A Statistical Perspective on the RoboCup Simulator League: Progress and Prospects

Kumiko TANAKA-Ishii, Ian FRANK
, Itsuki NODA, and Hitoshi MATSUBARA

Electrotechnical Laboratory
Umezono 1-1-4, Tsukuba
Ibaraki, JAPAN 305
{kumiko,ianf,noda,matsubar}@etl.go.jp

Abstract. This paper uses statistical analysis to demonstrate the progress to date in answering the RoboCup Synthetic Agent Challenge. We analyze the complete set of log data produced by the simulator tournaments of 1997 and 1998, applying techniques such as principal component analysis to identify precisely what has improved, and what requires further work. Since the code that implements our analysis produces its results in real-time, we propose releasing a proxy server that makes statistical analysis available to RoboCup developers. We believe such a server has a crucial role to play in facilitating and evaluating research on the three specific challenge problems of opponent modeling, teamwork and learning. We also suggest that — if RoboCup is to make the most of the efforts of participating researchers — the time is ripe for the institution of a modular team based on a common model.

1 Introduction

In the year between the first and second Robotic Soccer World Cups (RoboCup-97 and RoboCup-98), the level of play of the teams has unquestionably improved. This can be seen by the way that the best teams from 1998 beat the best teams from 1997. However, this kind of high-level observation does not answer the question of *why* the teams perform better. Our principal contribution in this paper is to qualitatively demonstrate the improvement in RoboCup soccer abilities by showing how a team's play can be analyzed with statistics. We analyze the entire set of log data produced by the 1997 and 1998 simulator tournaments in terms of 32 skill-based features.

The RoboCup Synthetic Agent Challenge [Kitano *et al.*1997] defines three separate challenges of *opponent modeling*, *teamwork* and *learning*. Our research in particular demonstrates that significant progress has been made towards meeting the teamwork challenge. For instance, our analysis of player correlations shows that forwards (and defenders) exhibit far greater teamwork in 1998. Further, our analysis techniques are also important for directing *future* research on the Synthetic Agent challenges. Most obviously, by highlighting both the aspects of play that have improved and also those that have not, our statistics enable

researchers to identify the most promising directions for further investigation. More importantly, however, the code that implements our analysis techniques produces its results in real-time, as games are running. We therefore propose offering this code to the RoboCup community in the form of a proxy server. We describe how such a proxy server can directly facilitate research on all three challenge problems, as well as on systems such as automated commentators and post-game analyzers. Further, we note that this kind of server has an obvious role in conducting the evaluation procedures set out in the challenge paper itself. This discussion leads us to suggest a new approach to RoboCup that would enable more effective evaluation and combination of the efforts of RoboCup researchers: the creation of a modular team based on a common model.

As far as we are aware, there have been few previous attempts to address the statistical analysis of soccer in an academic context. [Tanaka-Ishii *et al.*1998] have defined notions such as *pass bigrams* and suggested the use of Voronoi diagrams to measure the areas of influence of players. Also, [Takahashi *et al.*1998] and [Matsubara *et al.*1998] have carried out limited log analyses of the 1997 RoboCup and the 1998 Japan Open, respectively. To date, though, the only significant progress in the statistical analysis of soccer has been the coaching tools and analysis software produced by commercial companies (*e.g.*, see [SoftSport1998]). National teams are assumed to keep databases of statistics, but the details of such resources are a closely guarded secret.

2 The Data

Our analysis is based on all the log files of the first two RoboCups (55 games from 1997 and 81 games from 1998), and focuses on the differences between the winning and losing teams. Since team designers are allowed to manually change their team settings and code at half-time, we collect separately the statistics for the winners and losers of each *half-game* (excluding the 36 half-games that were drawn). The size of the data files for the 236 half-games that produced a winner is around 245Mbyte.

In interpreting this data, it is important to bear in mind that the simulator league also changed between the RoboCups of 1997 and 1998. Most notably:

- A goalkeeper (with a `catch` ability) was introduced.
- The stamina model was changed to make long-term resource management more important.
- An off-side rule was introduced.
- A limit was imposed on the maximum ball speed.

