Learning by Demonstration of Coordinated Plans in Multiagent Systems

Marco A. C. Simões[1,2], **Tatiane Nogueira**[1]

[1]Instituto de Computação – Universidade Federal da Bahia (UFBA)
Programa de Pós Graduação em Ciência da Computação (PGCOMP)
Salvador – BA – Brazil

[2]Universidade do Estado da Bahia (UNEB)
Centro de Pesquisa em Arquitetura de Computadores, Sistemas Inteligentes e Robótica (ACSO)
Salvador – BA – Brazil

msimoes@uneb.br, tatiane.nogueira@ufba.br

Abstract. *One of the significant challenges in Multiagent Systems (MAS) is the creation of cooperative plans to deal with the different scenarios that present themselves in a dynamic, real-time environment composed of teams of mobile robots. This work involves capturing human knowledge to demonstrate how robot teams can better cooperate in solving the problem they must solve. The research used the environment RoboCup 3D Soccer Simulation (3DSSIM) and the collection of human demonstrations were carried out through a set of tools developed from adapting existing solutions in the RoboCup community using a crowdsourcing strategy. In addition, fuzzy clustering was used to gather human demonstrations (setplays) with the same semantic meaning, even with minor differences. With the data organized, this work used a reinforcement learning mechanism to learn a classification policy that allows agents to decide which group of setplays is best suited to each situation that presents itself in the environment. The results show the ability of the robot team to evolve, from learning the suggested setplays and their use in an appropriate way to the skills of each robot.*

Student level: Ph.D.
Date of conclusion: July 5, 2022.
Link for full thesis: `http://www.acso.uneb.br/bahiart/uploads/Main/SimoesTesePhD.pdf`

Keywords— Multiagent Systems, Reinforcement Learning, Fuzzy Clustering, RoboCup, Robot Soccer.

1. Introduction

Machine learning meets an essential demand of robots to learn intuitive or even instinctive behaviors and knowledge of the human being. For example, although the movement "walking" is natural to the humans, the algorithmic understanding of how the organism executes it is not known. Machine learning allows robots to walk similarly to humans, learning to generalize movements from examples.

When we consider robots as a group or a team, we can model them as a *Multiagent Systems* (MAS). MASs are systems composed of multiple interactive computational elements known as agents [Wooldridge 2009]. An agent is an element capable of perceiving its environment through sensors and acting on this same environment through actuators. An agent is considered intelligent or autonomous if it has a fundamental characteristic: autonomy. An agent is autonomous if it can decide its actions on its own, without human intervention [Russell and Norvig 2021]. Another fundamental characteristic for an agent to be part of an MAS is the ability to interact with other agents. Interaction is not restricted to exchanging information; it must include some social activity such as cooperation, coordination, negotiation, etc [Wooldridge 2009].

This work focuses on a class of problems whose environments have the following properties: partially observable; stochastic; dynamic; continuous; multiagent; real-time [Russell and Norvig 2021]. We can reduce this class of problems to the standard challenge, chosen by scientists, of robot football [Kitano et al. 1998]. Robots must be able to make complex decisions in a short time, cooperating with allied robots and competing against robot or human opponents to meet this challenge. Since 1997, the RoboCup Federation[1] has promoted scientific development in artificial intelligence and robotics through scientific competitions between robots.

The thesis described by this paper aims to present experimental evidence that it is possible to capture humans' intuitive or unstructured knowledge when watching a robot game playing soccer to compose a dataset for training the robot team. Coordinated collective plays, called setplays, compose the dataset. We use a deep reinforcement learning solution for the team to learn a setplay selection policy from a large set of demonstrations performed by humans.

This paper was situated in state of the art through a systematic literature review, whose main results are described in section 2. In section 3, we present the solution built during the Ph.D. thesis summarized by this paper. Section 4 describes the assessment and results, and section 5 sets forth our conclusion and future work.

2. Related Work

When we look for works related to cooperative plans applied to robot soccer, we can find an important framework for designing use setplays in teams of robot soccer players [Mota et al. 2010]. This work was complemented later with a graphical interface [Reis et al. 2010], [Cravo et al. 2014]. Although these works represent an important landmark in the area of MAS applied to robot soccer, they do not use any machine learning approach. The designers should manually create each coordinated plan for the robots.

Learning from Demonstration (LfD) was used to learn low level robot skills [Freelan et al. 2014]. The behavior was decomposed in low level actions such as look for the ball, align to goal and kick. Despite the authors claiming that the solution applies to high-level collective behaviors, all the experiments described included only low-level robot skills. Another work, introduces the crowdsourcing approach in the context of robot soccer [Moradi et al. 2016]. However, it uses reinforcement learning only to train individual behaviors of the robot possessing the ball. We can find works that use reinforcement learning to learn the best transition in a state machine that represents a setplay [Fabro et al. 2014], or individual decision-making by robot soccer players [Shi et al. 2018]. Some works investigate the transfer of knowledge from the simulated environment to real robots [Bianchi et al. 2018].It is also worth mentioning the presence of many works that use deep reinforcement learning to train skills in soccer robots, such as walking, running, kicking [Abreu et al. 2019], [Melo et al. 2021], [Abreu et al. 2021], [Spitznagel et al. 2021], [Teixeira et al. 2020].

