

On Progress in RoboCup: The Simulation League Showcase

Thomas Gabel and Martin Riedmiller

Machine Learning Laboratory
University of Freiburg, 79110 Freiburg im Breisgau, Germany
{tgabel|riedmiller}@informatik.uni-freiburg.de

Abstract. The annual RoboCup competitions are certainly relevant to present times as they elect the best team of the current year. With respect to RoboCup’s well-known 2050 vision, however, it is crucial to assess the general progress being made not just qualitatively, but also in a quantitative manner. Although this is difficult to accomplish in most leagues, the recent development and circumstances in the soccer simulation league led to a unique situation which allowed us to perform an extensive experimental evaluation by means of which we can empirically measure the progress of playing performance made over a number of successive years. The findings we present in this paper, in particular the fact that significant improvements in the level of play can be verified quantitatively in the 2D soccer simulation league, may serve as a showcase for the progress made by the RoboCup initiative in general.

1 Introduction

The 2D soccer simulation league might be called the grandmother of all RoboCup leagues. Its origins date back to Itsuki Noda’s first implementations of the Soccer Server, a two-dimensional soccer simulation software in 1994 [1, 2] as well as to the Pre-RoboCup event at IROS 1996 in Osaka where simulated soccer agents competed with one another within official competitions for the first time. Ever since the 2D soccer simulation league has seen 14 successive versions of its simulation software and has competed in 13 world championship tournaments as well as in more than 50 major regional competitions. With such a long history within the RoboCup initiative, a natural question to ask is what progress has been made throughout the years.

This is the topic we want to focus on in this paper. In particular, we want to look back and investigate and reflect on the following questions:

- What are the exact conditions that are required to allow for a *meaningful* retrospection?
- If such conditions can be identified, has there been any *provable* progress in performance in soccer simulation 2D during the last few years?
- If so, how close are teams and how reliable are the results of a single game with respect to the noise in the simulation? Has there been something like

convergence in the teams' further development and has perhaps a certain saturation level in their performance been reached?

These questions are extremely hard to answer in other, hardware-based leagues. New technological solutions, new mechanical designs combined with strong rule changes introduced year by year are certainly necessary on the road towards RoboCup's 2050 vision, but they make it hard, if not impossible, to perform a meaningful comparison between the achievements and performance of RoboCup teams over the years. Indeed, sometimes it may have happened that the wheel has been invented twice, whereas at other instances former RoboCup team members minify or disdain contemporary approaches pointing out that certain problems had already been solved in their times. Hence, subjective assessments often superimpose and conceal the true progress.

By contrast, a stable platform allows for a reliable and meaningful evaluation of the true progress made. RoboCup in general, however, is known to be a highly dynamic and changing domain. The development in the soccer simulation league during the previous years, which we will summarize in more detail in Section 2, has created circumstances that must be described as unique in the realm of RoboCup. For a period of five consecutive years, the software platform used in the soccer simulation 2D league remained unchanged. As a consequence, it became possible for the first time to thoroughly and critically evaluate the development and progress of an entire RoboCup league over a period of five successive years. We present the results of a comprehensive analysis concerning these issues in Section 3. Moreover, given the fact that the competitive character of RoboCup competitions is nearly identical in all leagues, we may infer that the progress to be observed in the simulation league can also serve as a showcase for other RoboCup leagues even if their characteristics disallow for a quantitative comparison. Apart from these considerations, we also look at the current character of the competitions in the 2D league and analyze whether saturation in the level of performance has been reached by the teams (Section 4).

2 Evolution of Soccer Simulation

In its early years, the simulation software of the soccer simulation league (Soccer Server) underwent significant changes and extensions. The simulation grew more complex from year to year, e.g. tacklings were introduced, heterogeneous player types were created, players were equipped with a simulated arm and neck, to name just a few of the enhancements (see Figure 1 for an overview).