It is natural that client team programs change when the rules change. This is reflected in the statistics we present in the following section.

3 The Statistics

We analyzed the log files in terms of a set of thirty-two features that describe different aspects of a soccer game. These features are shown in Figure 1, along

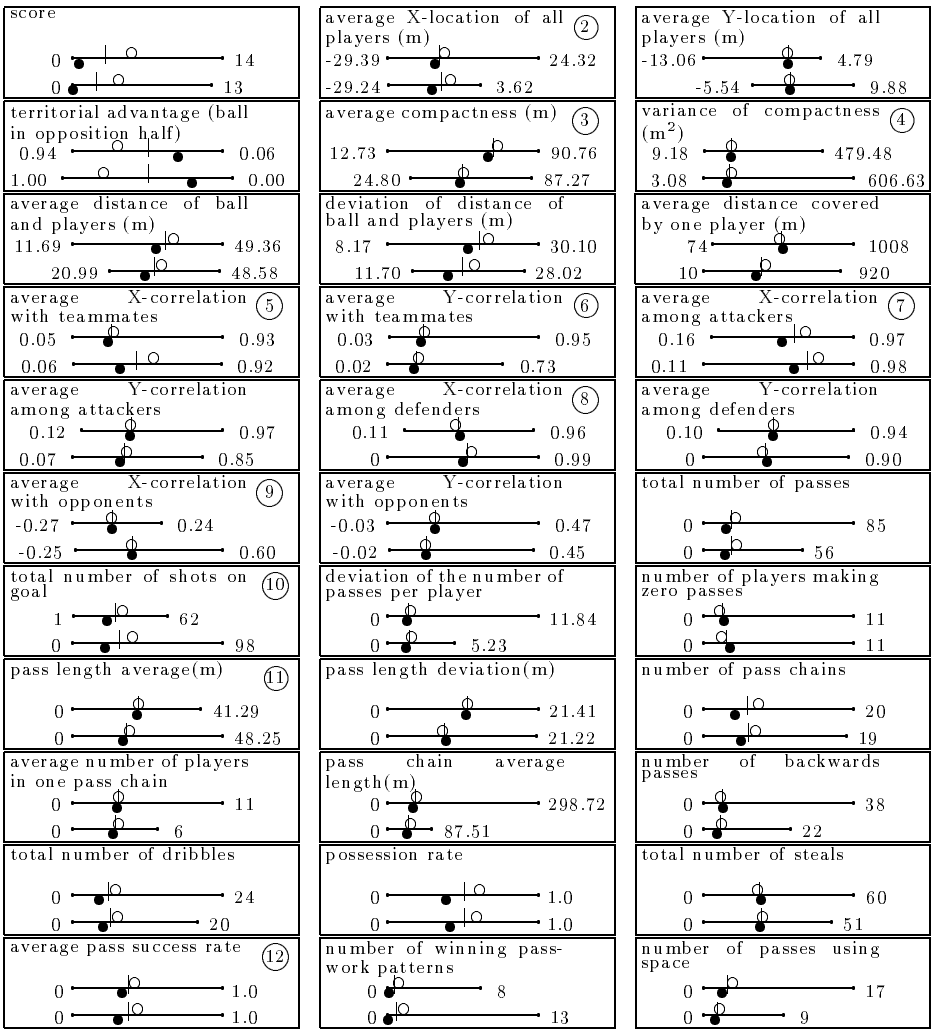

Fig. 1. The 32 evaluation features (plus 'score') that we used to analyze the RoboCup tournaments. The scales show the range of the values taken by these features for one team in one half game in 1997 (upper scale) and in 1998 (lower scale). The vertical bars denote the overall averages, and the ∘ and • marks denote the averages for the winning and losing teams. Circled numbers refer to the figure numbers of the graphs in the text.

with some simple scales giving information on the values they took in 1997 (upper line) and in 1998 (lower line). (Note that X values are measured along the longer side of the field and Y values along the shorter side, with zero in both cases being the center spot.) Broadly, the features fall into three categories that we will discuss in detail below: 8 that represent general aspects of a team's

formation (§3.1), 8 that represent formation in terms of the specific notion of player correlations (§3.2), and 16 that deal with passing and shooting skills (§3.3).

