On-Demand Skills Training to Support Regular Continuation Training for Fighter Pilots

Armon Toubman, Jelke van der Pal, Jur Crijnen
Royal Netherlands Aerospace Centre NLR
THE NETHERLANDS

armon.toubman@nlr.nl, pal@nlr.nl, jur.crijnen@nlr.nl

ABSTRACT

On-demand training relieves acute training needs, such as the operation of new systems or the defence against a new threat. It may also fulfil personal needs for refreshing existing skills. The current training regime is still based on a yearly schedule which may result in over-trained skills in some areas and considerable skill decay in other areas. Individual pilots differ considerably in their needs for refreshing their skills, but the amount of training exposure usually only differs between experienced and inexperienced pilots, the latter receiving more training. In this paper we present an approach for predicting a specific personal demand for training on specific skills. Depending on the nature and complexity of these skills, the personal demand may range from pure academics to part-task simulation to team training on realistic simulators of live systems. This implies advanced data logging from the training systems as well as advanced tooling for predicting and scheduling the personal training need. The approach will be tested by training former F-16 pilots with PC-based simulation according to their personal needs.

1.0 INTRODUCTION

On-demand training has been a dream of trainers and trainees alike for decades. Training needs coming from changes in the operational systems, procedures, doctrines, or from personal situations such as illness, leaves, deployments, are all often not satisfied at the most optimal timing or to the most optimal level of proficiency. Furthermore, the training available often does not meet personal needs. Timing, amount, and approach of required training may differ considerably between people, even in carefully selected professions such as fighter pilots. Personalised training concepts promise to achieve higher proficiency levels, better retention of skills, lower training costs and fostering higher job satisfaction which may in turn improve job retention. Obviously, personalised training is not achieved easily; it requires a more elaborate and transparent understanding of 1) the task demands, 2) the required competencies and proficiencies, 3) learner and performance models that recognise personal approaches and abilities, 4) an instructional model that selects the appropriate type of task and level of difficulty, 5) a range of training systems that are readily available, and 6) an elaborate data system to monitor and predict performance.

In this paper, we focus on personalisation of the refresher training for fighter pilots. Our basis is a performance-based training concept, building upon a competency-based training framework (for a more detailed description see [1]). Our first aim is to develop a predictive retention model that works for individual or small groups of pilots, which will be crucial for small air forces such as the Royal Netherlands Air Force. Retention is a person's ability to remain proficient over time while not practising the task. While the neurological mechanism of the decay process of knowledge and skills is not fully understood, there is an abundance of findings on the factors that influence this process [2]. The amount, quality, and spacing of training and practice, the type of task to perform, personal factors such as engagement during training, selfefficacy, and cognitive capabilities have all been found to affect the retention of skills. While these factors may describe the skill decay effects on a group level, predicting individual performance has not been achieved yet. This may be impossible if not taking into account the situational factors as well, e.g., task conditions (such as noise) and personal conditions (such as motivation or fatigue). Nevertheless, promising results with predictive performance models have been reported [3], [4].

Once predictive performance models are sufficiently robust, the next challenge is to determine the training required to restore the proficiency gap with minimal training time, costs and means. While combat readiness obviously relates to proficiency on the full mission given a certain range of operational conditions, this does not imply that the most effective approach to remain proficient is to focus on practicing the whole task in ever more complex simulation or LVC set ups. Especially experienced pilots could rely on a minimum of exposure to the full mission, while practicing specific part tasks on a range of simulators (from laptop simulation to Full Mission Simulation in a small network) and refreshing knowledge and procedures on a simple interactive tablet app. It may be the case that the complex competency related to the full mission is fairly robust and fades very slowly, except for a number of specific procedures and knowledge items which tends to decay more rapidly. The ultimate retention model needs to recognise the composition of competencies, from collective skills and high level non-technical competencies to very specific knowledge items and technical skills. A personalised continuation training plan may result from the full overview of retention processes, recognising a transfer of retention effect from competencies in one task (full mission or part task) to another. Obviously, the personal needs should cover the need to fly live under certain conditions. Balancing training media (live versus sim in all their versions) will be a key feature in a personal training plan.