On the one hand, these year-to-year changes were intended to make the simulation more complex and, eventually, more realistic. Moreover, yearly revisions kept up the pressure on all participating teams, so as to keep pace with the changes introduced. On the other hand, any modification to the simulation raised and still raises the question of backward compatibility. Will team binaries from a previous year still be usable for the current or yet-to-come Soccer Server version? Clearly, when certain features are removed from the simulation from one year to the next, then it is likely that older teams will malfunction given

that they relied on the feature removed. If existing parts of the simulation are altered, the impact on older team binaries will be less drastic, although they are in general expected to be impaired and play worse than they did with the previous Soccer Server version. Finally, the addition of entirely new features to the simulation is least critical. However, assuming that the use of a newly available feature (e.g. the use of heterogeneous players, introduced in 2001) can be beneficial with respect to team performance, a team making use of the new feature is supposed to outperform an older binary not aware of the innovation. In essence, by having annually modified the simulation environment, a meaningful comparison across teams from different years was impossible.

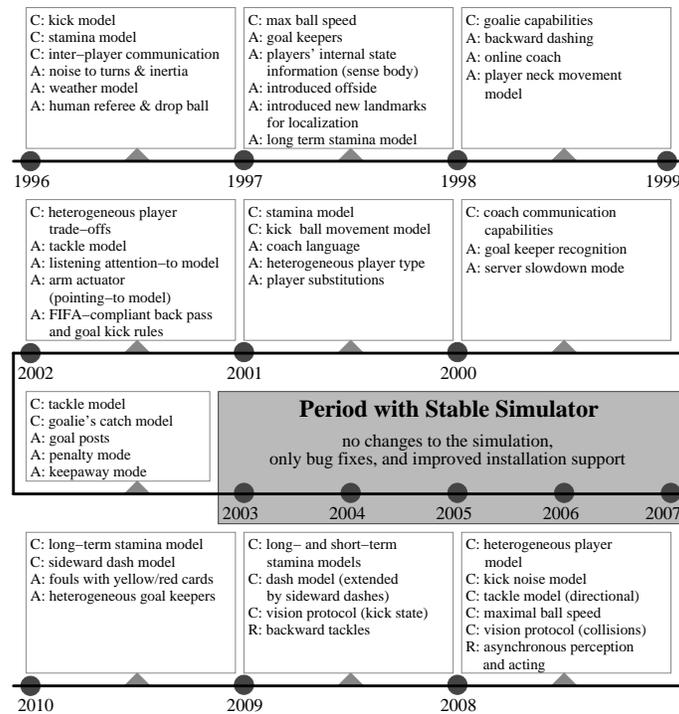


Fig. 1. Evolution of the Soccer Server: The 2D soccer simulation software has undergone various revisions over the years (**C**hanges, **A**dditions, and **R**emovals of features).

Besides the issues of simulator backward compatibility and cross-year comparisons of soccer simulation teams, the question on the long-term perspective of the simulation league was raised during the RoboCup Symposium 2002. To this end, it was envisioned to extend the existing 2D simulation towards a 3D scenario as well as to more realistic robot simulations. After RoboCup 2002, including the mentioned symposium milestone discussion [3], the proposal came up to freeze the 2D soccer simulator and to retain it as a testbed for multi-agent research.

In particular, the goal was to keep the simulated soccer environment stable in order to facilitate measuring scientific progress from year to year. Moreover, the idea was to start developing a new 3D simulator from scratch that should be based on correct physical modelling in a three-dimensional environment.

With almost eight years having passed since then, we can view the following historic developments in retrospect:

- First work on a 3D simulator began in August 2002 (as mailing list archives confirm¹) and the first RoboCup tournament of the newly founded 3D-subleague took place in 2004.
- The transition towards simulating humanoid robots was accomplished in 2007, though this meant abandoning playing 11 versus 11. In fact, as of now there exist multiple 3D robot simulators in parallel [4].
- The 2D simulator remained stable from 2003 onward until the beginning of 2008. Although the number of teams allowed to participate in the annual 2D competitions at RoboCup was strongly restricted, the interest in 2D soccer simulation did not decline.