No works were found that use approaches based on LfD and deep reinforcement learning

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

to take advantage of the critical view of human spectators to point out opportunities for better agents' performance in a robot soccer MAS. This thesis fills this gap in state-of-the-art, presenting experimental evidence that the intuitive knowledge of human spectators can improve robot soccer team performance.

3. Learning Setplays from Demonstration

The solution to the problem of learning setplays from demonstrations of human spectators was divided into three stages, as illustrated in Figure 1. In the first stage, volunteers are expected to watch the MAS *Bahia Robotics Team* (BahiaRT)[2] matches in recent official competitions using a version of the official *RoboCup 3D Soccer Simulation* (3DSSIM) *Logplayer* (*RoboViz*)[3] modified in this work, taking breaks in situations that they consider that the robots simulated in BahiaRT had an unwanted collective behavior, or underperform. The scenes captured in stage 1 are taken to stage 2, where LfD takes place. To support the first stage, we selected a subset of features present in the BahiaRT world model [Simões and Nogueira 2018].

In the second stage, the *Strategy Planner* (SPlanner) [Cravo et al. 2014] tool was changed so that it can be used effectively as a demo generator in the 3DSSIM environment, making it able to start a new *setplay* from the scene captured in *RoboViz*. In section 3.1, we describe our crowdsourcing strategies and the toolkit produced for this work.

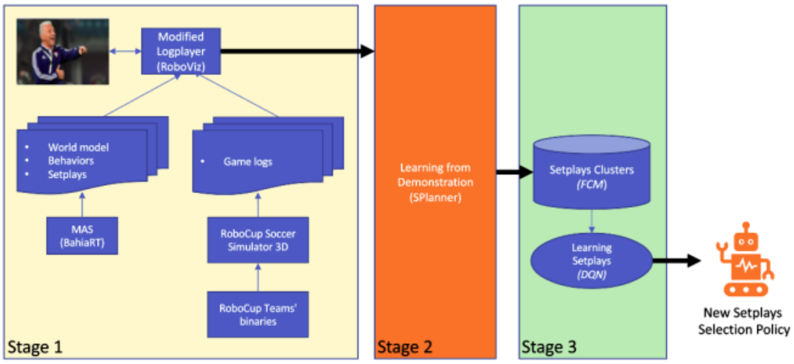


Figure 1. Learning Setplays from Demonstration: complete solution split into three stages.

In the third stage, we developed a fuzzy clustering engine to organize the dataset by setplays similarity. This is important to turn the dataset able to be used by agents in real-time, and to deal with semantic equivalence as defined in section 3.2.

The clustered dataset is used by a deep reinforcement learning solution based on algorithm *Deep Q Network* (DQN) [Mnih et al. 2015] to learn a setplays selection policy. The Section 3.3 describes the solution.

3.1. BahiaRT Collecting Setplays Toolkit

The construction of the database with *setplays* is based on a strategy of *crowdsourcing* in which people from anywhere in the world can contribute by playing the role of the human spectator.

We built a set of tools to support this strategy, bringing together the modified versions of RoboViz and SPlanner [Simões et al. 2021]. To make it easy to install and use the tools, they

[2] The BahiaRT is the scientific competition team from State of Bahia University. More details available at https://www.acso.uneb.br/bahiart

[3] https://github.com/magmaOffenburg/RoboViz

were organized into a *docker container*[4], preventing users from compiling the tools, installing libraries, and solving problems with dependencies. This organized set of tools for collecting *setplays* was called **BahiaRT Setplays Collecting Toolkit** [Simões et al. 2022] and made publicly available together with all the necessary documentation for publication in the repository `https://bitbucket.org/bahiart3d/setplaysdataset`.

We added a demonstration mode to RoboViz to allow users to watch game logs and pause and capture any scene in the game to start a new setplay demonstration. When a new demonstration is launched, the user can choose the team to whom the user will make the demonstration, the type of setplay (offensive or defensive), and the play mode when the setplay will start. Then he chooses the teammates and opponents players who will participate in the new setplay [Simões et al. 2021].

SPlanner has also gained a demo mode capable of importing the demo file generated by RoboViz. When starting, SPlanner already creates a new setplay using the parameters defined in RoboViz and positions the screen at step 0 with all participants selected. The user will then be able to use its graphical interface to create collective play suggestions for the robot soccer team. n this work, we also completed the SPlanner development by adding support for opponents' players, defensive setplays, and new offensive behaviors [Simoes et al. 2020].

The BahiaRT Setplays Collecting toolkit includes a submission form for users to send the text files generated by SPlanner containing all the information describing the setplays created as demonstrations [Simões et al. 2022]. During four months, we received 382 setplays' demonstrations. The following subsection describes our solution for organizing this dataset in clusters.

3.2. Organizing the Dataset

The crowdsourcing strategy adopted in this work potentially generates an arbitrarily large number of instances in the dataset. This fact can make it unfeasible to use in the 3DSSIM environment that uses a $20ms$ simulation cycle. However, the chances of having equivalent *setplays* are high, with many different people generating demos from robot soccer games. They would not be *setplays* precisely the same since many of the attributes that make up a *setplay* have continuous values, reducing the probability of absolute equality. Then, we define that this equality between the *setplays* will exist when there is a **semantic equivalence** [Simões et al. 2020].

