RoboCup 2D Soccer Simulation League: Evaluation Challenges

Mikhail Prokopenko[1], Peter Wang[2], Sebastian Marian[3], Aijun Bai[4], Xiao Li[5] and Xiaoping Chen[5]

[1] Complex Systems Research Group, Faculty of Engineering and IT
The University of Sydney, NSW 2006, Australia
mikhail.prokopenko@sydney.edu.au
[2] Data Mining, CSIRO Data61, PO Box 76, Epping, NSW 1710, Australia
[3] Compa-IT, Romania
[4] Department of Electrical Engineering and Computer Sciences
University of California Berkeley
[5] Multi-Agent Systems Lab, School of Computer Science and Technology,
University of Science and Technology of China

Abstract. We summarise the results of RoboCup 2D Soccer Simulation League in 2016 (Leipzig), including the main competition and the evaluation round. The evaluation round held in Leipzig confirmed the strength of RoboCup-2015 champion (WrightEagle, i.e. WE2015) in the League, with only eventual finalists of 2016 competition capable of defeating WE2015. An extended, post-Leipzig, round-robin tournament which included the top 8 teams of 2016, as well as WE2015, with over 1000 games played for each pair, placed WE2015 third behind the champion team (Gliders2016) and the runner-up (HELIOS2016). This establishes WE2015 as a stable benchmark for the 2D Simulation League. We then contrast two ranking methods and suggest two options for future evaluation challenges. The first one, "The Champions Simulation League", is proposed to include 6 previous champions, directly competing against each other in a round-robin tournament, with the view to systematically trace the advancements in the League. The second proposal, "The Global Challenge", is aimed to increase the realism of the environmental conditions during the simulated games, by simulating specific features of different participating countries.

1 Introduction

The International RoboCup Federation's Millennium challenge sets an inspirational target that by mid-21st century, a team of fully autonomous humanoid soccer players shall win the soccer game, complying with the official rule of the FIFA, against the winner of the most recent World Cup [1]. In pursuit of this goal, the RoboCup Federation has introduced multiple leagues, with both physical robots and simulation agents, which have developed different measures of their progress over the years. The main mode, of course, is running competitions at the national, regional and world cup levels. In addition, however, various leagues have included specific evaluation challenges which not only complement the competitions, but also advance the scientific and technological

base of RoboCup and Artificial Intelligence in general. Typically, a challenge introduces some new features into the standard competition environment, and then evaluates how the teams perform under the new circumstances.

For example, during an evaluation round of RoboCup 2001 the rules of the soccer simulator were modified in such a way that "dashing on the upper half of the field resulted in only half of normal speed for all the players" [2]. This modification was not announced in advance, and while the changed conditions were obvious to human spectators, none of the simulation agents could diagnose the problem [2].

A specific technical challenge was presented by the so-called Keepaway problem [3], when one team (the "keepers") attempt to keep the ball away from the other team (the "takers") for as long as possible.

Later on, the focus of evaluation in RoboCup 2D Soccer Simulation League shifted from changing the physics of the simulation or the tactics of the game, to studying the diverse "eco-system" of the League itself, which has grown to include multiple teams. The Simulated Soccer Internet League (SSIL) was designed to allow a continual evaluation of the participating teams during the time between annual RoboCup events: pre-registered teams could upload their binaries to a server on which games were played automatically [4]. The SSIL was used at some stage as a qualification pathway to the annual RoboCup, but this practice was discontinued due to verification problems.

Several other challenges and technical innovations introduced in Soccer Simulation Leagues (both 2D and 3D), including heterogeneous players, stamina capacity model, and tackles, are described in [5]. This study further pointed out the importance of the online game analysis and online adaptation.

More recently, a series of "drop-in player challenges" was introduced by [6] in order to investigate how real or simulated robots from teams from around the world can cooperate with a variety of unknown teammates. In each evaluation game, robots/agents are drawn from the participating teams and combined to form a new team, in the hope that the agents would be able to quickly adapt to meaningfully play together without pre-coordination. The "drop-in" challenge was adopted by RoboCup Standard Platform League (SPL) and both RoboCup Soccer Simulation Leagues, 2D and 3D. In all the considered leagues, the study observed "a trend for agents that perform better at standard team soccer to also perform better at the drop-in player challenge" [6].

At RoboCup-2016 in Leipzig, several soccer and rescue leagues increased realism of the competition by holding their competitions outdoors. In the SPL, a separate competition was successfully held not on the customary green carpet but rather on an artificial turf, under diverse natural lighting conditions. Similarly, Middle Size Soccer League also successfully implemented a Technical Challenge under these difficult conditions, while the Humanoid League used artificial turf and real soccer balls[6].

In this paper, we describe the latest evaluation challenge, introduced by RoboCup 2D Soccer Simulation League [7, 8] in 2016, in order to trace the progress of the overall League. Furthermore, we describe two possibilities for future challenges: one intended to systematically trace the advancements in the League ("The Champions Simulation League"), and the other aimed to increase the realism of the environmental conditions during the simulated games ("The Global Challenge").

[6] http://robocup2016.org/press-releases/leipzig-best-place-for-robots-and-friends/452749

2 Methodology and Results

2.1 Actual competition

The RoboCup-2016 Soccer Simulation 2D League included 18 teams from 9 countries: Australia, Brazil, China, Egypt, Germany, Iran, Japan, Portugal and Romania. The last group stage was a round-robin tournament for top 8 teams. It was followed by the two-game semi-final round, a single-game final, and 3 more playoffs between third and fourth, fifth and sixth, and seventh and eighth places.