In 2008, the freezing policy was finally abandoned and, once again, significant changes were introduced into the 2D simulation in that year. There were two main reasons for this change of policy:

- The 3D league had started off adopting agents in the form of spheres allowing for access to abstract commands like kicking, turning, or dashing, not unsimilar to those known from the 2D simulation. With the above-mentioned transition to modelling humanoid robot models in 2007, the 3D league entered a lower level where the simulated robots had to be programmed in terms of controlling motor torques. As a consequence, the performance level of the games dropped significantly [4] as well as the number of players per team, that could be simulated simultaneously, did decline rapidly (2 vs. 2). Thus, at this point, the 2D simulation again represented the only branch of RoboCup focussing on issues like multi-agent coordination, team-play, and strategic reasoning.
- Given the special role resulting from the previous point, there was a growing demand within the simulation community to further develop the 2D simulator. Among other things, it was aimed at making the 2D simulation resemble real soccer, at least from a bird’s eye perspective, as much as possible.
- As pointed out, the soccer simulation 2D league had experienced a period of stableness lasting five consecutive years. However, contrary to the original intention, year-to-year progress was not measured in the time window in which the 2D soccer simulation represented a stable multi-agent testbed.

With the paper at hand, we aim at filling the gap mentioned last. Since the policy of keeping a stable 2D simulator was irreversibly revoked in 2008, we consider it important to grasp the opportunity of performing a meaningful evaluation (at least covering the stable period visualized in Figure 1) and prove the progress of this league quantitatively, not just qualitatively.

¹ http://sourceforge.net/mailarchive/forum.php?forum_name=sserver-three-d

3 On Progress and Continuity

In human soccer playing, it is impossible to make sound statements about the development of the playing performance level over years. Although many experts will certainly agree that contemporary soccer exhibits more speed, dynamism, and superior athleticism than a few decades ago, empirical proof on its superior performance is inconceivable.

3.1 RoboCup Leagues' Further Development

With respect to long-term comparisons, one might argue that the situation is better in robot soccer. Robots are patient, can be stored away after a competition, and be unpacked after a few years in order to compete against a contemporary team. This is a nice thought experiment, but experience in RoboCup has proven that such an approach is infeasible. Hence, a formal analysis of the progress of soccer-playing robots made throughout the years is difficult to establish. Nevertheless, most researchers who have been involved in RoboCup activities over several years will unanimously agree that the progress made in all hardware-oriented RoboCup soccer leagues is not to be questioned. Despite this, assessments concerning the exact year-to-year progress remain qualitative in nature.

RoboCup's simulation leagues adopt a special role. No hardware development and maintenance is required and software agents do not age. However, the performance of soccer-playing software agents strongly relies on the simulation of the physical world they are placed in. If the characteristics of that simulation change, a meaningful evaluation of the progress made is rendered impossible. As delineated (Section 2), the simulation league has experienced a highly dynamic history, where the simulators used have undergone tremendous changes over the years. So, statements about the general level of playing performance frequently remained vague and on a qualitative level, without empirical verification. For example, already in 2002 it was claimed that the "overall playing strength of the teams in the tournament was quite impressive" [3]. Moreover, "the playing level of the tournament showed increased and consistent improvement as compared to last year's tournament". Similarly, "the 2003 tournament again showed a big advance in the performance of the teams" [5] and the "level of play of the last twelve teams this year was very mature and close to each other" [6].

While we subscribe to these assessments, we need to stress that no empirical proof for their correctness exists and, in fact, cannot exist. As shown in Figure 1, from 2001 to 2002 and 2002 to 2003 several changes were introduced into the 2D soccer simulation, which is why a meaningful evaluation of the objective progress made in those days is infeasible. By contrast, our focus is on the subsequent period of stability (2003-2007). The results of a large-scale analysis of the progress made during that time interval shall be presented in the next section.

3.2 Empirical Proof of Progress

Platform As pointed out, of crucial importance to a meaningful comparison is the use of a simulation platform that is equally supported by all teams un-

der consideration. During the stable period sketched in Figure 1, four different versions of the Soccer Server were used for the RoboCup world championship tournaments in the respective years. Though differing in their version numbers and deviating slightly in platform support and ease of installation, the simulation is entirely identical and allows all team binaries published during that time to work properly with any of the Soccer Server versions released in those years. Therefore, we conducted all empirical analyses using the 2007 version 11.1.0 of the Soccer Server, as this version is fully downward compatible and represents the one that is expected to have applied all relevant bug fixes.