We illustrate our discussion with 11 graphs that detail the frequency distribution of selected features. In each of these graphs, the x-axis is discretized into 30 equal and contiguous intervals that span the observed maximum and minimum values of the feature in question. Considering the losers and winners in 1997 and 1998 separately, the percentage of games for which the value of the feature falls within each interval is then counted, and a smoothed curve fitted through the 30 points.

We concentrate throughout this section on highlighting the overall changes in playing skills from 1997 to 1998. We should point out, though, that the real-time nature of our code (discussed in §4) also supports the investigation of other aspects, such as the way that individual statistics change dynamically as a game develops.

3.1 Formation

Figure 2 shows the average of all the players' X-positions. Players in the opposition's half have a positive X-value, so the larger the values in the figure, the more attacking the formation. Both in 1997 and 1998, the winners were more attacking than the losers (the curves are further to the right).

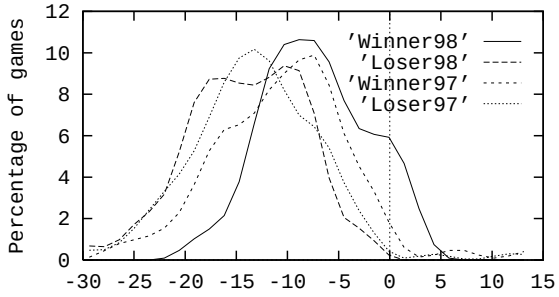

Fig. 2. The average X-location of all players (m)

In real soccer, the entire formation of a team follows the movement of the ball across the field. We define the *compactness* of a team as the X-distance between its front-most player and its rear-most player (excluding the goalkeeper). Figure 3 shows the distribution of average compactness in RoboCup, demonstrating clearly that teams in 1998 were more compact than in 1997. This improvement is probably a result of the introduction of the off-side rule. The presence of off-sides allows a team to protect itself against counter-attack if they move up-field

as a unit with the ball. This becomes an aspect of the game that teams cannot afford to neglect.

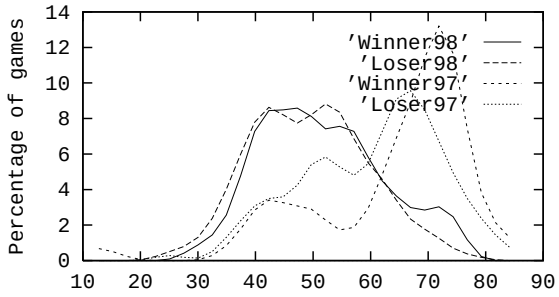

Fig. 3. The average compactness (m)

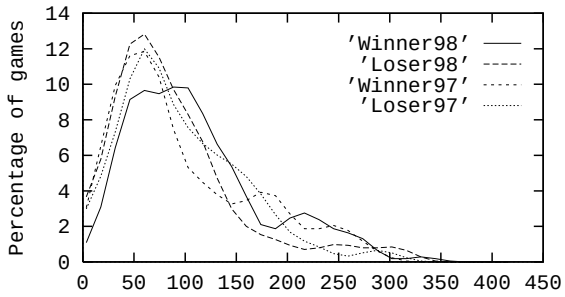

Fig. 4. The variance of compactness (m^2)

Figure 4 shows the variance of compactness during games. This does not show a remarkable difference between winners and losers or over the two years. So, although team formation has improved to dynamically follow the ball, the degree of compactness does not yet change significantly during the game. That the players follow the ball more closely in '98 can also be seen from Figure 1: the average distance between the ball and players decreases in '98. At the same time, though, the average distance covered by a single player has also decreased, indicating that teams have improved their formations whilst also handling their resources better.