In the two-game semi-final round, HELIOS2016 (Japan) [9] defeated team Ri-one (Japan) [10], 3:0 and 4:0, while Gliders2016 (Australia) [11, 12] defeated team CSU_Yunlu (China) [13], winning both games with the same score 2:1.

The single-game final between HELIOS2016 and Gliders2016 went into the extra time, and ended with Gliders2016 winning 2:1.

The third place was taken by team Ri-one which won against CSU_Yunlu 3:0.

Oxsy (Romania) [14] took the fifth place, winning 4:0 against Shiraz (Iran) [15]; and MT2016 (China) [16] became seventh, winning against FURY (Iran) [17] on penalties 4:2. The final ranking of RoboCup-2016 (Leipzig, Germany) is shown in the left column of Table 1.

2.2 Ranking Estimation

Using the ranking estimation methodology established by [18, 19], we conducted an 8-team round-robin tournament for top 8 teams from RoboCup-2016. The estimation process used the released binaries of top RoboCup-2016 teams[7], where all 28 pairs of teams play approximately 4000 games against one another. The following *discrete* scheme was used for discrete point calculation:

- Firstly, the average score between each pair of teams (across all 4000 games) is rounded to the nearest integer (e.g. "1.2 : 0.5" is rounded to "1 : 1").
- Next, points are allocated for each pairing based on these rounded results: 3 for a win, 1 for a draw and 0 for a loss.
- Teams are then ranked by the sum of the points received against each opponent. The total goal difference of the rounded scores is used as a tie-breaker.

The final ranking $\mathbf{r^d}$ under this scheme is presented in Table 1.

In order to capture the overall difference between any two rankings $\mathbf{r^a}$ and $\mathbf{r^b}$, the L_1 distance is utilised [18]:

$$d_1(\mathbf{r^a}, \mathbf{r^b}) = \|\mathbf{r^a} - \mathbf{r^b}\|_1 = \sum_{i=1}^{n} |r_i^a - r_i^b| \,, \tag{1}$$

where i is the index of the i-th team in each ranking, $1 \leq i \leq 8$.

The distance between the actual ranking $\mathbf{r^a}$ and the estimated ranking $\mathbf{r^d}$ is

$$d_1(\mathbf{r^a}, \mathbf{r^d}) = |1-1|+|2-2|+|3-6|+|4-4|+|5-3|+|6-5|+|7-8|+|8-7| = 8.$$

[7] https://chaosscripting.net/files/competitions/RoboCup/WorldCup/2016/2DSim/binaries/

	Gliders	HELIOS	Ri-one	CSU_Yunlu	Oxsy	Shiraz	MT2016	FURY	Goals	Points	$\mathbf{r^d}$
Gliders		0.3 : 0.4	2.8 : 0.3	1.9 : 0.3	0.7 : 0.8	3.8 : 0.4	5.0 : 0.0	2.5 : 0.2	18 : 1	17	1
HELIOS	0.4 : 0.3		1.8 : 0.1	3.0 : 0.2	1.2 : 0.5	4.3 : 0.3	3.6 : 0.0	2.5 : 0.0	17 : 1	17	2
Ri-one	0.3 : 2.8	0.1 : 1.8		1.1 : 1.1	0.2 : 1.8	0.6 : 0.5	0.4 : 0.0	0.6 : 0.5	3 : 10	4	6
CSU_Yunlu	0.3 : 1.9	0.2 : 3.0	1.1 : 1.1		0.5 : 1.2	2.0 : 0.7	1.4 : 0.0	1.2 : 0.4	6 : 8	11	4
Oxsy	0.8 : 0.7	0.5 : 1.2	1.8 : 0.2	1.2 : 0.5		3.5 : 0.5	4.4 : 0.0	3.0 : 0.1	16 : 4	15	3
Shiraz	0.4 : 3.8	0.3 : 4.3	0.5 : 0.6	0.7 : 2.0	0.5 : 3.5		0.5 : 0.1	0.8 : 1.0	5 : 16	5	5
MT2016	0.0 : 5.0	0.0 : 3.6	0.0 : 0.4	0.0 : 1.4	0.0 : 4.4	0.1 : 0.5		0.0 : 0.0	0 : 15	2	8
FURY	0.2 : 2.5	0.0 : 2.5	0.5 : 0.6	0.4 : 1.2	0.1 : 3.0	1.0 : 0.8	0.0 : 0.0		2 : 12	3	7

Table 1. Round-robin results (average goals scored and points allocated with the discrete scheme) for the top 8 teams from RoboCup 2016, ordered according to their final actual competition rank, $\mathbf{r^a}$. The scores are determined by calculating the average number of goals scored over approximately 4000 games rounded to the nearest integer, then awarding 3 points for a win, 1 point for a draw and 0 points for a loss. The resultant ranking is marked with $\mathbf{r^d}$.

The top two teams were fairly close in their performance (confirmed by the final game, which needed extra time). Similarly the 7th and 8th teams were similar in strength too (not surprisingly their playoff ended up with penalties). The main discrepancy between the actual and estimated results is due to performances of two teams: Oxsy (whose rank is estimated as third, while the actual rank was only fifth) and Ri-one (which finished the competition as third, while its average rank is estimated to be sixth).

2.3 A critique of the continuous ranking scheme

There exists another ranking method: *continuous* scheme [18, 19]:

- Teams are ranked by the sum of average points obtained against each opponent across all 4000 games.
- The total goal difference of the non-rounded scores is used as a tie-breaker.