Team Selection Our aim has been to assess the progress made over the years in soccer simulation 2D. We are convinced that a year’s level of play can be validly read from the performance of that season’s top teams, i.e. the teams placing best during a year’s RoboCup world championships. We refrain, however, from saying that a year’s champion alone is a reliable representative of its volume’s strength. For these reasons, we decided to have each of the years from 2003 to 2007 be represented by all teams that made it into the respective year’s semi-finals (top four teams). In reporting results, we explicitly do not provide the exact team names, but use identifiers of the form “year_place” (e.g. 2007_2 to indicate the runner-up of RoboCup 2007) for the sake of better interpretability. A matching from these identifiers to team names is given in the Appendix.

Experiments We retrieved the original top four binaries from the five consecutive years under consideration and prepared them for use in the scope of an extensive evaluation involving more than 4000 regular matches. In order to ensure identical prerequisites, we conducted all matches consecutively on exactly the same hardware and used Soccer Server version 11.1.0. For the league progress analysis, we let each team face each other team fifteen times in a standard match of 6000 simulation cycles length.

Results In the top right part of Figure 2, we can see how well each team performed in this comprehensive evaluation. The bar chart shows the share of points each team carved out, when repeatedly facing any of the other 19 teams involved in this study. As usual, a team is awarded 3 points for a victory, 1 for a draw, and 0 for a defeat, thus 100% would indicate winning each match, whereas 33.3% would be equivalent to drawing each match. The 2007 champion obviously comes off best in this regard with a share of 86.0%. Moreover, we can observe that older teams (darker shades of gray) are placed in the second half of this score board-like visualization.

The main part of Figure 2 focuses on the scores (each data point is averaged over 270 matches) the teams yielded when repeatedly playing against all other teams considered in this evaluation. By connecting the data points belonging to representatives of the same year, the steady increase in performance becomes clear. Apparently, the polygons formed are shifted from the chart’s top left to its bottom right region with increasing years (where the average number of goals scored is larger than the number of goals received).

While the previous analysis has focused on individual teams, Figure 3 examines the performance level of different years against one another. Here, we let

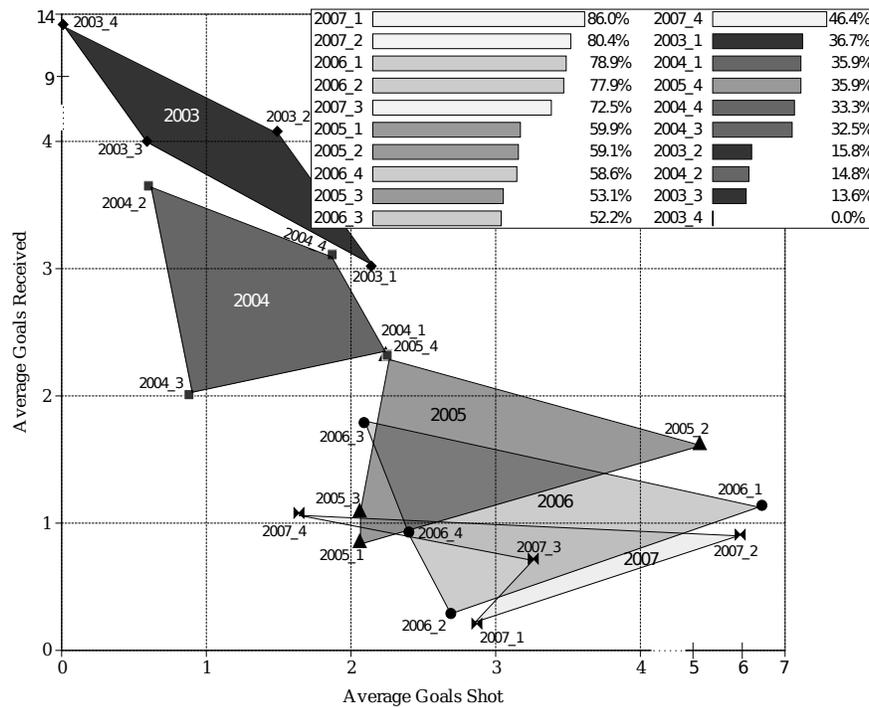


Fig. 2. Top right: Relative strengths of the 20 teams considered in this evaluative study. Main: Average scores yielded by each of these representatives, when repeatedly playing against one another.

