Transfer Learning of Air Combat Behavior

Armon Toubman,
Jan Joris Roessingh
Department of Training, Simulation,
and Operator Performance
National Aerospace Laboratory NLR
Amsterdam, Netherlands
{Armon.Toubman,
Jan.Joris.Roessingh}@nlr.nl

Pieter Spronck
Tilburg center for Cognition
and Communication
Tilburg University
Tilburg, Netherlands
p.spronck@gmail.com

Aske Plaat,
Jaap van den Herik
Leiden Institute of Advanced
Computer Science
Leiden University
Leiden, Netherlands
{aske.plaat,
jaapvandenherik}@gmail.com

***Abstract*—Machine learning techniques can help to automatically generate behavior for computer generated forces inhabiting air combat training simulations. However, as the complexity of scenarios increases, so does the time to learn optimal behavior. Transfer learning has the potential to significantly shorten the learning time between domains that are sufficiently similar. In this paper, we transfer air combat agents with experience fighting in 2-versus-1 scenarios to various 2-versus-2 scenarios. The performance of the transferred agents is compared to that of agents that learn from scratch in the 2v2 scenarios. The experiments show that the experience gained in the 2v1 scenarios is very beneficial in the plain 2v2 scenarios, where further learning is minimal. In difficult 2v2 scenarios transfer also occurs, and further learning ensues. The results pave the way for fast generation of behavior rules for air combat agents for new, complex scenarios using existing behavior models.**

Keywords-reinforcement learning; transfer learning; air combat; training simulations; computer generated forces

I. INTRODUCTION

Training air combat maneuvers in simulations is less expensive, safer and more flexible than in live training. The computer generated forces (CGFs) inhabiting these simulations, e.g. in the enemy role, require good behavior models for optimal training efficacy. Traditionally, the behavior of the CGFs is scripted. However, writing scripts is costly, as it requires time and domain expertise. Also, the resulting scripts are rigid and tailored to a specific CGF in a specific scenario. Machine learning techniques may offer more efficient, adaptive, and effective solutions to these problems.

Earlier work investigated the use of reinforcement learning methods [1] to generate air combat behavior (see, e.g., [2, 3, 4]). Toubman et al. investigated air combat behavior in 2v1 scenarios, in which a team of two agents learned to fight a single, statically scripted enemy. Realistic training scenarios involve more aircraft, and therefore we would like to move to more complex scenarios. For efficiency, we aim at reusing the experience that agents have gained in previous scenarios. Therefore, we turn to applying transfer learning methods [5, 6].

Transfer learning is a range of techniques for reusing experience gained in one scenario in another. They may provide a solution to scaling the CGFs from straightforward to more difficult scenarios. Such scenarios frequently share a common task, such as 'eliminate all enemies'. However, the circumstances are different in each scenario: a different number of enemies, the enemies use different tactics, etc.

In this paper, we apply transfer learning to air combat behavior. We transfer agents that have learned to defeat single enemies in air combat (the *source task*) to a scenario that contains two enemies (the *target task*). The single enemy uses a mix of three tactics, so that the learning agents have to come up with a generalized counter-tactic. The learning agents are transferred to scenarios in which the two enemies use tactics based on the tactics of the original single enemy. Furthermore, the learning agents are also transferred to a scenario with two enemies using a new, unseen tactic, in which the enemies specifically collaborate. We find that the agents with 2v1 experience transfer successfully to the 2v2 scenarios.

The main contribution of this paper is a demonstration of a transfer of reinforcement learning agents between a 2v1 and a 2v2 air combat scenario. To the best of our knowledge this has not been reported before.

The paper is structured as follows: Section II gives an overview of related work. Section III describes our transfer learning method, and section IV describes the experiment we used to test our technique. In Section V we show the results of this experiment. The paper is completed by a discussion of the results in Section VI. Some concluding remarks follow in Section VII.

II. RELATED WORK

Transfer learning methods have been successfully applied in classification, regression and clustering tasks [6]. In these tasks, due to model availability and the time it takes to train new models, it can be desirable to reuse old models on new data. However, when the new data has different features or a different distribution, the models will have to be adapted. In these cases, the expertise stored in the old models should be reused as efficiently as possible. The field of transfer learning concerns itself with studying effective ways for this reuse of knowledge.