3.2 Player Correlations

Another way to analyze players' movements is in terms of correlations between player locations. We calculate these correlations in the form of two 22×22 symmetric matrices, one for the correlation of players' X-locations and the other for

the players' Y-locations. These matrices are generated for each time step of the game.

Figure 5 and Figure 6 show the average of the values in the X and Y correlation matrices over entire half-games. These averages are calculated for the entries in the 11×11 sub-matrices giving correlations between teammates (there were no negative elements in these sub-matrices, making the average a realistic measure). Although the Y-correlations show no significant change, the X-correlations show a substantial improvement in the movements of the 1998 winners. In fact, since this curve has two peaks it shows that there are some winners that could correlate and others that could not.

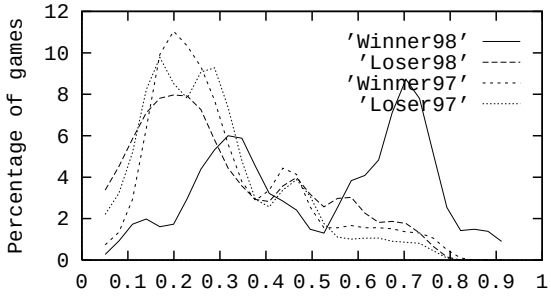

Fig. 5. The average X-correlation with teammates

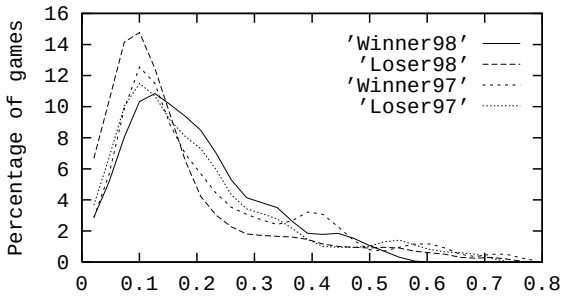

Fig. 6. The average Y-correlation with teammates

To examine teamwork in smaller groups, we also calculated the average X-correlation values for just a team's front 3 attackers and for its back 3 defenders (excluding the goalkeeper). Again, these results (Figure 7 and Figure 8) show a marked improvement in the winning teams of 1998. In contrast to this, the Y-correlations again do not

show an improvement. In fact, *all* the average values for Y-correlations in Figure 1 show a decrease in 1998. This apparent *worsening* of play is partially due to the skewing of the 1997 figures by a small number of games that exhibited

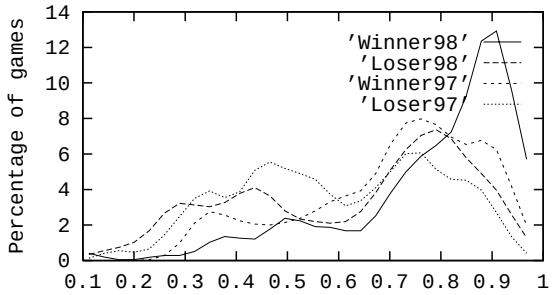

Fig. 7. The average X-correlation among attackers

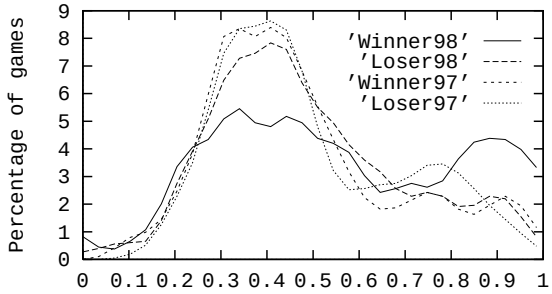

Fig. 8. The average X-correlation among defenders

the 'rugby effect': teams simply surrounding the ball with their players in a scrum that slowly moved across the field. Motivated by the introduction of the off-side rule, teams have improved their X-correlations to the extent that the 1998 average is higher than the biased 1997 value. For the Y-correlations, such improvements remain as further research.