the representatives from each season play 240 matches each against the representatives from any other year. The matrix-like result presentation in that figure highlights the fact that in any of the 10 combinations, the season listed in the row was inferior to the column season. Generally, when comparing year x and year y (with x, y within $\{2003, \dots, 2007\}$), it holds that the difference in the number of points achieved against one another as well as the difference in the average scores becomes more pronounced the larger $|x - y|$.

In addition to the year to year-based comparisons, Figure 4 summarizes the progress analysis of the stable period by visualizing the joint performance of representatives from the same year playing against teams from all other seasons. As a matter of fact, the share of matches won is increasing continuously over the years, while the share of games lost is dropping at the same time. For example, the 2007 representatives lost only 12.9% of their matches against binaries from all recent years. The chart to the right in Figure 4 shows that a similar trend holds for the average scores resulting from this study and, thus, provides further empirical proof for the verifiable progress of soccer simulation 2D made from 2003 to 2007.

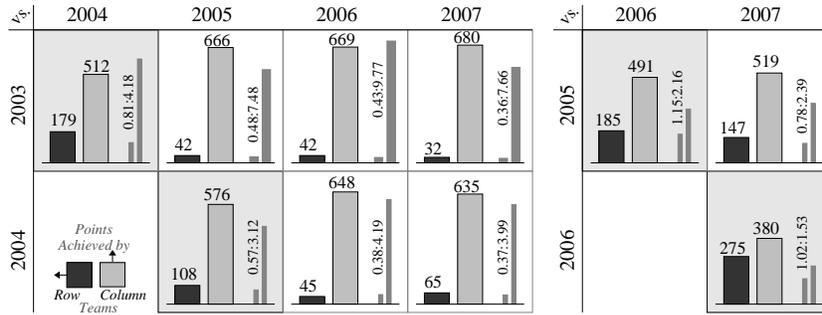


Fig. 3. Representatives from each year of the stable period (2003-2007) played 240 matches against representatives from any other year. This figure summarizes the numbers of points (maximum: 720) as well as average scores yielded in this experiment.

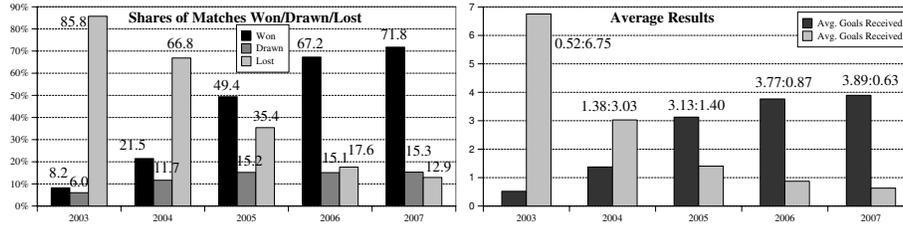


Fig. 4. Joint performance of representatives from the same year when playing against teams from all other seasons: Both average results (over 2400 matches) as well as the development of the shares of matches won/drawn/lost hint to the fact that substantial progress in playing performance had been made.

4 On the Character of 2D Competitions

With the development delineated, the question arises whether we can observe some kind of saturation in the further development of the 2D league teams' performance and whether we can see the top teams playing at approximately the same level.