Definition 1 (Semantic Equivalence) *Two setplaysSP_i and $SP_j, i \neq j$, are considered semantically equivalent if they represent the same play at the abstract domain knowledge level.*

To organize the dataset of setplays, we used a two-level strategy, splitting the set of features that describe a setplay into two subsets. The first subset has four features: (i) our players number; (ii) their players number; (iii) abort condition; (iv) number of steps. These features are integer values, except for *abort condition* which is a boolean expression represented here as a binary tree. We extracted these features after analysis of setplay files generated by SPlanner [Simões et al. 2020]. A 5th feature might complete this set: (v) the list of steps. However, this feature is a list of objects of type *Step* described by nine additional features. So, we expanded the list of steps in the second subset of features. We use the algorithm *Fuzzy C-Means* (FCM) to organize the dataset in clusters considering only the features (i)...(iv) in the first level. The Figure 2 shows the two-level FCM approach we use in this work.

We used the feature (v) in the second round of execution of algorithm FCM, applying it to each cluster generated in the first level. So, the feature (v) expands to second level: (v.i) our players in Step; (v.ii) their Players in Step; (v.iii) wait time; (v.iv) abort time; (v.v) our players list; (v.vi) their players list; (v.vii) next step; (v.viii) transition condition; (v.ix) behaviors list.

[4] `https://www.docker.com/`

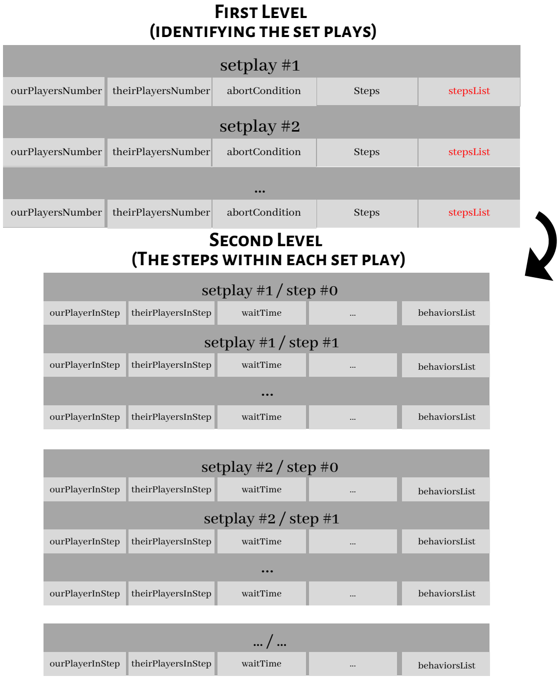

FIRST LEVEL
(IDENTIFYING THE SET PLAYS)
setplay #1
ourPlayersNumber | theirPlayersNumber | abortCondition | Steps | stepsList
setplay #2
ourPlayersNumber | theirPlayersNumber | abortCondition | Steps | stepsList
...
ourPlayersNumber | theirPlayersNumber | abortCondition | Steps | stepsList
SECOND LEVEL
(THE STEPS WITHIN EACH SET PLAY)
setplay #1 / step #0
ourPlayerInStep | theirPlayersInStep | waitTime | ... | behaviorsList
setplay #1 / step #1
ourPlayerInStep | theirPlayersInStep | waitTime | ... | behaviorsList
...
ourPlayerInStep | theirPlayersInStep | waitTime | ... | behaviorsList
setplay #2 / step #0
ourPlayerInStep | theirPlayersInStep | waitTime | ... | behaviorsList
setplay #2 / step #1
ourPlayerInStep | theirPlayersInStep | waitTime | ... | behaviorsList
...
ourPlayerInStep | theirPlayersInStep | waitTime | ... | behaviorsList
... / ...
ourPlayerInStep | theirPlayersInStep | waitTime | ... | behaviorsList

Figure 2. The two-level FCM architecture organizes the dataset in clusters.

The features (v.i). . .(v.iv), and (v.vii) are scalar values. The properties (v.v) and (v.vi) are lists of pairs of Cartesian coordinates which identify all players' positions in the current step of a setplay. The feature (v.viii) is a boolean expression represented as binary tree and the feature (v.ix) is a list of strings describing the behaviors executed by each teammate on current step.