Both schemes, discrete and continuous, were introduced in order to evaluate different competition formats, using the top 8 teams of 2012 and 2013 [18, 19]. However, over the years it has become apparent that the continuous scheme suffers from two major drawbacks, violating the balance of points (3 for a win, 1 for a draw and 0 for a loss) and overestimating the points for draws and losses. Specifically, under the continuous scheme:

1. there is a bias to attribute more points to draws with higher scores.
2. there is a bias to reduce the advantage of the three-points-for-a-win standard.

1. Let us consider two opposite scenarios: (i) two teams A and B of equal strengths, denoted $A \Leftrightarrow B$, but with a stronger defensive capability, play N games resulting in the average $0 : 0$ score; and (ii) two teams X and Y of equal strengths $X \Leftrightarrow Y$, but with a stronger attacking capability, play N games resulting in the average $q : q$ score, where $q > 0$ is sufficiently large, e.g., $q = 3$. In the first pair, the scores of individual games, which may or may not be draws, do not diverge much from $0 : 0$, as the teams are defensive. And so the actual drawn scores $0 : 0$ dominate among the results, with

large outliers $k:0$ or $0:k$, for $k>0$ being relatively rare. Thus, the continuous points p attained by teams A and B stay close to 1.0, for example, $p_A \approx p_B \approx 1.2$.

In the second pair, the scores of individual games, which again may or may not be draws, diverge more from the average $q:q$, due to a higher variability of possible high scores. Consequently, the proportion of actual draws among N games is much smaller in comparison to the first pair, and the large outliers $k:0$ or $0:k$, even for $k>q$, are more numerous. As a result, the teams X and Y exchange wins and losses more often than teams A and B, acquiring more points for their respective wins. This yields the continuous points p_X and p_Y significantly higher than 1.0, for example, $p_X \approx p_Y \approx 1.4$, creating a general bias to attribute more points to the drawn contests with higher scores: $p_A \approx p_B < p_X \approx p_Y$. A typical sample of 10,000 scores $q_1 : q_2$, where both q_1 and q_2 are normally distributed around the same mean q, with the standard deviation $\sigma = 1.0$, results in the following continuous points $p_{\Leftrightarrow}(q)$ for different draws around q: $p_{\Leftrightarrow}(0) = 1.23$ for draws $0.38:0.38$, $p_{\Leftrightarrow}(1) = 1.33$ for draws $1.07:1.08$, $p_{\Leftrightarrow}(2) = 1.36$ for draws $1.99:2.00$, and $p_{\Leftrightarrow}(3) = 1.38$ for draws $3.02:3.00$.

While the higher scoring teams may be expected to get an advantage at a tie-breaker stage, getting more continuous points for the same outcome is obviously an unfair bias. The discrete scheme does not suffer from this drawback as the average scores are converted into the identical discrete points immediately, i.e., $p_A = p_B = p_X = p_Y = 1.0$.

It is easy to see that the lower bound for the continuous points shared by any two teams of equal strength is $\inf_{\Leftrightarrow} = 1.0$ (attainable only if all N games are drawn), while the upper bound is $\sup_{\Leftrightarrow} = 1.5$ (attained in the extreme case when all N games are non-draws, with wins and losses split equally). Consequently, under the continuous scheme, the points attributed to equal teams drawing on average, are overestimated, being somewhere between the lower and upper bounds: $\inf_{\Leftrightarrow} < p < \sup_{\Leftrightarrow}$, while the expected result (one point) sits only at exactly the lower bound.

2. The "three-points-for-a-win" standard which was widely adopted since FIFA 1994 World Cup finals "places additional value on wins with respect to draws such that teams with a higher number of wins may rank higher in tables than teams with a lower number of wins but more draws"[8]. To illustrate the second drawback of the continuous scheme we will contrast two scenarios, comparing the combined points of two drawn contests against the combination of one-won and one-lost contests.

Firstly, we consider a case when team Q is paired with teams U and Z, such that $Q \Leftrightarrow U$ and $Q \Leftrightarrow Z$. We do not expect transitivity, and so $U \Leftrightarrow Z$ is not assumed. The continuous points for team Q resulting from these two iterated match-ups, both drawn, could vary between these lower bound ($\inf_{\Leftrightarrow,\Leftrightarrow}$) and upper bound ($\sup_{\Leftrightarrow,\Leftrightarrow}$):

$$\inf_{\Leftrightarrow,\Leftrightarrow} = \inf_{\Leftrightarrow} + \inf_{\Leftrightarrow} = 2.0$$

$$\sup_{\Leftrightarrow,\Leftrightarrow} = \sup_{\Leftrightarrow} + \sup_{\Leftrightarrow} = 3.0$$

Typically the combined points vary around the level of $p_Q \approx 2.6$, which is an overestimation of the ideal outcome by more than half-a-point.