A first answer to this question can be read from the charts in Figure 3. Teams from a successor season do always have a clear edge over the preceding year's teams. While 2004 teams would lose 20.8% of the matches against 2003 representatives and teams from 2005 only 14.7% against 2004 representatives, this number increased slightly to 19.6% as well as 29.2% in 2006 and 2007, respectively. The progress made from year to year, however, has slowed down recently as can be read from the average scores (e.g. 1.02:1.53 for 2006 vs. 2007 teams). A second answer to this question relates once again to the issues mentioned in Section 2. With the re-initiated introduction of changes and extensions to the simulator in 2008, any form of convergence became basically infeasible, since the modifications of the simulation forced the teams to reimplement considerable

parts of their teams or to add functionality that may turn out to be crucial for becoming or staying competitive. There are, however, two further answers to the above question, both of which we will discuss and underpin empirically in the following sections.

4.1 Randomness and Repetability

The question on convergence of performance is closely related to a question that is frequently brought up by spectators during competitions, especially, when it comes to the final matches, “If you replayed this match, would the result be the same?” Of course, the general answer is “no”. But, if we could observe something like convergence in the performance development of the teams, then this “no” should become evidently more pronounced over the years.

Experiments In order to investigate this question, we focused on the final matches of the seven recent RoboCup tournaments (2003-2009). Under the same conditions as the experiments in Section 3, we let the final matches of those years be replayed for 100 times each, calculating average results as well as corresponding standard deviations (outcomes are plotted in Figure 5).

Results This series of 700 matches reveals two interesting facts:

- An apparent (and appreciative) observation is that in any year (since 2003) the 2D simulation league found a worthy champion. The analysis shows that, from the perspective of the winning team, the average score over many repetitions of the final match is $x : y$ with $x > y$ for all years (in fact, $\min_{2003, \dots, 2009} x - y = 0.31$). Also, it never occurred that the probability for the runner-up to win a match against the champion was higher than the probability for the champion to win. To this end, the closest situation was in 2006, where the runner-up would have even won one out of three matches.
- There is no clear trend of converging performance (i.e. with more and more equally well-performing final match participants), not even during the stable period. At best, one may infer that in 2003 and 2004 the domination of the champion team used to be more pronounced than in later years and that the total number of goals scored in the final had been dropping until 2008. Thus, the development observed does not allow us to conclude that a state of saturation had or has been reached.

4.2 The Effect of “Minor” Changes: A RoboCup 2009 Example

Another indicator for saturation of progress is that significant improvements to a team’s chances of winning the next game can be realized only by investing an exponential amount of effort and would, thus, be impossible to accomplish during competitions. In RoboCup 2009, our team (Brainstormers) experienced a textbook counter example which is unrelated to any recent simulator changes and, hence, descriptive for the state of the league.

Initial Situation: Our team’s defense strategy strongly relies on an effective man-to-man marking. The exact matching from our players to their direct, i.e. to

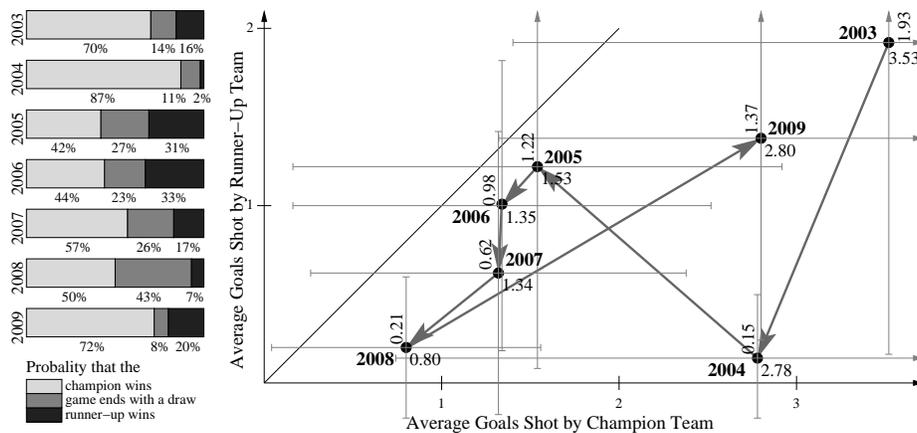


Fig. 5. Replaying the final matches from 2003-2009: Although no clear trend is manifested, in each year the superior team has won the World Champion title.

be marked, opponents are computed by the coach program based on an opponent player role determination, using gliding position windows with exponential weighting as well as a greedy graph-matching algorithm. Having determined an appropriate man-to-man marking, the coach communicates the assignment of direct opponents to the players on demand.