One final use of correlations is to demonstrate the marking abilities of teams. The average of a team's X-correlations with opponents is shown in Figure 9 (the correlation matrix is symmetric, so there is only one graph for each year). This demonstrates that marking has increased from '97 to '98. In general, however, the symmetric nature of player correlations makes it difficult to identify *which* team's players are actually doing the marking. We are investigating how to collect more statistics on marking by examining the dependencies between the directions and the timings of players' movements.

3.3 Passing, Shooting & Dribbling Skills

Here, we look at the remaining sixteen features from Figure 1, which all deal with ball-handling skills. In RoboCup, client programs issue large numbers of kick commands. To distinguish kicks that have a genuine effect, we define a

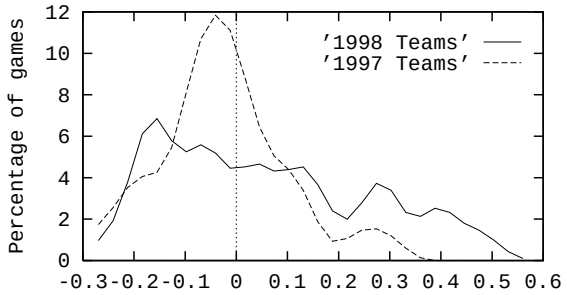

Fig. 9. The average X-correlation with opponents

genuine kick sequence as two successive successful kicks of the ball that occur with a time lag of more than 0.5s and at a physical separation of more than 0.5m. A successful pass is then a genuine kick sequence in which the two kicks are made by players of the same team; a dribble is a genuine kick sequence in which the two kicks are made by the same player.

A team's possession time is defined as starting with the first kick of a team's kick sequence, and continuing until either an opponent kicks the ball, or the ball leaves open play (a free kick, goal kick, throw in, corner, or center kick is awarded). We measure the possession of one team as the rate of the total possession time in the game. A *steal* occurs when the possession transfers from one team to another without the ball leaving open play. The pass success rate of a team is then defined as the ratio of its passes to the total of its number of passes plus the number of the opponent's steals. A *winning passwork pattern* is defined as any chain of three players A, B, C from the same team such that passes from A to B and from B to C occur at least once, and C scores at least one goal. Finally, a *pass using space* is a pass where the ball never comes within 5m of an opponent.

Let us look at some of the graphs for the ball-handling features. One of the most interesting is Figure 10, which shows the total number of shots on goal. The curves for the 1997 and 1998 winners both have multiple peaks, suggesting that there are two types of strategies for shooting. Another interesting conclusion can be drawn from Figure 11, which shows the average pass length. This suggests that shorter, more stable passwork is one of the keys for strong play. In Figure 12 we also show the average pass success rate. This is slightly better for winners than for losers, and has a small tendency to increase from 1997 to 1998.

In fact, most of the graphs for the ball-handling features show small improvements such as that of Figure 12 rather than the larger differences of Figure 10 and Figure 11. To provide a better picture of the true importance of ball-handling skills, we therefore carried out the *principal component analysis* described in the following section.

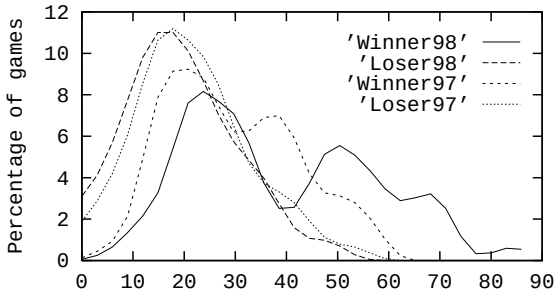

Fig. 10. The total number of shots on goal

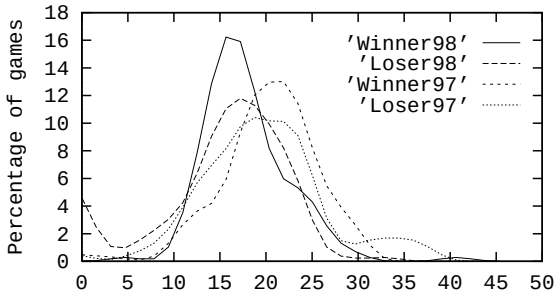

Fig. 11. The pass length average(m)