Transfer learning methods have also been identified as a useful tool in reinforcement learning [5, 7]. Reinforcement learning is a technique through which agents learn by operating in some environment [1]. Through feedback from the

environment, the agents are rewarded for some of their actions, and punished for other actions. Repeated trial and error leads to the generation of some optimal policy. See [8] and [9] for some recent reinforcement learning application examples of transfer learning.

In this paper, we are interested in applying transfer learning methods to reinforcement learning agents learning air combat behavior. Learning air combat behavior is a non-trivial task, as air combat involves multiple agents, team behavior, and limited resources. Previous attempts at generating air combat behavior using machine learning have included learning classifier systems [10], behavior mining [11], and neuro-evolution [12]. However, such techniques create opaque behavior models that are hard to review after the learning process has completed. Therefore, Toubman et al. [2, 3, 4] applied a reinforcement learning technique that uses behavior rules that are themselves left unchanged during the learning process.

Transfer learning has been applied to the generation of behavior for other types of CGFs, based on various types of machine learning techniques. For example, Gorski and Laird [13] applied three different transfer learning methods to agents in an urban combat environment. The three methods all showed an improved task completion time, except for one case, which was attributed to the method's inability to scale to more complex tasks. As another example, Sharma et al. [14] applied transfer learning to an agent playing real-time strategy games, using a combination of case-based reasoning and reinforcement learning. They report finding equal or higher performance using their transfer method in complex scenarios, compared to the agent playing without transferred experience.

Several studies have been performed on how well experience gained by human trainees in air combat simulations transfers to real world settings (see, e.g., [15] for a comprehensive literature review). However, to the best of our knowledge, this is the first study on a transfer of the entities inside air combat simulations.

Here, we build on the work presented in [2, 3, 4] which used Dynamic Scripting (DS) as the reinforcement learning method of choice. DS [16] iteratively recombines behavior rules from some rule base into scripts (policies) which are executed by some agent. The rules in the pre-existing rule base have a certain weight, which constitutes each rule's probability of being selected for a newly generated script. Feedback from the agent's environment is used to update the weights of rules that were used in the environment, thereby altering their probabilities to be selected again. A particular advantage of DS is a high learning speed [16].

DS easily facilitates transfer learning as it stores knowledge in the form of weights that are assigned to rules. The simplest form of knowledge transfer using DS is therefore to reuse a rule base that was used for some source task (the original task), including the learned weights, for a new target task (the new task to which the learning agents are transferred).

III. Transfer Learning of Air Combat Behavior

In this paper we are interested in transferring air combat agents that learn using the DS method. As mentioned in the previous section, each agent has a rule base with behavior rules, which have associated weights. In essence, DS stores the experience gained by the agents in these weights. Transferring the agents therefore means copying the rule bases, including the weights, and assigning them to the same agents.

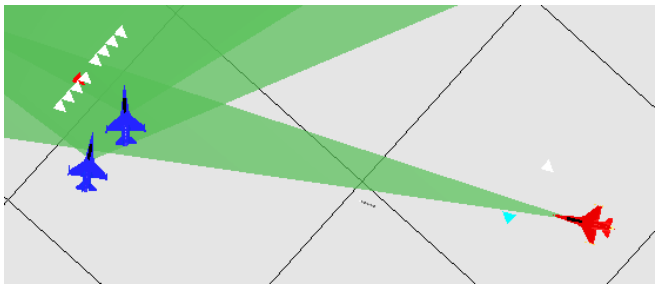

Figure 1. Screenshot of the simulation.

The goal of the transfer described in this paper is more effective and efficient learning of behavior for the target task, i.e., 2v2 air combat encounters. The benefit of this transfer should be faster development time of CGF behavior for complex scenarios, by reusing behavior models developed for different, simpler scenarios.

Taylor and Stone [5] enumerate five metrics that can be used to measure the success of a knowledge transfer. Below, they are listed with a brief description.

The first measure, *jumpstart*, shows the difference in performance on the first trial, with and without transfer.

The second measure, *asymptotic performance*, shows the difference in final performance with and without transfer, after the learning phase is over.

The third measure, *total reward*, shows the difference in accumulated reward during learning, with and without transfer.