Secondly, we consider a scenario with team R matched-up against teams V and W, with team V being weaker than R, denoted $R \Rightarrow V$, while the team W is stronger

[8] https://en.wikipedia.org/wiki/Three_points_for_a_win

than R, denoted $R \Leftarrow W$. The relative strength of V and W is not important for our comparison. The continuous points that team R attains from the first pair, against the weaker opponent V, are bounded by $\inf_{\Rightarrow} = 1.5$ (just a slight over-performance) and $\sup_{\Rightarrow} = 3.0$ (the total dominance with all N games won):

$$1.5 = \inf_{\Rightarrow} < p_R < \sup_{\Rightarrow} = 3.0 \; .$$

In practice, the stronger team rarely drops below $p_R \approx 2.0$ points. In the second pair, team R is weaker, and its continuous points are bounded by $\inf_{\Leftarrow} = 0.0$ (the total inferiority with all N games lost) and $\sup_{\Leftarrow} = 1.5$ (getting almost to an equal standing):

$$0.0 = \inf_{\Leftarrow} < p_R < \sup_{\Leftarrow} = 1.5 \; .$$

In practice, the weaker team rarely reaches beyond $p_R \approx 1.0$ points. A typical sample of 10,000 scores $q_1 : q_2$, where q_1 and q_2 are normally distributed around the means q and 0.0 respectively, with the standard deviation $\sigma = 1.0$, results in the following continuous points $p_{\Rightarrow}(q)$ for different won contests around q: $p_{\Rightarrow}(1) = 2.31$ for wins $1.07 : 0.38$, $p_{\Rightarrow}(2) = 2.75$ for wins $2.00 : 0.38$, and $p_{\Rightarrow}(3) = 2.94$ for wins $2.97 : 0.38$. Correspondingly, the continuous points $p(q)$ for the respective lost contests sampled under the same distribution are overestimated above 0.0 as follows: $p_{\Leftarrow}(1) = 0.32$, $p_{\Leftarrow}(2) = 0.13$, and $p_{\Leftarrow}(3) = 0.04$.

The combined continuous points for team R after these match-ups, one won and one lost, could vary between the lower bound of and the upper bound of

$$\inf_{\Rightarrow,\Leftarrow} = \inf_{\Rightarrow} + \inf_{\Leftarrow} = 1.5$$
$$\sup_{\Rightarrow,\Leftarrow} = \sup_{\Rightarrow} + \sup_{\Leftarrow} = 4.5$$

In practice, $2.0 < p_R < 4.0$. That is, the combined continuous points of a win and a loss typically vary around $p_R \approx 3.0$, which is an appropriate outcome.

Contrasting the possible bounded intervals and typical outcomes of two contests (two draws versus one win and one loss) immediately highlights that the continuous points do not differentiate these scenarios sufficiently well. The intention of the three-points-for-a-win standard was precisely to preference the one-win-and-one-loss scenario over the two-draws scenario, $p_{\Rightarrow,\Leftarrow} = 3 > p_{\Leftrightarrow,\Leftrightarrow} = 2$. In other words, team Q with two drawn contests should achieve a lower rank than team R with a won and a lost contest, with the difference being the cost of a single drawn game. The continuous scheme fails in this regard, by producing, on average, less than half-a-point difference, $p_{\Rightarrow,\Leftarrow} \approx 3.0 > p_{\Leftrightarrow,\Leftrightarrow} \approx 2.6$. In fact, it is quite conceivable that $p_{\Rightarrow,\Leftarrow}$ could happen to be less than $p_{\Leftrightarrow,\Leftrightarrow}$ under the continuous scheme in some cases, as $\inf_{\Rightarrow,\Leftarrow} < \sup_{\Leftrightarrow,\Leftrightarrow}$. In other words, one hard-won contest, e.g. $p_{\Rightarrow}(1) = 2.31$, coupled with a serious loss, e.g., $p_{\Leftarrow}(3) = 0.04$ could earn less points (e.g., $p_{\Rightarrow,\Leftarrow} \approx 2.35$) than two high-scoring draws, e.g. $p_{\Leftrightarrow}(3) = 1.38$ (resulting in $p_{\Leftrightarrow,\Leftrightarrow} \approx 2.76$) — definitely, something not intended by the three-points-for-a-win standard.

Again, the discrete scheme easily overcomes this drawback as the average scores are converted into the appropriate discrete points for each contest (3 for a win, 1 for a draw and 0 for a loss), and combined only afterwards.

The two problems identified for the continuous scheme may amplify over many match-ups in a 8-teams round-robin, especially when there are many teams of similar strength (which is the case in the Simulation League in recent years). The biases become even more pronounced in the absence of transitivity in teams' relative strengths. In light of these concerns, we suggest that some recent works employing the continuous scheme, e.g. [20], would benefit from re-evaluation.

2.4 Evaluation round

The 2016 competition also included an evaluation round, where all 18 participating teams played one game each against the champion of RoboCup-2015, team WrightEagle (China), i.e., WE2015 [21]. Only two teams, the eventual finalists Gliders2016 and HELIOS2016, managed to win against the previous year champion, with Gliders defeating WrightEagle 1:0, and HELIOS2016 producing the top score 2:1.

We extended this evaluation over 1000 games, again playing WE2015 against the top 8 teams from RoboCup-2016. Table 2 summarises the evaluation for RoboCup-2016: both actual scores obtained in Leipzig and the averages over 1000 games.

	Gliders2016	HELIOS2016	Ri-one	CSU_Yunlu	Oxsy	Shiraz2016	MT2016	FURY
WE2015	0 : 1	1 : 2	7 : 1	2 : 0	4 : 1	3 : 2	4 : 0	11 : 2
WE2015	1.4 : 1.8	1.3 : 1.7	5.0 : 0.5	2.7 : 0.5	3.5 : 1.3	4.0 : 0.8	5.9 : 0.0	4.8 : 0.4

Table 2. Evaluation round results for the top 8 teams playing against WE2015. Top row: actual scores obtained at RoboCup-2016 in Leipzig; bottom row: average scores over 1000 games.