3.4 Principal Component Analysis

To use principal component analysis (PCA) in our analysis, we calculated the eigenvectors of the 32×32 correlation matrix of the features in Table 1. These eigenvectors give the directions in which the data cloud is stretched most, and so

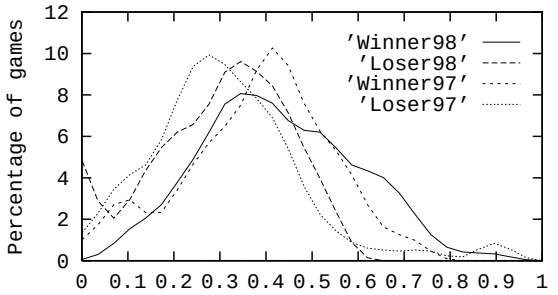

Fig. 12. The average pass success rate

allow us to identify the features that play the most important roles in RoboCup games. For the data from the 1998 RoboCup (including the drawn half-games), Table 1 shows the ten features that had the largest elements in the eigenvector corresponding to the largest eigenvalue. Note that the features related to ball handling are prominent. So, whereas the previous sections revealed that substantial progress has been made between '97 and '98 in team formation and co-ordinated movement, here we see that an underlying ability to handle the ball well is also crucial for success.

Table 1. The RoboCup98 features highlighted by PCA

Feature	Value
total number of passes	0.297
deviation of the number of passes per player	0.287
average pass success rate	0.271
number of pass chains	0.269
number of backwards passes	0.234
average distance covered by one player	0.230
number of players making zero passes	-0.222
total number of shots on goal	0.219
average X-correlation with teammates	0.214
average X-correlation among attackers	0.194

We also used principal component analysis to identify the 1998 RoboCup games that produced the best play. For each half-game, we found the inner product of the vector of the feature values (for each team) with the vector of largest eigenvalues. Since this produces larger values for teams that played well, we can identify the good teams independently of the actual tournament results. Table 2 shows the ten teams that produced the highest scoring games. Note that these teams all fared well in the actual tournament itself. Our analysis demonstrates in detail the reasons for this success.

Table 2. PCA top ten half-games played in RoboCup98

	Team (final rank)	Game
1	AT Humboldt (2)	1st half *vs* Gemini
2	CMUnited (1)	1st half *vs* Andhill98
3	CMUnited (1)	2nd half *vs* Rolling Brains
4	CMUnited (1)	2nd half *vs* Kasugabito II
5	AT Humboldt (2)	2nd half *vs* Andhill98
6	CMUnited (1)	2nd half *vs* Andhill98
7	Tumsa (Best 12)	1st half *vs* Gemini
8	Windmills (3)	2nd half *vs* AT Humboldt
9	Andhill98 (Best 6)	1st half *vs* Gemini
10	Windmills (3)	1st half *vs* AT Humboldt

3.5 Statistics Summary

The above sections have demonstrated the progress in the RoboCup simulator league from 1997 to 1998. Our overall conclusions can be summarized as follows:

- Teams work together to follow the play better in '98, despite having to manage stamina more carefully.
- Teamwork has improved significantly along the long side of the field, but not along the shorter.
- Improved teamwork can be seen in '98 in the way that forwards (and defenders) share their roles.
- Teams carry out more marking in '98.
- Ball-handling skills have not improved significantly on average, but principal component analysis shows that they are crucial for strong play.

These results demonstrate that progress has been made towards meeting the teamwork Synthetic Agent challenge. To a lesser extent, an improvement in opponent modeling is also demonstrated by our analysis of marking.

4 Prospects: A statistics proxy server

The statistical analyses we have described above answer the question of the extent to which soccer skills in RoboCup have progressed *to date*. In addition to this, however, the code that we have implemented to carry out this analysis also has a direct bearing on the *future prospects* of research on the Synthetic Agent Challenge.