The fourth measure, *transfer ratio*, shows the ratio of the reward accumulated by the agents using transfer learning, to the reward accumulated by the agents without transfer.

The fifth measure, *time to threshold*, shows the difference in the amount of trials needed to reach a certain level of performance, with and without transfer.

None of these metrics provides conclusive results by itself. For example, the jumpstart measure only shows how well the transferred agents do on the target task, without describing the effect of any learning. A second example is in the asymptotic performance and the total reward which both depend on the amount of time the agents spend learning.

For the work described here, we will use the five metrics mentioned above to measure the effects of the transfer method. Their application will be further discussed in Section IV.

IV. Method

The transfer method was tested in an air combat simulation. In the first simulation, one red agent performs a Combat Air Patrol (CAP) in a section of airspace. Two blue agents (the learning agents) enter this airspace with the goal of defeating the red agent. Red uses various tactics, meaning the blues will have to come up with a generalized solution. This scenario constitutes the *source task* for the blues. The blues learn until

their performance asymptote is reached, which was determined in [4] to be after 100 trials. Figure 1 shows a screenshot of the simulation.

Then, we add the second red agent. The red fighters now use one of four different tactics (see Section IV-B). This scenario contains the *target tasks* for the blues.

The blues and the reds are further described in Sections IV-A and IV-B respectively. Section IV-C presents the learning parameters that are used in the experiments. Section IV-D describes the transfer method that is used. Finally, the analysis of the results from the experiments is described in Section IV-E.

A. CGFs

As in [4], the fighter jets in the simulation are based on the F-16, and are equipped with radar, a radar warning receiver (a device that detects incoming radar signals), and four missiles based on the AIM-120B AMRAAM. Each jet is controlled by an agent. The behaviour of each agent is governed by scripts. For the agents in the blue team, the scripts are generated using DS (see Section IV-B). The agents in the red team have a script per tactic that they use (see Section IV-C).

The behavior rules in each script are if-then rules which map observations to actions. Each time step, a matching rule is selected and executed. Examples of the rules are "if I see an enemy on my radar, and this enemy is within 80 kilometers of me, and I have missiles left, fire a missile at this enemy" and "if I detect an incoming missile, turn right 180 degrees".

B. Blue team

The blue team consists of two agents, a 'lead' and a 'wingman', controlling simulated F-16 fighter jets. The blues learn using the DS technique, as explained earlier. The blues communicate and are able to coordinate their actions through the use of a decentralized coordination scheme based on [2].

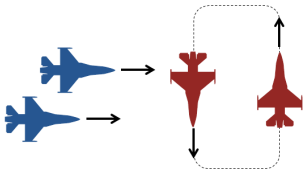

Figure 2. The reds (right) fly a CAP while the blues approach (left).

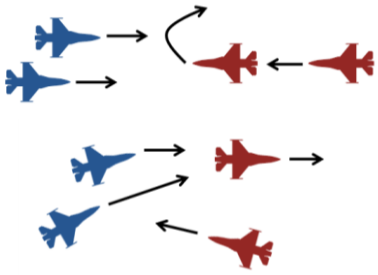

Figure 3. The lead-trail tactic. The red lead draws away the blues (above), giving the red wingman an opportunity to fire (below).

At the start of each trial, the blues are positioned so that they fly will into the portion of airspace where the red team is performing its CAP. The rule bases used by the blue agents are mostly identical, except for the inclusion of extra rules in the wingman's rule base that allow it to fly in formation with the blue lead.

C. Red team

Depending on the task, the red team consists of one or two agents, also controlling F-16 fighter jets.

In the target tasks, the red agents use one of four tactics:

- The **default** tactic: the red lead performs a CAP and engages the blues upon detection. The red wingman flies the same CAP, lagging half a pattern behind the lead (see Fig. 2);
- The **evading** tactic: the same as the default tactic, but the reds actively try to evade incoming missiles;
- The **close range** tactic: the same as the default tactic, but the reds engage blue at a closer range;
- The **lead-trail** tactic. This is a well-known tactic used by two-ship formations. The red wingman flies directly behind the red lead, as they approach the blues. When the blues detect the red lead, the lead turns away. As the blues follow the red lead, the red wingman has the opportunity to directly fire at the blues (see Fig. 3).