The evaluation round confirmed the strength of RoboCup-2015 champion in the League. It is evident that WE2015, if entered in 2016, would likely have achieved third rank. To confirm this conjecture we combined the estimation results presented in Table 1 with the estimates of WE2015 scores from Table 2, summarised in Table 3.

	Gliders	HELIOS	WE2015	Ri-one	CSU_Yunlu	Oxsy	Shiraz	MT2016	FURY	Goals	Points	$\mathbf{r^e}$
Gliders		0.3 : 0.4	1.8 : 1.4	2.8 : 0.3	1.9 : 0.3	0.7 : 0.8	3.8 : 0.4	5.0 : 0.0	2.5 : 0.2	20 : 2	20	1
HELIOS	0.4 : 0.3		1.7 : 1.3	1.8 : 0.1	3.0 : 0.2	1.2 : 0.5	4.3 : 0.3	3.6 : 0.0	2.5 : 0.0	19 : 2	20	2
WE2015	1.4 : 1.8	1.3 : 1.7		5.0 : 0.5	2.7 : 0.5	3.5 : 1.3	4.0 : 0.8	5.9 : 0.0	4.8 : 0.4	29 : 8	18	3
Ri-one	0.3 : 2.8	0.1 : 1.8	0.5 : 5.0		1.1 : 1.1	0.2 : 1.8	0.6 : 0.5	0.4 : 0.0	0.6 : 0.5	4 : 15	4	7
CSU_Yunlu	0.3 : 1.9	0.2 : 3.0	0.5 : 2.7	1.1 : 1.1		0.5 : 1.2	2.0 : 0.7	1.4 : 0.0	1.2 : 0.4	7 : 11	11	5
Oxsy	0.8 : 0.7	0.5 : 1.2	1.3 : 3.5	1.8 : 0.2	1.2 : 0.5		3.5 : 0.5	4.4 : 0.0	3.0 : 0.1	17 : 8	15	4
Shiraz	0.4 : 3.8	0.3 : 4.3	0.8 : 4.0	0.5 : 0.6	0.7 : 2.0	0.5 : 3.5		0.5 : 0.1	0.8 : 1.0	6 : 20	5	6
MT2016	0.0 : 5.0	0.0 : 3.6	0.0 : 5.9	0.0 : 0.4	0.0 : 1.4	0.0 : 4.4	0.1 : 0.5		0.0 : 0.0	0 : 21	2	9
FURY	0.2 : 2.5	0.0 : 2.5	0.4 : 4.8	0.5 : 0.6	0.4 : 1.2	0.1 : 3.0	1.0 : 0.8	0.0 : 0.0		2 : 17	3	8

Table 3. Evaluation round-robin results (average goals scored and points allocated with discrete scheme), combined for the top 8 teams from RoboCup 2016 and the RoboCup-2015 champion (WE2015). The resultant evaluation ranking is marked with $\mathbf{r^e}$.

	Gliders2016	WE2015	WE2014	WE2013	HELIOS2012	WE2011	Goals	Points	$\mathbf{r^l}$
Gliders2016		1.8 : 1.4	1.8 : 1.3	1.7 : 0.9	1.2 : 0.1	2.0 : 1.0	9 : 4	15	1
WE2015	1.4 : 1.8		2.5 : 2.5	3.0 : 2.5	2.2 : 0.9	4.0 : 2.9	13 : 12	8	2
WE2014	1.3 : 1.8	2.5 : 2.5		2.8 : 2.6	2.3 : 0.8	3.9 : 3.0	13 : 12	8	3
WE2013	0.9 : 1.7	2.5 : 3.0	2.6 : 2.8		1.9 : 0.9	2.9 : 3.2	12 : 12	6	4
HELIOS2012	0.1 : 1.2	0.9 : 2.2	0.8 : 2.3	0.9 : 1.9		2.6 : 1.8	6 : 9	3	5
WE2011	1.0 : 2.0	2.9 : 4.0	3.0 : 3.9	3.2 : 2.9	1.8 : 2.6		12 : 16	1	6

Table 4. Champions Simulation League round-robin results (average goals scored and points allocated with discrete scheme), for six champions of RoboCup 2011 to 2016. To distinguish WE2015 and WE2014 results, non-rounded scores were used as a tie-breaker. The resultant league ranking with discrete point allocation scheme is marked with $\mathbf{r^l}$.

3 Proposed challenges

3.1 Champions Simulation League

In order to trace the progress of the League over time it is interesting to compare performance of several previous champions, directly competing against each other in a round-robin tournament. For example, we evaluated relative performance of six champions of RoboCup-2011 to RoboCup-2016: WrightEagle (WE2011 [22, 23], WE2013 [24, 25], WE2014 [26], WE2015 [21]), HELIOS2012 [27] and Gliders2016 [11, 12].

The round-robin results over 1000 games, presented in Table 4, confirmed the progress of the League over the last six years, with the resultant ranking $\mathbf{r^l}$ completely concurring with the chronological ranking $\mathbf{r^t}$, i.e., $d_1(\mathbf{r^l}, \mathbf{r^t}) = 0$.

3.2 Global Challenge

Another proposal suggests to pit together the best teams from each of the top 6 or 8 participating countries (for example, in 2016 it would have been Australia, Brazil, China, Egypt, Germany, Iran, Japan, Romania), with two "home-and-away" games between opponents. There can be 14 games for a home-and-away single-elimination round with 8 teams; or 30 games for a home-and-away double round-robin with 6 teams. The "Global Challenge" will be distinguished from the main competition by playing the games with different parameters, for example, higher noise, or even with random player(s) disconnecting. In other words, the Global Challenge will focus on resilience of the teams in the face of unexpected conditions.