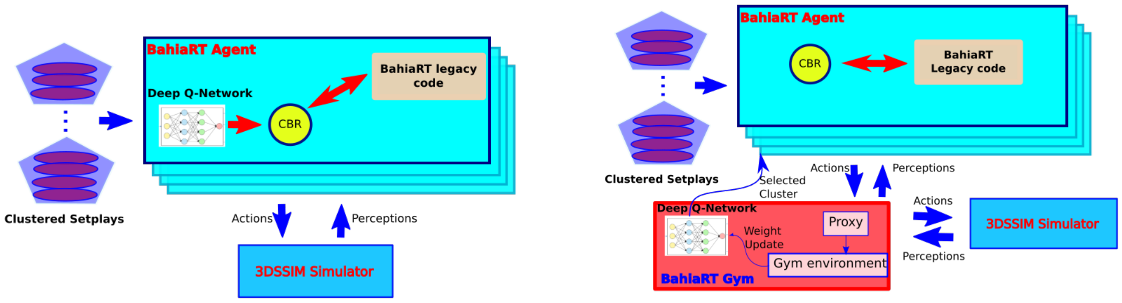
The FCM requires the definition of an appropriate distance measure to measure the similarity between the dataset instances. Euclidean distance is commonly used to estimate the distance between two instances with scalar properties. However, the proposed dataset schema contains some non-scalar data types. In this work, we defined new distance norms for features represented as a list of Steps, binary trees, a list of Cartesian pairs, and a list of strings. We used the new norms to modify the standard FCM distance calculus between instances and centroids [Simões et al. 2021].

We used the FCM with the new distance norms to organize the dataset of setplays in clusters to be used in the reinforcement learning solution described in the following subsection.

3.3. Learning a Setplays Selection Policy

FCPortugal Setplays Framework (FSF) [Mota et al. 2010] uses a *setplays* manager based on the approach of *Case-Based Reasoning* (CBR) [Wangenheim and Wangenheim 2003]. This approach builds a case history from the agents' use of *setplays*. The team BahiaRT extends the FSF to support setplays execution. However, the CBR solution is not scalable to a large dataset of setplays. This work presents a Deep Reinforcement Learning (DRL) strategy to learn a new setplays select policy to choose one of the clusters of the dataset generated by the solution described in subsection 3.2. So, the CBR applies on the setplays in the selected cluster. The complete solution is exhibited in Figure 3a.

The strategy uses the DQN algorithm using a Deep Q-Network to represent the learned policy. The DQN receives the clustered dataset and the properties from BahiaRT's world model as input and generates a cluster number as output. The CBR loads the setplays definitions of this



(a) BahiaRT's new architecture using the learned setplays selection Policy implemented in a Deep Q-Network.

(b) BahiaRT's training architecture uses BahiaRT Gym

Figure 3. DQN solution to learn and execute a new setplays selection policy.

cluster and selects the best setplay to use in the current situation.

We use Open AI Gym [Brockman et al. 2016] for training the Deep Q Network and the DQN implementation in stable baselines 3 [Raffin et al. 2021]. Figure 3b shows the complete training architecture.

As Open AI Gym does not offer a 3DSSIM environment, we developed our environment named BahiaRT Gym [Simões et al. 2022]. This environment is uncoupled from the BahiaRT team's code. Any 3DSSIM team can use BahiaRT Gym for DRL experiments. Both single agent and multiagent training are available. All perceptions sent by the simulator to the agents and the actions sent from agents to the simulator are available in BahiaRT Gym for use in the observation space or reward calculus. BahiaRT Gym also connects the agents to send exploratory actions during training. The observation space used for training the setplays selection policy is defined from the results of preliminary experiments [Simões and Nogueira 2018] and the results of the setplays files collected using the BahiaRT Setplays Collecting Toolkit [Simões et al. 2022]. Players positions, ball position, play modes and ball's field zone are the features used in the observation space. The action space A is defined by an integer representing the identification of the group of *setplays* generated by the dataset organizer:

$$A = \{1, \dots, C^*\}, \tag{1}$$

where C^* is the total number of groups found by the *fuzzy* organizer of the dataset defined in the section 3.2. The algorithm updates the weights of the Q-Network when a training episode ends using the reward function defined as

$$r(s_i, a_i) = \frac{\Delta x_B}{|\Delta x_B|} \times 2^{|\Delta x_B|} \times \left[1 + 10 \times (\text{flag}_{\text{GS}} + \text{flag}_{\text{GC}}) + 3 \times \text{flag}_{\text{SS}} + \frac{\text{flag}_{\text{FS}}}{2}\right], \tag{2}$$

where $s_i \in O$ is the state observed at the instant i at the beginning of the training episode and $a_i \in A$ is the action chosen by the agent at the instant i. $\Delta x_B = x_B^f - X_B^s$, where x_B^f and X_B^s are the coordinates on the x axis of the ball at the final and initial instants of the episode, respectively. The flags define boolean conditions and receive value 1 for *true* and 0 for *false*. flag_{GS} is true when the BahiaRT scores a goal in the episode, and flag_{GC} is true if BahiaRT concedes a goal in the episode. flag_{SS} is true when a successful setplay finishes in the episode, and flag_{FS} is true if a failed setplay ends in the episode.

This subsection describes a DRL for training a setplays selection policy represented by a Q-Network. The following section presents the results of the experiments used for assessment.