In the source task, the red agents use a mixed tactic. In each encounter, the reds pick one of the tactics at random (except for the lead-trail tactic), and use that tactic until they lose an encounter. At that point, they again select one of the tactics at random.

D. Learning parameters

A learning episode consists of 150 consecutive trials. For each task, 150 learning episodes are performed.

The reward function used during learning was the probability-of-kill reward function described by [4]. In essence, this function rewards agents for fired missiles with a reward value proportional to the probability with which their missile would have hit (reducing the learning time compared to a reward only in the actual occurrence of a hit).

To make maximum use of the probability-of-kill (pK) reward function, all fighter jets were armed with four 'dummy' missiles that did not explode on impact. This made sure each agent had enough opportunities to fire missiles during a trial.

The winner of each trial is determined by evaluating the pK of each missile fired in that trial. A random value is generated and compared to the pK value, which determines whether that missile would have hit or not. In case of a hit, the team that fired that missile wins that trial.

DS requires an adjustment function to translate the reward that an agent collects to changes to the weights of its rules. The adjustment function used here is shown in Eq. 1.

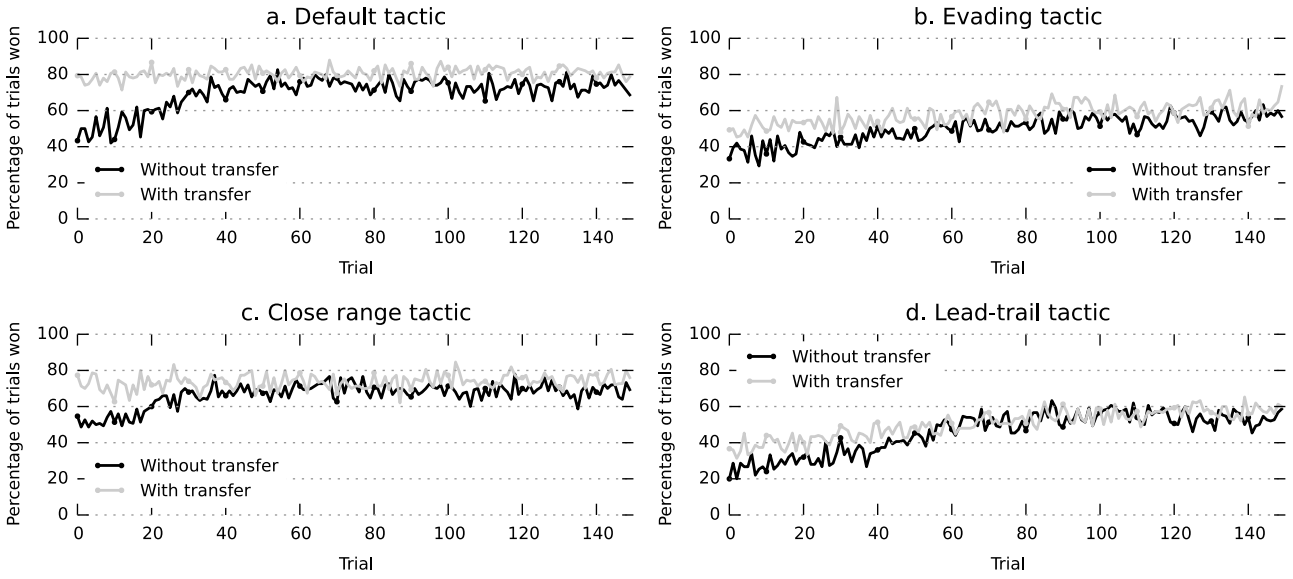

Figure 4. Percentages of trials won by the blue team against each of red's tactics, with and without knowledge transfer from the source task.

$$adjustment(\, r \,) = max(\, 50 * (\, 2r\text{-}1 \,), \text{-}25 \,) \qquad (1)$$

Equation (1) shows the calculation of the weight adjustments based on the agent's reward *r*. The maximum increase in weight is 50, while the maximum decrease is -25. This allows weights to climb more rapidly than they can fall, making it harder for rules to lose weight because of a chance failure.

E. Transfer between tasks

The two blue agents learn behavior in the source task until they reach their performance asymptote. At this point, the weights of the rules in the rule bases of the blue lead and wingman are optimized for the source task. Keeping the rules with these weights, the blues are presented the target task.