In each game, the home side would choose a hidden parameter to vary, in order to represent some features of their country (like high altitude in Bolivia or long-distance travel to Australia). These parameters will not be known to the opposition, but would be the same for both teams in that game.

The full list of possible hidden server parameters may include a significant number (currently, the number of server parameters is 27) and the set of changeable parameters will be agreed in advance. The global challenge mode will be selected via a new parameter, for example, server::global_challenge_mode, introduced in the simulation server (server.conf). When the global_challenge_mode parameter is set to true, the server will permit the left side coach (the home side) to send a command like this: (change_player_param (param_1 value) (param_2 value). . . (param_N value)).

For example, if the home side chooses to simulate some bad weather conditions or a soggy pitch, these server parameters can be changed: ball_accel_max, ball_decay, ball_rand, ball_speed_max, catch_probability, inertia_moment, kick_rand, player_rand.

Exploiting their own strong points, and possibly trying to exploit some weak points of the opponent, the home side could change some of the available parameters in a way that creates an advantage. While the adjusted environment will be applied equally to the both teams, the task of the left side coach (the home team) will be to optimise the choice of the adjusted parameters to maximise the home side advantage.

4 Conclusion

We summarised the results of RoboCup-2016 competition in the 2D Soccer Simulation League, including the main competition and the evaluation round. The evaluation round confirmed the strength of RoboCup-2015 champion (WrightEagle, i.e. WE2015) in the League, with only eventual finalists of 2016 (Gliders2016 and HELIOS2016) capable of winning against WE2015. After the RoboCup-2016, we extended this evaluation, over 1000 games for each pair, in a multi-game round-robin tournament which included the top 8 teams of 2016, as well as WE2015. The round-robin results confirmed that WE2015 would take third place, behind the champion team (Gliders2016) and the runner-up (HELIOS2016). This establishes WE2015 as a stable benchmark for the 2D Simulation League. In doing so we offered a critique of a particular ranking method (the *continuous* scheme), arguing that the *discrete* scheme is more appropriate.

We then followed with proposing two options to develop the evaluation challenge further. The first such possibility introduces "The Champions Simulation League", comprising several previous champions, directly competing against each other in a round-robin tournament. "The Champions Simulation League" can systematically trace the advancements in the League, measuring the progress of each new champion over its predecessors. We evaluated The Champions Simulation League with the champions from 2011 to 2016, producing a ranking which completely concurs with the chronological order, and confirming a steady progress in the League. Arguably, simulation leagues are the only ones in RoboCup where such an evaluation is possible, given the obvious constraints and difficulties with running such a tournament in robotic leagues.

Tracing such advances is especially important because different champion teams usually employ different approaches, often achieving a high degree of specialisation in a sub-field of AI, for example, automated hierarchical planning developed by WrightEagle [23, 24, 26, 21, 28], opponent modelling studied by HELIOS [27], and human-based evolutionary computation adopted by Gliders [11, 12]. Many more research areas are likely to contribute towards improving the League, and several general research directions are recognised as particularly promising: nature-inspired collective intelligence [29–31], embodied intelligence [32–35], information theory of distributed cognitive systems [36–41], guided self-organisation [42–44], and deep learning [45–47].

The other proposed evaluation challenge ("The Global Challenge") aims to model environmental conditions during the games by simulating specific features of different participating countries, such as climate, infrastructure, travel distance, etc. This, arguably, may increase the realism of the simulated competition, making another small step toward the ultimate Millennium challenge.

5 Acknowledgments

A majority of RoboCup 2D Soccer Simulation teams, including the 2016 champion team, Gliders2016, are based on the well-developed code base *agent2d* [48], release of which has greatly benefited the RoboCup 2D Simulation community. Several teams, including WrightEagle and Oxsy, are independent of *agent2d*.