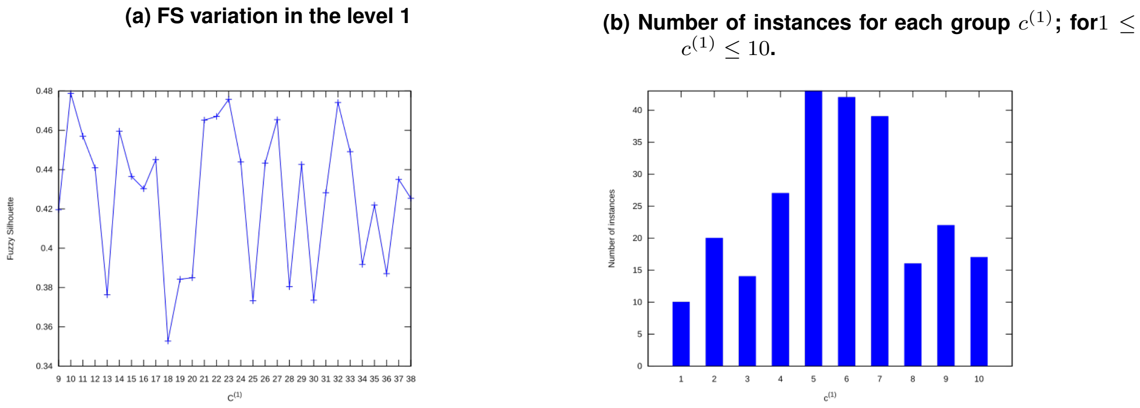
4. Assessment

We assessed the results of our solutions in two parts. In the first one, we evaluated the appropriateness of the dataset clustering (see subsection 4.1). Then we assessed the learned setplays selection policy regarding its effects on BahiaRT's performance, as described in subsection 4.2.

4.1. Assessing dataset clustering

One of the main challenges in using clustering algorithms is defining the number of clusters. We defined two number of clusters, one for the first level $C^{(1)}$, and the other for the second level $C^{(2)}$.

We executed the FCM algorithm using values of $C^{(1)} = \frac{\sqrt{n}}{2}, \ldots, 2 \times \sqrt{n}$, where $n = 382$ is the total number of setplays in the dataset. For each value of $C^{(1)}$, we run the algorithm 10 times to minimize the effects of random initialization of centroids. We use a Cluster Validation Index (CVI) named Fuzzy Silhouette (FS) [Eustáquio et al. 2018] to assess the best value for $C^{(1)}$ regarding the dataset used in this experiment. The FS is a maximization index in the interval $[-1; 1]$.



(a) FS variation in the level 1

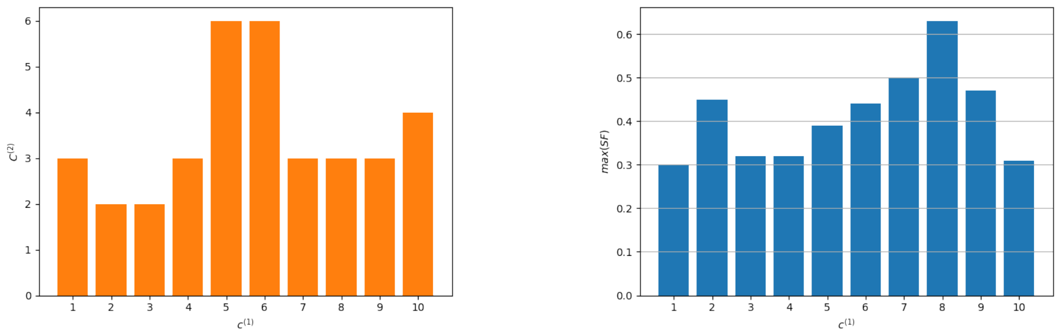
(b) Number of instances for each group $c^{(1)}$; for$1 \leq c^{(1)} \leq 10$.

Figure 4. Results of experiments for a dataset with $n = 382$ instances, and $9 \leq C^{(1)} \leq 38$.

Figure 4a shows the variation of FS for $9 \leq C^{(1)} \leq 38$. The higher value of FS is obtained for $C^{(1)} = 10$. We split the dataset into $C^{(1)} = 10$ clusters and got the distribution of setplays exhibited in the figure 4b. Each group $c^{(1)} = 1, \ldots, 10$ has a different number of setplays with a considerable difference. While the average is 25 setplays per group, there are groups in the order of 40 instances.

We followed the same procedure for the second level executing 10 instances of the FCM algorithm for each cluster $1 \leq c^{(1)} \leq 10$, considering $\frac{\sqrt{n_{c^{(1)}}}}{2} \leq C^{(2)} \leq 2 \times \sqrt{n_{c^{(1)}}}$. $n_{c^{(1)}}$ is the number of instances in cluster $c^{(1)}$.

Figure 5a shows that each group $c^{(1)}$ can be split into 2 to 6 subgroups. When we sum the number of groups $C^{(2)}$, for all $c^{(1)} = 1, \ldots, 10$ clusters, we get a total of $C^{*} = 35$ groups, which we use to define the action space for learning the setplays selection policy, as described in section 3. The distribution of groups is a consequence of the values of FS found in the second level as illustrated in figure 5b. It is noticeable that, for some groups, most values of FS increased compared to level 1.



(a) Number of clusters on level 2($C^{(2)}$) to split each group ($c^{(1)}$) from level 1.

(b) Maximum FS value for each cluster $c^{(1)}$ after execution of FCM on level 2.