Having a look at the distribution of results that is created when the assignment mechanism works properly, we can see (Figure 6, left column) that our team would win/draw/lose 30/30/40% of the matches against the 2009 champion WrightEagle and 32/29/39% against Oxy (3rd place). Taking into account the average results and their respective standard deviations plotted in that figure, it is obvious that the starting situation could be described as head to head.

Preparation for RoboCup 2009: WrightEagle decided to develop a player role-changing mechanism that turned out to make the described man-to-man marking get out of step. Although these role changes were detrimental to WrightEagle's own performance (e.g. because a defender now had to act as an attacker) and they increased the number of goals received from 1.01 ± 0.975 to 1.45 ± 2.06 , the damage to the Brainstormers' defense strategy was more pronounced. The number of goals shot by WrightEagle more than doubled as can be read from Figure 6 (top row, middle column). Consequently, when both teams faced each other at the semi-final the match ended 0:4.

On-Site Modifications: After having observed and analyzed the above-mentioned semi-final, team Oxy copied WrightEagle's role-changing strategy in preparation for the matches on the following day. Simultaneously, our team took counter measures and improved the man-to-man marking determination. The effect of Oxy's overnight clone of WrightEagle's role-changing strategy can be read from Figure 6 (bottom row, middle column). Apparently, applying the mentioned strategy blended very well with Oxy's general way of playing, which is

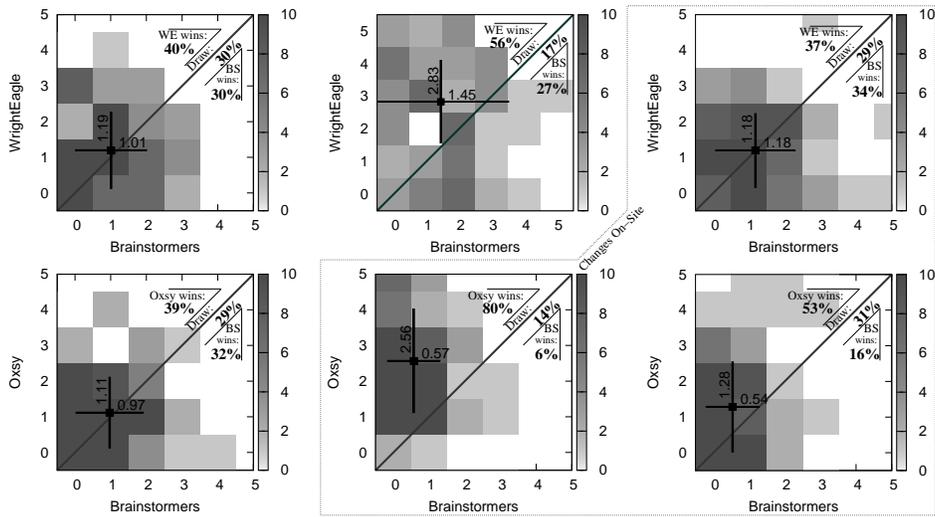


Fig. 6. Average goals shot and received with corresponding standard deviations as well as result distributions of matches between the Brainstormers and WrightEagle (top) and Oxy (bottom) during RoboCup 2009. The initial situation (left column) is contrasted with the impact of a single winning feature added by WrightEagle and Oxy (middle column) and counter measures taken by the Brainstormers (right column), with the latter realized during the competition.

why the impact on the average result yielded against the Brainstormers was substantial (improvement from 0.97:1.11 to 0.57:2.56 on average). By contrast, the impact of our team’s counter measures (also implemented overnight), which happened to play no further role during the rest of the tournament, are visualized in the right column of Figure 6. Apparently, the impact of the role-changing strategy could be fully annihilated when playing against WrightEagle, and partially when playing against Oxy.