In particular, our code produces its results in real-time, as a game is running. Thus, we suggest offering this code to the RoboCup community in the form of an independent proxy server. In our prototype server, we have incorporated two further techniques developed by [Tanaka-Ishii *et al.*1998]: the representation of ball-play chains as first order Markov processes, and the calculation of players' defensive areas with Voronoi diagrams. Figure 13 shows the likely uses of such a proxy server.

In addition to facilitating research on automated commentators and post-game analyzers, then, we envisage that a proxy server will directly enable research on the three challenge problems. Obvious applications are:

- facilitate off-line learning from game logs and on-line learning via the touch-line coach planned for 1999,
- identify the opponent's key players, passwork patterns and styles of play (opponent modeling),
- identify players not fulfilling their assigned roles, enabling suitable adaptations to be made (teamwork).

Another application we foresee is related to the specific evaluation procedures laid out in the challenge paper itself. To our knowledge, only one of these evaluations — a test of team robustness — has so far been carried out. This test took place at RoboCup 1998, and involved 13 of the 37 participating teams. With the 1997 champions as the fixed opposition, each of these 13 teams played a single

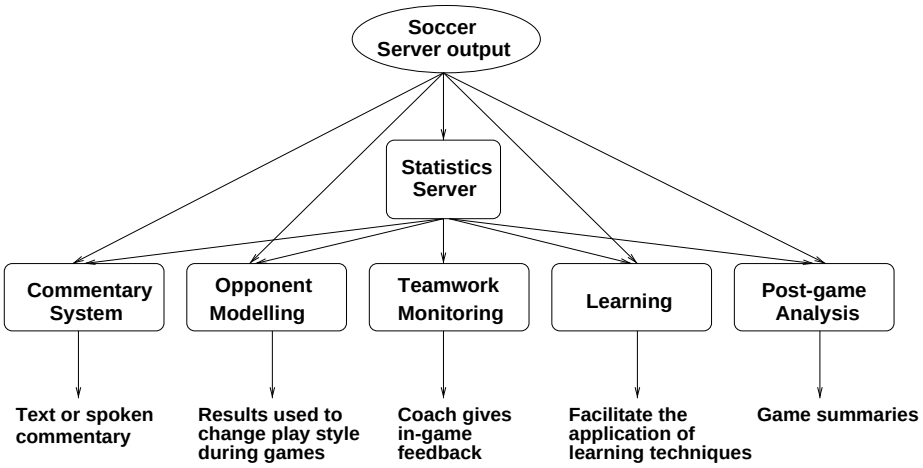

Fig. 13. Likely uses of a statistics proxy server

half-game with 11, 10, 9 and then 8 players. The logs of this test can be found at `http://www.isi.edu/soar/galk/RoboCup/Eval/`. We are not aware of any detailed analysis, but the logs clearly demonstrate that evaluation will not be straightforward. As might be expected, conducting just one half-game with each number of players results in noisy data. For example, the CAT-Finland team lost 0-1 when playing with 11 players, but with just 10 players it won 1-0, and with 9 players it drew 1-1.

Thus, to generate meaningful results from the robustness test, statistical analyses of the kind presented in this paper will be important. Indeed, since almost all of the evaluation procedures laid out in the challenge paper involve the playing of games, we expect to see our proxy server used in this way very often.

5 Prospects: A modular common model

We feel that what is ultimately most important in RoboCup evaluation is identifying the *AI techniques* that produce good performance; only by identifying the best techniques can they be effectively re-used in the future to advance RoboCup research. Within the current framework of RoboCup, however, this kind of evaluation is hard to carry out. Almost all teams use a combination of a number of techniques. So, even with a detailed statistical analysis of full games, it is in general difficult to identify which AI techniques worked best for particular aspects of the game. Similarly, the evaluation procedures described in the challenge paper focus on varying the conditions under which a team plays a game rather than varying the AI techniques used for particular sub-problems. The prevailing tendency in RoboCup, then, is simply to assume that teams that win are the ones that employ good techniques. This must be true to some extent,

but the reverse is certainly not true: teams that lose do *not* necessarily employ bad techniques. They may actually excel at some aspects of soccer whilst being let down by others.