Figure 5. Experimental results for clustering on level 2 using a dataset with $n = 382$ instances split into ten groups in level 1.

The next subsection presents the assessment of the learned setplays selection policy regarding its effects on the BahiaRT's overall performance.

4.2. Assessing the Learned Setplays Selection Policy

The learned setplays selection policy assessment starts with selecting three opponent teams for training and evaluation. The selected teams are a sample representing teams with historical performance better (magmaOffenburg[5]), similar (ITAndroids[6]), and worse (WITS-FC[7]) than BahiaRT.

We executed BahiaRT using the training architecture defined in section 3. The BahiaRT Gym [Simões et al. 2022] controls the episodes start and reset and updates the reward to the Q-Network. We used a set of 100 matches against each selected opponent, alternating the opponents to avoid overfitting in the Q-Network.

After training the Q-Network, we executed another series of 300 matches (100 for each opponent) to measure the overall performance when compared to the team before using the new learned policy. As the reward function consider the ball displacement as one of its main variables, we used heat maps of the ball's position to assess the results. We also used some game statistics as shown in Tables 1 (baseline before training) and 2 (after training).

Table 1. BahiaRT's games before using the learned setplays selection policy.

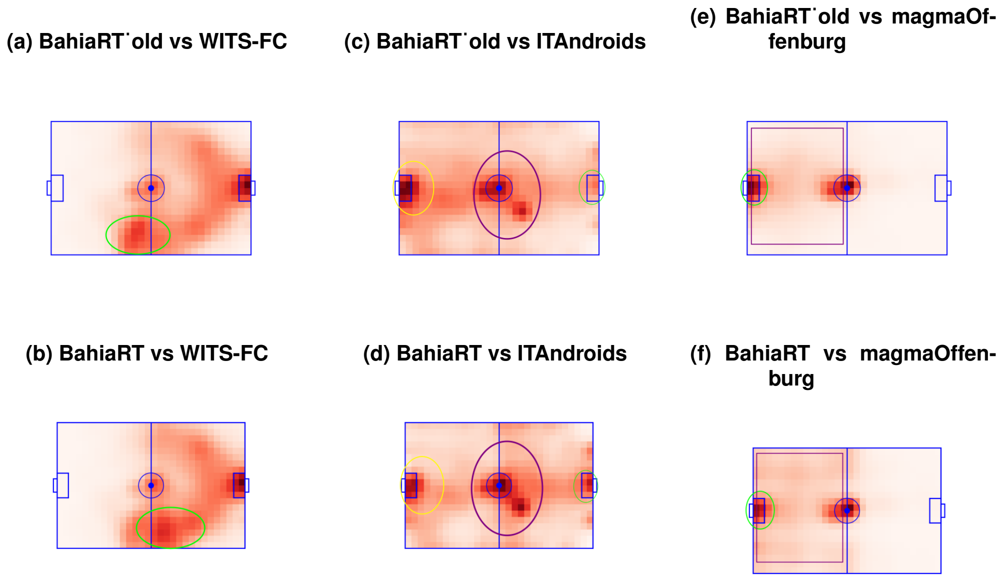
Opponents	Wins	Draws	Loses	Avg. GS	Avg. GC
ITAndroids	16	58	26	0,31	0,41
WITS_FC	59	41	0	0,87	0
magmaOffenburg	0	0	100	0,01	5,64

When we compare each line of the two Tables, it is clear that some increase in the number of Wins is detected in matches against ITAndroids and WITS-FC. The increase in the average of

[5] Team from Hochschule Offenburg University, Germany. https://robocup.hs-offenburg.de/

[6] Team from Instituto Tecnológico da Aeronáutica (ITA), Brazil. http://www.itandroids.com.br/

[7] Team from Wits University, South Africa. https://www.wits.ac.za/



(a) BahiaRT˙old vs WITS-FC

(c) BahiaRT˙old vs ITAndroids

(e) BahiaRT˙old vs magmaOffenburg

(b) BahiaRT vs WITS-FC

(d) BahiaRT vs ITAndroids

(f) BahiaRT vs magmaOffenburg

Figure 6. Heat maps of ball's position in 100 matches against each opponent before and after setplays selection policy training.

Goals Scored (GS) and the decrease in the average of Goals Conceded (GC) explain the results. On the clash against magmaOffenburg, we noticed a slight reduction in the average of GC. Still, this result is insufficient to turn the overall performance of team BahiaRT similar to magmaOffenburg.

Table 2. BahiaRT' game after training the new setplays selection policy.

Opponents	Wins	Draws	Loses	Avg. GS	Avg. GC
ITAndroids	26	53	21	0,45	0,28
WITS_FC	67	33	0	1,03	0
magmaOffenburg	0	0	100	0,01	5,14

A deeper analysis is possible when we evaluate the heat maps of the ball's position. We call *BahiaRT˙old* the original team before training and *BahiaRT* is the team with the new learned policy. In all heat maps, the BahiaRT defense in on the left side and the attack is on the right side.