F. Analysis of results

We obtain 150 win/loss sequences per task by recording whether the blues win or lose each trial.

As mentioned earlier, Taylor and Stone [5] defined five metrics for determining the success of a transfer method. We apply these metrics to our performance results. However, instead of comparing the amount of rewards gathered by the agents, as is traditionally done in reinforcement learning research, we use the number of trials won by the blue agents as our performance scores.[1] For the purposes of our application, we are most interested in the number of trials won, as this number shows the performance as it would be in a live environment.

We apply the jumpstart metric through direct comparison of the performance of the blue agents in the first trial of the source and target tasks. For the asymptotic performance metric, we calculate the mean performance of the last 30 trials per task. The transfer ratio metric, applied to the trials won by blue with and without transfer, gives an impression of the difference in performance throughout the learning process. Finally, the time to threshold metric shows how long it takes the agents to reach a certain level of performance.

[1] The probability-of-kill reward function, together with the dummy missiles and the use of the probability-of-kill to determine the winner of a trial, creates a discrepancy between the collected rewards and winning/losing trials (even though the two are strongly related, see (Toubman et al., 2015).

V. RESULTS

The four target tasks were run twice, once with transfer of the blue agents and once without. For each target task, the blues were allowed to learn over 150 trials. Each task was repeated 150 times.

The results of applying the five performance metrics as described in sections 2 and 4 are shown in Fig. 4, 5, and 6, and Table 1.

Fig. 4 shows the initial performance on the tasks with and without transfer (the jumpstart metric). For each tactic, the transferred agents reached a higher initial performance than the agents that were not transferred.

Table 1 shows the asymptotic performance on the target tasks. Unpaired *t*-tests show that the blues reached significantly higher final performance with transfer, on all tasks.

Fig. 5 shows the total amount of trials won by the blues with and without transfer. Using the results shown in Fig. 5, the transfer ratios are determined as +15.3% for the default tactic, +16.1% for the evading tactic, +10.9% for the close range tactic, and -9.6% for the lead-trail tactic.

Fig. 6 shows the time to threshold, which was determined as the first trial at which the blues reached a performance higher than the mean performance of the final 30 trials.

The blues' performance on all four of the target tasks is shown in Fig. 7.

Barring the time to threshold and the total number of trials won in the case of the lead-trail tactic, we clearly see that transfer learning improves performance calculated with all five metrics, for all the tasks.

VI. Discussion

The results show clearly that a transfer of experience gained in earlier, easier 2v1 scenarios provides an advantage in more difficult 2v2 scenarios. This advantage is present from the start, as can be seen in Fig. 4 which shows a higher initial performance against each tactic. The advantage even has an effect on the final performance, as can be seen in Table 1. Judging from these results, earlier air combat experience is reusable in different scenarios, and provides a benefit over starting with learning in those other scenarios. The main finding of this research is that the blues reach a higher initial and final performance with transfer against the lead-trail tactic, a tactic they did not encounter in their source task.

The results in Fig. 5 show that the transferred agents are more effective throughout the entire learning process, as they reach a higher number of trials won against three out of the four tactics. As can be seen in Fig. 6, the blues reach a performance level equal to their final performance sooner with transfer than without transfer, except in the case of the lead-trail tactic.

Regarding the total number of trials won and the time to threshold metric, the transferred agents perform slightly worse against the lead-trail tactic than without transfer. Even though the transferred agents have a higher initial and final performance, it seems that the learning process advances slower when the agents try to adapt their prior knowledge. While the agents may have learned some rules on the source task that are also applicable extent on the target task to some, these rules may have been a local optima that the agents had to climb out of slowly to reach even better results.

It is interesting to see that learning appears to halt after transfer against the default and close range tactics, while against the evading and the lead-trail tactics, learning continues. This confirms the relative difficulty of the latter two tactics. Toubman et al. [4] reported the need for a relatively large number of trials until the performance asymptote was reached by the blues against the evading tactic of the reds, which seems to corroborate this finding.