While the example outlined in this section represents only a snapshot from last year’s RoboCup tournament, it serves as evidence to support the claim that decisive and potentially winning modifications to a 2D soccer simulation team’s strategy can still be achieved in little time and, in particular, even on-site. This fact is not only a counter example against the saturation hypothesis, but it also stresses the exciting character of the competitions.

5 Conclusion

In this paper, we have targeted questions that relate to the progress made in RoboCup across years. While qualitative assessments of the further development exist in abundance for many leagues, in general it is impossible to make distinctive and verifiable measurements of the playing performance between participants of different years. Starting with the observation that recent circumstances

in the soccer simulation 2D league allowed us to perform a meaningful evaluation, we presented the results of an extensive study that, in a retrospective manner, investigated the further development of the playing strengt in soccer simulation.

We found that the progress made during the so-called “stable period” (2003-2007) is astonishing. While admired for their sophisticated play in 2003 and 2004 [5], world champions of those times played at the level of a low-class team a few years later, sometimes losing in the double digits against top teams from 2006 or 2007. In an additional analysis, we found that up to now no saturation has been reached in the level of play of the teams competing in the soccer simulation 2D league. In particular, we found that even minor changes or extensions to a team’s repertoire of capabilities, i.e. modifications that can be realized even on-site during competitions, may suffice to bring about significant and game-winning advantages. Needless to say this feature strongly contributes to the exciting character of the competitions in this league.

Acknowledgements The authors would like to thank Andreas Hechenblaickner for providing the HLM League Manager sources as well as Hidehisa Akiyama, Ke Shi, and C. Fan for their help in retrieving and refloating old team binaries.

Appendix

	1	2	3	4
2003	UvATrilearn (NED)	TsinghuAeolus (CHN)	Brainstormers (GER)	Everest (CHN)
2004	STEP (RUS)	Brainstormers (GER)	Mersad (IRN)	TsinghuAeolus (CHN)
2005	Brainstormers (GER)	WrightEagle (CHN)	TokyoTech (JPN)	STEP (RUS)
2006	WrightEagle (CHN)	Brainstormers (GER)	Ri-One (JPN)	TokyoTech (JPN)
2007	Brainstormers (GER)	WrightEagle (CHN)	HELIOS (JPN)	OPU-Hana (JPN)
2008	Brainstormers (GER)	WrightEagle (CHN)	HELIOS (JPN)	AmoyNQ (CHN)
2009	WrightEagle (CHN)	HELIOS (JPN)	Oxxy (ROM)	Brainstormers (GER)

References

1. Noda, I.: Soccer Server: A Simulator of RoboCup. In: Proceedings of the AI Symposium 1995, Japanese Society for Artificial Intelligence (1995) 29–34
2. Noda, I., Matsubara, H., Hiraki, K., Frank, I.: Soccer Server: A Tool for Research on Multi-Agent Systems. *Applied Artificial Intelligence* **12** (1998) 233–250
3. Obst, O.: Simulation League – League Summary. In: G. Kaminka, P. Lima, R. Rojas (Eds.): *RoboCup 2002: Robot Soccer World Cup VI*. Springer, Melbourne, Australia (2003) 443–452
4. Kalyanakrishnan, S., Hester, T., Quinlan, M., Bentor, Y., Stone, P.: Three Humanoid Soccer Platforms: Comparison and Synthesis. In: Proceedings of the RoboCup International Symposium 2009. Springer, Graz, Austria (2009)
5. Pagello, E., Menegatti, E., Bredendfeld, A., Costa, P., Christaller, T., Jacoff, A., Polani, D., Riedmiller, M., Safiotti, A., Sklar, E., Tomoichi, T.: RoboCup-2003. *New Scientific and Technical Advances. AI Magazine* **25** (2004) 81–98
6. Pagello, E., Menegatti, E., Bredendfeld, A., Costa, P., Christaller, T., Jacoff, A., Johnson, J., Riedmiller, M., Safiotti, A., Tomoichi, T.: Overview of RoboCup 2003 Competition and Conferences. In: D. Polani et al. (Eds.): *RoboCup 2003: Robot Soccer World Cup VII*. Springer, Padova, Italy (2004) 1–14