When we compare the ball's position heat map resulting from 100 matches against WITS-FC, we can see a difference in the region inside the green circle in Figures 6a and 6b. The ball's position concentration in this area in Figure 6a moved forward to the attack field in Figure 6b. From this highlighted region in Figure 6b, the BahiaRT's agents can perform a direct shot to goal and score. These results explain the increase in the number of goals scored in matches against WITS-FC in Table 2.

The clash against ITAndroids presents three points of attention. The first point is the region within the yellow circle in Figures 6c e 6d. The yellow circle shows that the ball positions are less concentrated in the BahiaRT's defensive goal area in Figure 6d than in Figure 6c. These results explain the decrease in the number of goals conceded in Table 2. The highlighted regions in the purple and green circles show an increase in the concentration of ball positions in the offensive midfield and BahiaRT's attack goal area. This fact explains the increase in the number of goals verified in Table 2.

The Figures 6e e 6f shows two highlighted zones: the BahiaRT's defensive goal area (green circle) and the entire defensive midfield (purple rectangle). In the green circle, we can note a decrease in the ball's position concentration in the BahiaRT's defensive goal area. The purple rectangle shows the ball positions more spread out in the defensive midfield. These two facts explain the reduction in the number of goals conceded. The use of defensive setplays in the learned setplays selection policy allows the team BahiaRT to use more efficient defensive behavior, turning harder to magmaOffenburg to perform fast passes and score goals. In the next section, we discuss the results, present our conclusions and some possible future work.

5. Conclusion and Future Work

The results presented in section 4 show evidence that it is possible to collect demonstrations from human spectators to teach a MAS of robot soccer players better setplays to be used in the diverse situation. The evolution of game statistics is explained by analyzing the ball's position heat maps and exposing the influence of the reward function used in the DRL strategy. We highlight that the team BahiaRT is not in the state-of-the-art of basic skills like walking and kicking. However, it presents a clear evolution when playing against opponents of different levels. It is clear evidence that the high-level strategy, using intuitive human knowledge, can increase the overall MAS performance.

The thesis summarized in this paper generated several products: (1) a DRL solution to learn a new setplays selection policy using demonstrations collected using a crowdsourcing strategy [Simões 2022]; (2) BahiaRT Gym: an open software available for the community to execute any DRL experiment using the 3DSSIM simulator [Simões et al. 2022]; (3) BahiaRT Setplays Selection Toolkit: a set of tools to allow users to watch games and build setplays demonstrations to send to the authors [Simoes et al. 2020] [Simões et al. 2022]; (4) A new FCM organizer to the setplays dataset able to deal with non-scalar data types [Simões et al. 2020] [Simões et al. 2021]; (5) a dataset of almost 400 setplays that will be published to be used by the community [Simões 2022].

These results show a clear contribution of this work to the advancement of the state-of-the-art and some technical contribution to the research community that can use the tools and dataset produced on this work for their research projects. The thesis produced five publications [Simões and Nogueira 2018], [Simoes et al. 2020], [Simões et al. 2020], [Simões et al. 2021], and [Simões et al. 2022]. The work was also awarded as best oral presentation(2020), second best oral presentation(2018) and third best oral presentation(2021) in the Workshop de Estudantes de Pós-Graduação em Ciência da Computação (WE.PGCOMP/UFBA). The BahiaRT Gym was presented and awarded second place in the RoboCup 2022 3DSSIM Free Scientific Challenge, which contributed to the third place for team BahiaRT in the Technical Challenge[8].

This thesis opens some future work opportunities: applying the same solution to other problem domains(e.g., Unmanned Aerial Vehicle (UAV)'s MAS); investigating the effects of basic skills (e.g., walking, kicking) on the learning of the setplays policy; studies of other solutions to evaluate the dataset clustering or organizing the dataset; assess a solution to use DRL to select an individual setplay without using CBR; investigating a new set of non-default hyperparameters for DQN or other DRL algorithms to learn the policy; evolution of the BahiaRT Gym. The opportunities for future research are diverse.

[8] https://ssim.robocup.org/robocup-2022-soccer-simulation-3d-results/

References

Abreu, M., Lau, N., Sousa, A., and Reis, L. P. (2019). Learning low level skills from scratch for humanoid robot soccer using deep reinforcement learning. In *2019 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pages 1–8.

Abreu, M., Silva, T., Teixeira, H., Reis, L. P., and Lau, N. (2021). 6D Localization and Kicking for Humanoid Robotic Soccer. *Journal of Intelligent & Robotic Systems*, 102(2):30.

Bianchi, R. A., Santos, P. E., da Silva, I. J., Celiberto, L. A., and de Mantaras, R. L. (2018). Heuristically accelerated reinforcement learning by means of case-based reasoning and transfer learning. *Journal of Intelligent & Robotic Systems*, 91(2):301–312.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). *OpenAI Gym*. arXiv.org. _eprint: arXiv:1606.01540.

Cravo, J., Almeida, F., Abreu, P. H., Reis, L. P., Lau, N., and Mota, L. (2014). Strategy planner: Graphical definition of soccer set-plays. *Data & Knowledge Engineering*, 94:110–131.