Is there any benefit of the transfer for the default and close range tactics? From the first to the last trial, the performance remains largely stationary, with some random variation. While these changes in performance in performance are minimal, they are still higher than the performance without transfer. Also, the non-transferred agents take around 40 trials to reach their optimal performance. We therefore argue that even in these straightforward cases, better performance is obtained after transfer. However, the stationary performance rates on the default and close range tactic may also indicate that in the case of these tactics, the learning problem becomes easier when a second enemy is added. Simply put, the blues may have been able to collect more reward (i.e., fire missiles with a higher P_k) because of the addition of a second easy-to-hit target. Further research should point out whether this is in fact the case.

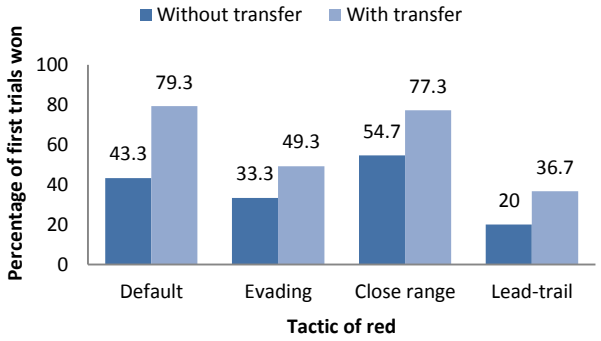

Figure 5. Mean initial performance per tactic. Higher is better.

Table I. Mean asymptotic performance over last 30 trials, in terms of percentage of trials won. Higher is better

	Without transfer		With transfer		Unpaired *t*-test
Tactic	**μ**	**σ**	**μ**	**σ**	***p***
Default	73.5	3.8	80.8	2.6	< .0001
Evading	56.3	3.6	61.8	4.8	< .0001
Close range	69.0	3.5	74.9	3.1	< .0001
Lead-trail	53.1	3.7	58.1	3.1	< .0001

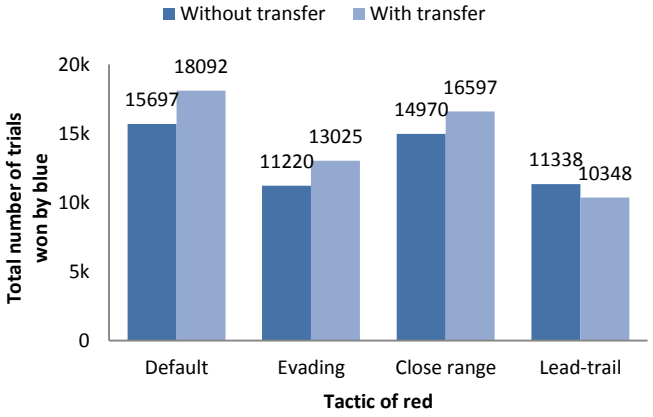

Figure 6. Total number of trials won by blue per tactic. Higher is better.

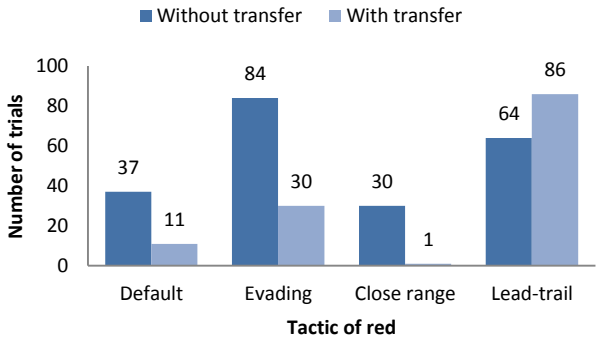

Figure 7. Time to threshold (mean performance > asymptotic performance). Lower is better.

The transfer method described in this paper assumes that all learning agents use the same rule bases across all tasks (see section II). However, this does not have to be the case. For example, it is possible to transfer rules with optimized weights into a new rule base containing rules with default starting weights. Even when the optimized rules are not needed in a new rule base as is, a custom mapping between rules with some similarity may be able to convert gained experience to a new domain.

VII. Conclusions

By transferring air combat agents from 2v1 scenarios to 2v2 scenarios, we find that in general the agents achieve higher performance than agents without prior experience. This increase in performance appears using five different performance metrics. In particular, the transferred agents reach a higher mean final performance on all of the target tasks. Against the most difficult tactic, the transfer learners reach higher performance from the onset, although they take longer to reach their maximum performance level than without transfer. In conclusion, the transfer method presented in this paper is beneficial to agents learning air combat behavior, although the size of the benefits appears to depend on the difficulty of the presented tasks.

