningful and fair performance evaluation of what the teams achieved in the field. In this paper, several approaches to scoring are presented and compared. Our focus is on search and rescue competitions. Each approach is critically analysed and its features comparatively discussed. This analysis is beneficial to provide an overview of scoring systems tailored to different kinds of competitions. The reported examples can be used as building blocks to improve existing scoring systems or to design new approaches.

Keywords

Issuing Organization

NATO Science and Technology Organization
Centre for Maritime Research and Experimentation

Viale San Bartolomeo 400, 19126 La Spezia, Italy

[From N. America:
STO CMRE
Unit 31318, Box 19, APO AE 09613-1318]

Tel: +39 0187 527 361
Fax:+39 0187 527 700

E-mail: library@cmre.nato.int