References

1. Burkhard, H.D., Duhaut, D., Fujita, M., Lima, P., Murphy, R., Rojas, R.: The road to RoboCup 2050. IEEE Robotics Automation Magazine **9**(2) (Jun 2002) 31–38
2. Obst, O.: Using model-based diagnosis to build hypotheses about spatial environments: A response to a technical challenge. In Polani, D., Bonarini, A., Browning, B., Yoshida, K., eds.: RoboCup 2003: Robot Soccer World Cup VII. Lecture Notes in Artificial Intelligence. Springer, Berlin, Heidelberg, New York (2004) 518 – 525
3. Stone, P., Kuhlmann, G., Taylor, M.E., Liu, Y.: Keepaway soccer: From machine learning testbed to benchmark. In Noda, I., Jacoff, A., Bredenfeld, A., Takahashi, Y., eds.: RoboCup-2005: Robot Soccer World Cup IX. Volume 4020. Springer Verlag, Berlin (2006) 93–105
4. Obst, O.: Simulation league - league summary. In Kaminka, G.A., Lima, P.U., Rojas, R., eds.: RoboCup 2002: Robot Soccer World Cup VI. Volume 2752 of Lecture Notes in Computer Science., Springer (2002) 443–452
5. Akiyama, H., Dorer, K., Lau, N.: On the progress of soccer simulation leagues. In Bianchi, R.A.C., Akin, H.L., Ramamoorthy, S., Sugiura, K., eds.: RoboCup 2014: Robot World Cup XVIII. Volume 8992 of Lecture Notes in Computer Science., Springer (2014) 599–610
6. MacAlpine, P., Genter, K., Barrett, S., Stone, P.: The robocup 2013 drop-in player challenges: A testbed for ad hoc teamwork. In: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems. AAMAS'14, International Foundation for Autonomous Agents and Multiagent Systems (2014) 1461–1462
7. Kitano, H., Tambe, M., Stone, P., Veloso, M.M., Coradeschi, S., Osawa, E., Matsubara, H., Noda, I., Asada, M.: The RoboCup Synthetic Agent Challenge 97. In: RoboCup-97: Robot Soccer World Cup I, London, UK, Springer (1998) 62–73
8. Noda, I., Stone, P.: The RoboCup Soccer Server and CMUnited Clients: Implemented Infrastructure for MAS Research. Autonomous Agents and Multi-Agent Systems **7**(1–2) (July–September 2003) 101–120
9. Akiyama, H., Nakashima, T., Henrio, J., Henn, T., Tanaka, S., Nakade, T., Fukushima, T.: HELIOS2016: Team Description Paper. In: RoboCup 2016 Symposium and Competitions: Team Description Papers, Leipzig, Germany, July 2016. (2016)
10. Asai, K., Katsumata, Y., Shibayama, T., Nomura, H., Kondo, R., Tanaka, H., Uchinishi, K., Mizumoto, M., Fuzimitsu, T., Sei, M., Tani, Y., Kubo, S., Matsushita, Y.: RoboCup 2016 - 2D Soccer Simulation League Team Description Ri-one (Japan). In: RoboCup 2016 Symposium and Competitions: Team Description Papers, Leipzig, Germany, July 2016. (2016)
11. Prokopenko, M., Wang, P.and Obst, O., Jaurgeui, V.: Gliders2016: Integrating multi-agent approaches to tactical diversity. In: RoboCup 2016 Symposium and Competitions: Team Description Papers, Leipzig, Germany, July 2016. (2016)
12. Prokopenko, M., Wang, P.: Disruptive innovations in RoboCup 2D Soccer Simulation League: from Cyberoos'98 to Gliders2016. In Behnke, S., Lee, D.D., Sariel, S., Sheh, R., eds.: RoboCup 2016: Robot Soccer World Cup XX. Lecture Notes in Artificial Intelligence. Springer, Berlin (2016)

13. Li, P., Ma, X., Jiang, F., Zhang, X., Peng, J.: CSU_Yunlu 2D Soccer Simulation Team Description Paper 2016. In: RoboCup 2016 Symposium and Competitions: Team Description Papers, Leipzig, Germany, July 2016. (2016)
14. Marian, S., Luca, D., Sarac, B., Cotarlea, O.: OXSY 2016 Team Description. In: RoboCup 2016 Symposium and Competitions: Team Description Papers, Leipzig, Germany, July 2016. (2016)
15. Asali, E., Valipour, M., Afshar, A., Asali, O., Katebzadeh, M., Tafazol, S., Moravej, A., Salehi, S., Karami, H., Mohammadi, M.: Shiraz Soccer 2D Simulation Team Description Paper 2016. In: RoboCup 2016 Symposium and Competitions: Team Description Papers, Leipzig, Germany, July 2016. (2016)
16. Zhang, L., Yao, B., Chen, S., Lv, G.: MT2016 Robocup Simulation 2D Team Description. In: RoboCup 2016 Symposium and Competitions: Team Description Papers, Leipzig, Germany, July 2016. (2016)
17. Darijani, A., Mostaejeran, A., Jamali, M.R., Sayareh, A., Salehi, M.J., Barahimi, B.: FURY 2D Simulation Team Description Paper 2016. In: RoboCup 2016 Symposium and Competitions: Team Description Papers, Leipzig, Germany, July 2016. (2016)
18. Budden, D., Wang, P., Obst, O., Prokopenko, M.: Simulation leagues: Analysis of competition formats. In: RoboCup 2014: Robot Soccer World Cup XVIII, Springer (2014)
19. Budden, D.M., Wang, P., Obst, O., Prokopenko, M.: Robocup simulation leagues: Enabling replicable and robust investigation of complex robotic systems. IEEE Robotics and Automation Magazine **22**(3) (2015) 140–146
20. Gabel, T., Falkenberg, E., Godehardt, E.: Progress in RoboCup Revisited: The State of Soccer Simulation 2D. In Behnke, S., Lee, D.D., Sariel, S., Sheh, R., eds.: RoboCup 2016: Robot Soccer World Cup XX. LNAI. Springer, Berlin (2016)
21. Li, X., Chen, R., Chen, X.: WrightEagle 2D Soccer Simulation Team Description 2015. In: RoboCup 2015 Symposium and Competitions: Team Description Papers, Hefei, China, July 2015. (2015)
22. Bai, A., Lu, G., Zhang, H., Chen, X.: WrightEagle 2D Soccer Simulation Team Description 2011. In: RoboCup 2011 Symposium and Competitions: Team Description Papers, Istanbul, Turkey, July 2011. (2011)
23. Bai, A., Chen, X., MacAlpine, P., Urieli, D., Barrett, S., Stone, P.: WrightEagle and UT Austin Villa: RoboCup 2011 Simulation League Champions. In: RoboCup 2011: Robot Soccer World Cup XV. Lecture Notes in Artificial Intelligence. Springer (2012)
24. Zhang, H., Jiang, M., Dai, H., Bai, A., Chen, X.: WrightEagle 2D Soccer Simulation Team Description 2013. In: RoboCup 2013 Symposium and Competitions: Team Description Papers, Eindhoven, The Netherlands, June 2013. (2013)
25. Zhang, H., Chen, X.: The decision-making framework of WrightEagle, the RoboCup 2013 soccer simulation 2D league champion team. In: Robot Soccer World Cup, Springer (2013) 114–124
26. Zhang, H., Lu, G., Chen, R., Li, X., Chen, X.: WrightEagle 2D Soccer Simulation Team Description 2014. In: RoboCup 2014 Symposium and Competitions: Team Description Papers, Joao Pessoa, Brazil, July 2014. (2014)
27. Akiyama, H., Shimora, H., Nakashima, T., Narimoto, Y., Yamashita, K.: HELIOS2012: Team Description Paper. In: RoboCup 2012 Symposium and Competitions: Team Description Papers, Mexico City, Mexico, June 2012. (2012)
28. Bai, A., Wu, F., Chen, X.: Online planning for large Markov decision processes with hierarchical decomposition. ACM Transactions on Intelligent Systems and Technology **6**(4) (July 2015) 45:1–45:28
29. Sayama, H.: Guiding designs of self-organizing swarms: Interactive and automated approaches. In Prokopenko, M., ed.: Guided Self-Organization: Inception. Volume 9 of Emergence, Complexity and Computation. Springer Berlin Heidelberg (2014) 365–387