Further research may include extending the transfer method used in this paper to asymmetric rule bases, to see if more general transfer of training is possible with the rule-based approach. Additionally, it would be interesting to apply the same transfer method applied to different domains, which may involve different numbers and types of rules.

Acknowledgment

The authors thank Lt Col Roel Rijken (Royal Netherlands Air Force) for the first version of the simulation environment and advice regarding air combat tactics, and Pieter Huibers and Xander Wilcke for further work on the simulation. The work in this paper was performed in the context of the agreement between NLR and LCDS.

References

[1] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, Cambridge, MA, USA: MIT Press, 1998.

[2] A. Toubman, J. J. Roessingh, P. Spronck, A. Plaat and J. van den Herik, "Dynamic Scripting with Team Coordination in Air Combat Simulation," in *27th Int. Conf. Ind. Eng. and Other Applicat. of Appl. Intell. Syst.*, Kaohsiung, Taiwan, 2014.

[3] A. Toubman, J. J. Roessingh, P. Spronck, A. Plaat and J. van den Herik, "Centralized Versus Decentralized Team Coordination Using Dynamic Scripting," in *Proc. 28th Eur. Simulation and Modelling Conf.*, Porto, Portugal, 2014.

[4] A. Toubman, J. J. Roessingh, P. Spronck, A. Plaat and J. van den Herik, "Rewarding Air Combat Behavior in Training Simulations," in *IEEE Int. Conf. Syst., Man and Cybern.*, Hong Kong, China, 2015, to be published.

[5] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *J. of Mach. Learning Res.*, vol. 10, no. 1, pp. 1633-1685, 2009.

[6] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, pp. 1345-1359, Oct. 2010.

[7] A. Lazaric, "Transfer in Reinforcement Learning: A Framework and a Survey," in *Reinforcement Learning*, 2012, pp. 143-173.

[8] T. Takano, H. Takase, H. Kawanaka and S. Tsuruoka, "Transfer Method for Reinforcement Learning in Same Transition Model -- Quick Approach and Preferential Exploration," in *10th Int. Conf. Mach. Learning and Applicat.*, Honolulu, HI, 2011.

[9] S. Proper and P. Tadepalli, "Multiagent Transfer Learning via Assignment-Based Decomposition," in *8th Int. Conf. Mach. Learning and Applicat.*, Miami Beach, FL, 2009.

[10] R. E. Smith, B. A. Dike, R. K. Mehra, B. Ravichandran and A. El-Fallah, "Classifier systems in combat: two-sided learning of maneuvers for advanced fighter aircraft," *Compu. Methods in Appl. Mechanics and Eng.*, vol. 186, no. 2-4, pp. 421-437, 2000.

[11] Y. Yin, G. Gong and L. Han, "Experimental study on fighters behaviors mining," *Expert Syst. with Applicat.*, vol. 38, no. 5, pp. 5737-5747, 2011.

[12] R. Koopmanschap, M. Hoogendoorn and J. J. Roessingh, "Tailoring a cognitive model for situation awareness using machine learning," *Appl. Intell.*, vol. 42, no. 1, pp. 36-48, 2015.

[13] N. A. Gorski and J. E. Laird, "Experiments in Transfer Across Multiple Learning Mechanisms," in *Int. Conf. Mach. Learning, Workshop Structural Knowledge Transfer Mach. Learning*, Pittsburgh, PA, 2006.

[14] M. Sharma, M. Holmes, J. Santamaria, A. Irani, C. Isbell and A. Ram, "Transfer Learning in Real-Time Strategy Games Using Hybrid CBR/RL," in *20th Int. Joint Conf. Artificial Int.*, Hyderabad, India, 2007.

[15] J. Borgvall, M. Castor, S. Nählinder, P.-A. Oskarsson and E. Svensson, "Transfer of Training in Military Aviation," FOI, Linköping, Sweden, FOI-R--2378-SE, Dec. 2007.

[16] P. Spronck, M. Ponsen, I. Sprinkhuizen-Kuyper and E. Postma, "Adaptive game AI with dynamic scripting," *Mach. Learning*, vol. 63, no. 3, pp. 217-248, 2006.