We suggest that the current situation in the RoboCup simulator league is similar to that in another research field often regarded as a challenge problem: Computer Go. Superficially, this field appears very healthy, with over 100 active programs being developed (including about ten that are commercial), and a number of computer tournaments (two with the status of World Championships). However, it was recently pointed out by [Müller1998] that progress is much slower than it could be. A typical program has between 50 and 100 modules, and it is generally the *weakest* of these that determines the strength of the program. Thus, the field progresses by all of the program developers simultaneously trying to improve the level of their worst modules. Müller proposed that the overall rate of progress would be significantly improved if researchers worked together on the implementation of a *common model*.

Although RoboCup is young compared to Computer Go, we believe that a similar argument applies. A common model would improve the rate of progress, provided it has the essential feature of being *modular*. Modularity is important because it is this that facilitates the isolation and direct evaluation of competing techniques for any individual aspect of the game. This in turn increases the chances of being able to identify and retain the best aspects of every RoboCup team, including those with poor overall match results.

Thus, we propose the institution of at least one modular RoboCup team based on a common model. To function successfully and to benefit most from the available manpower, we suggest that such a team should be:

- open source; anyone is free to use the code, and to contribute improved versions of any module.
- well-tested; contributors demonstrate the advantages of their modules with match results or statistical analysis such as that described in this paper.
- well-documented; the learning curve is as painless as possible.
- updated frequently; there is an efficient organization for selecting and integrating currently best modules that keeps all contributors informed of progress.

Already, the first two RoboCups have identified important sub-tasks of soccer. A small sample of these tasks includes the automatic learning of passwork skills, the pre-compilation of communication options, static evaluation functions for player positioning, and the specification of a small number of distinct 'player modes'. We believe that formalizing a common model incorporating such modules will expedite RoboCup research. Ultimately, it will also lead to a deeper understanding of soccer itself, and of the multiple challenges it represents.

6 Conclusions

We have used statistics to qualitatively demonstrate the progress to date in the simulator league of RoboCup. Our results highlighted the changes in play

from 1997 to 1998 in terms of 32 features. We also demonstrated the utility of examining the correlation of players' movements and of applying principal component analysis.

Further, we emphasized that our analysis is important to the *future* of the RoboCup research. We proposed the free distribution of a proxy server that can supply in real-time the full range of statistical analyses described in this paper, and pointed out that such analyses will have a crucial role in facilitating and evaluating research on the challenge problems of opponent modeling, teamwork, and learning. Finally, we suggested that to make the best use of the efforts of RoboCup researchers, at least one modular team based on a common model should be instituted.

References

[Kitano *et al.*1997] H. Kitano, M. Tambe, P. Stone, M. Veloso, S. Coradeschi, E. Osawa, H. Matsubara, I. Noda, and M. Asada. The RoboCup synthetic agent challenge 97. In *Proceedings of IJCAI-97*, pages 24–29, Nagoya, Japan, 1997.

[Matsubara *et al.*1998] H. Matsubara, I. Frank, K. Tanaka-Ishii, I. Noda, and H. Nakashima Automatic soccer commentary and RoboCup. In *Proc. of the 2nd Intl. Wkshop on RoboCup*, pages 7–22, 1998.

[Müller1998] M. Müller. Computer go: a research agenda. In H.J. vd Herik and H. Iida, eds., *Proc. of The 1st Intl. Conf. on Computers and Games*, Japan, 1998. Springer-Verlag. LNCS series, vol. 1558.

[SoftSport1998] SECOND LOOK soccer analysis software. The makers, SoftSport Inc., can be found on the Web at `http://www.softsport.com/`, 1998.

[Takahashi *et al.*1998] T. Takahashi, T. Naruse. From play recognition to good play detection. *Proceedings of Robocup98 Workshop*, pages 129–134, 1998.

[Tanaka-Ishii *et al.*1998] K. Tanaka-Ishii, I. Noda, I. Frank, H. Nakashima, K. Hasida, and H. Matsubara. Mike: An automatic commentary system for soccer. In *Proceedings of ICMAS98*, Paris, France, 1998.