30. Nallaperuma, S., Wagner, M., Neumann, F.: Analyzing the effects of instance features and algorithm parameters for maxmin ant system and the traveling salesperson problem. Frontiers in Robotics and AI **2** (2015) 18
31. Hamann, H., Khaluf, Y., Botev, J., Divband Soorati, M., Ferrante, E., Kosak, O., Montanier, J.M., Mostaghim, S., Redpath, R., Timmis, J., Veenstra, F., Wahby, M., Zamuda, A.: Hybrid societies: Challenges and perspectives in the design of collective behavior in self-organizing systems. Frontiers in Robotics and AI **3** (2016) 14
32. Pfeifer, R., Bongard, J.C.: How the body shapes the way we think: A new view of intelligence. The MIT Press (November 2006)
33. Polani, D., Sporns, O., Lungarella, M.: How information and embodiment shape intelligent information processing. In Lungarella, M., Iida, F., Bongard, J., Pfeifer, R., eds.: Proceedings of the 50th Anniversary Summit of Artificial Intelligence, New York. Volume 4850 of Lecture Notes in Computer Science., Berlin / Heidelberg, Springer (2007) 99–111
34. Der, R.: On the role of embodiment for Self-Organizing robots: Behavior as broken symmetry. In Prokopenko, M., ed.: Guided Self-Organization: Inception. Volume 9 of Emergence, Complexity and Computation. Springer Berlin Heidelberg (2014) 193–221
35. Ghazi-Zahedi, K., Haeufle, D.F.B., Montfar, G., Schmitt, S., Ay, N.: Evaluating morphological computation in muscle and dc-motor driven models of hopping movements. Frontiers in Robotics and AI **3** (2016) 42
36. Ay, N., Bertschinger, N., Der, R., Guttler, F., Olbrich, E.: Predictive information and explorative behavior of autonomous robots. The European Physical Journal B - Condensed Matter **63** (2008) 329–339(11)
37. Tishby, N., Polani, D.: Information theory of decisions and actions. In Cutsuridis, V., Hussain, A., Taylor, J.G., eds.: Perception-Action Cycle: Models, Architectures, and Hardware. Springer New York, New York, NY (2011) 601–636
38. Cliff, O.M., Lizier, J., Wang, R., Wang, P., Obst, O., Prokopenko, M.: Towards quantifying interaction networks in a football match. In Behnke, S., Veloso, M., Visser, A., Xiong, R., eds.: RoboCup 2013: Robot Soccer World Cup XVII, Springer (2013) 1–12
39. Lizier, J.T., Prokopenko, M., Zomaya, A.Y.: A framework for the local information dynamics of distributed computation in complex systems. In Prokopenko, M., ed.: Guided Self-Organization: Inception. Volume 9 of Emergence, Complexity and Computation. Springer Berlin Heidelberg (2014) 115–158
40. Cliff, O.M., Prokopenko, M., Fitch, R.: An information criterion for inferring coupling of distributed dynamical systems. Frontiers in Robotics and AI **3** (2016) 71
41. Cliff, O.M., Lizier, J.T., Wang, P., Wang, X.R., Obst, O., Prokopenko, M.: Quantifying long-range interactions and coherent structure in multi-agent dynamics. Artificial Life **23**(1) (2017) 34–57
42. Prokopenko, M.: Guided self-organization. HFSP Journal **3**(5) (2009) 287–289
43. Der, R., Martius, G.: The Playful Machine – Theoretical Foundation and Practical Realization of Self-Organizing Robots. Springer (2012)
44. Prokopenko, M.: Guided Self-Organization: Inception. Springer, Berlin (2014)
45. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8) (August 2013) 1798–1828
46. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural Networks **61** (2015) 85–117
47. Greenwald, H.S., Oertel, C.K.: Future directions in machine learning. Frontiers in Robotics and AI **3** (2017) 79
48. Akiyama, H.: Agent2D Base Code. `http://www.rctools.sourceforge.jp` (2010)