Eustáquio, F., Camargo, H., Rezende, S., and Nogueira, T. (2018). On Fuzzy Cluster Validity Indexes for High Dimensional Feature Space. In Kacprzyk, J., Szmidt, E., Zadrożny, S., Atanassov, K. T., and Krawczak, M., editors, *Advances in Fuzzy Logic and Technology 2017*, Advances in Intelligent Systems and Computing, pages 12–23. Springer International Publishing.

Fabro, J. A., Reis, L. P., and Lau, N. (2014). Using Reinforcement Learning Techniques to Select the Best Action in Setplays with Multiple Possibilities in Robocup Soccer Simulation Teams. In *2014 Joint Conference on Robotics: SBR-LARS Robotics Symposium and Robocontrol*, pages 85–90, Sao Carlos, Sao Paulo, Brazil. IEEE.

Freelan, D., Wicke, D., Sullivan, K., and Luke, S. (2014). Towards Rapid Multi-robot Learning from Demonstration at the RoboCup Competition. In *RoboCup 2014: Robot World Cup XVIII*, Lecture Notes in Computer Science, pages 369–382. Springer, Cham.

Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I., Osawai, E., and Matsubara, H. (1998). RoboCup: A challenge problem for AI and robotics. In Kitano, H., editor, *RoboCup-97: Robot soccer world cup I*, pages 1–19, Berlin, Heidelberg. Springer Berlin Heidelberg.

Melo, L. C., Melo, D. C., and Maximo, M. R. O. A. (2021). Learning Humanoid Robot Running Motions with Symmetry Incentive through Proximal Policy Optimization. *Journal of Intelligent & Robotic Systems*, 102(3):54.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Moradi, M., Ardestani, M. A., and Moradi, M. (2016). Learning decision making for Soccer Robots: A crowdsourcing-based approach. In *2016 Artificial Intelligence and Robotics (IRANOPEN)*, pages 25–29.

Mota, L., Lau, N., and Reis, L. P. (2010). Co-ordination in RoboCup's 2D simulation league: Setplays as flexible, multi-robot plans. In *2010 IEEE Conference on Robotics, Automation and Mechatronics*, pages 362–367.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. (2021). Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research*. Publisher: MIT Press.

Reis, L. P., Lopes, R., Mota, L., and Lau, N. (2010). Playmaker: Graphical definition of formations and setplays. In *5th Iberian Conference on Information Systems and Technologies*, pages 1–6.

Russell, S. J. and Norvig, P. (2021). *Artificial intelligence: a modern approach*. Pearson Series in Artificial Intelligence. Pearson, Hoboken, NJ, fourth edition edition.

Shi, H., Lin, Z., Hwang, K., Yang, S., and Chen, J. (2018). An Adaptive Strategy Selection Method With Reinforcement Learning for Robotic Soccer Games. *IEEE Access*, 6:8376–8386.

Simoes, M. A. C., Nobre, J., Sousa, G., Souza, C., Silva, R. M., Campos, J., Souza, J. R., and Nogueira, T. (2020). Strategy Planner: Enhancements to support better defense and pass strategies within an LfD approach. In *2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pages 46–52, Ponta Delgada, Portugal. IEEE. tex.ids= Simoes_2020.

Simões, M. A. C. (2022). *Aprendizagem por Demonstração de Planos Coordenados em Sistemas Multiagentes*. Ph.D., Universidade Federal da Bahia.

Simões, M. A. C., da Silva, R. M., and Nogueira, T. (2020). A Dataset Schema for Cooperative Learning from Demonstration in Multi-robot Systems. *Journal of Intelligent & Robotic Systems*, 99(3-4):589–608. tex.ids= Simoes_2019 publisher: Springer Science and Business Media LLC.

Simões, M. A. C., Nobre, J., Sousa, G., Souza, C., Silva, R. M., Campos, J., Souza, J. R., and Nogueira, T. (2021). Generating a dataset for learning setplays from demonstration. *SN Applied Sciences*, 3(6):608. tex.ids= Simoes_2021.

Simões, M. A. C. and Nogueira, T. (2018). Towards setplays learning in a multiagent robotic soccer team. In *2018 latin american robotic symposium, 2018 brazilian symposium on robotics (SBR) and 2018 workshop on robotics in education (WRE)*, pages 277–282. tex.ids= Simoes_2018 tex.copyright: All rights reserved.

Simões, M. A., Mascarenhas, G., Fonseca, R., dos Santos, V. M., Mascarenhas, F., and Nogueira, T. (2022). BahiaRT Setplays Collecting Toolkit and BahiaRT Gym. *Software Impacts*, 14:100401.

Spitznagel, M., Weiler, D., and Dorer, K. (2021). Deep Reinforcement Multi-Directional Kick-Learning of a Simulated Robot with Toes. In *2021 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pages 104–110.

Teixeira, H., Silva, T., Abreu, M., and Reis, L. P. (2020). Humanoid Robot Kick in Motion Ability for Playing Robotic Soccer. In *2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pages 34–39, Ponta Delgada, Portugal. IEEE.

Wangenheim, C. G. v. and Wangenheim, A. v. (2003). *Raciocínio baseado em casos*. Manole, Barueri. OCLC: 69935690.

Wooldridge, M. (2009). *An Introduction to Multiagent Systems*. Wiley, Chichester, UK, 2 edition.