A first step towards this ultimate personalised training plan is advancing the predictive power of retention models. The next sections outline the challenges to this goal and ongoing work. In section 4 we present our effort to develop a personal predictive retention model and how we intent to validate it.

2.0 MODELING RETENTION

The goal of modelling retention is to predict, for an individual, the amount of skill that is retained after a certain period of non-use. While the act of modelling retention is easily defined, the actual modelling is quite difficult. Here, we identify two main challenges.

The first challenge of modelling retention is the sparsity of measurements over time. For modelling the retention of complex skills, a complicating factor is that assessment of the skill requires the measurement of multiple variables. Each of the variables must be modelled. However, there may be complex interactions between the variables that must be modelled as well for accurate predictions. The second challenge is thus dealing with high-dimensional data. We discuss the two challenges below.

The concept of retention has an inherent notion of time. A retained skill is what remains of a previously acquired skill after some time period (i.e., the retention interval). We thus require a measurement of the skill before, and after the retention interval. Since skills decay gradually over time, it would be ideal to repeatedly measure the skill during the retention interval as well. However, since the execution of a task for a measurement also has a training effect, such measurements interfere with the measurement after the retention interval, which is the measurement we aim to model. Thus, the first challenge of modelling retention is inherently sparse data of a temporal nature.

For simple skills, such as remembering a predefined set of words, the retention of the skill is easily measured as the number of words that is remembered. It is our experience that as the complexity of the skill increases, it becomes (a) harder to define a successful application of the skill, and (b) of interest to determine which sub-skills (e.g., visual or aural detection in a cognitive task) are retained at a higher level than other sub-skills. Consequently, the second challenge of modelling the retention of complex skills is the high dimensionality of the data.

To illustrate, Table 1-1 shows example measurements of a skill for a particular person. During the skill acquisition phase, the skill is repeatedly measured, e.g., by means of tests. Next, during the retention interval, the skill remains unused, and therefore decays to some extent. During the retention interval, the skill cannot be measured. After the retention interval, the retained skill is measured by the retention test. Ideally, the retention model is able to predict the retained skill (indicated by the question marks), so that the skill level of the person will not drop below a certain point. Furthermore, re-training may be planned so that the skill level can be brought back to its maximum in the least amount of time.

Table 1-1. Example of skill measurements in subsequent phases.

Day	Phase	Variable 1	Variable 2	…	Variable n
1	Skill acquisition	0.2	0.3	…	0.5
2		0.4	0.4	…	0.7
3		0.6	0.8	…	0.9
…	Retention interval				
t	Retention test	?	?	…	?

3.0 RELATED WORK

The study of the retention of skills traces back to the 'forgetting curve' proposed by Ebbinghaus, which has maintained its scientific relevance over the years (see, e.g., [5]). Since the forgetting curve, advances in statistics and the introduction of computerized techniques such as machine learning have led to new methods by which retention can be modelled. The most relevant methods are found in the research areas of *time series forecasting* (section 3.1) and *knowledge tracing* (section 3.2). We discuss the two methods below.

3.1 Time series forecasting

Time series forecasting (TSF) is the prediction of future values, by modelling previously observed values. The autoregressive moving average (ARMA) family of models are classical statistical TSF tools. ARIMA models have been shown to work well on both one-step (i.e., one prediction) and multi-step (i.e., prediction upon prediction) forecasting on univariate datasets [6], [7]. Their performance on multivariate datasets has not yet conclusively been determined.

Machine learning techniques such as recurrent neural networks (RNNs), and the related long short-term memory (LSTM) and gated recurrent unit (GRU), are popular alternatives to regular statistical methods. A benefit of machine learning techniques over classical methods is that machine learning is able to learn a single, perhaps non-linear, model over multiple datasets. This is useful in the case of retention modelling, where a single individual will not yield enough data for a model. Furthermore, these techniques have been shown to be robust in cases of sparse multivariate time series data, including missing data [8]. Recently, a RNN augmented with an exponential smoothing method has won the M4 competition for TSF methods [9].

3.2 Knowledge tracing

The research area of knowledge tracing is concerned with the modelling of the changing knowledge state of students. With such a model, the student's performance on the next exercise or test can be predicted [10]. In its original form, a Bayesian model tracks the skills that the student has or has not yet acquired. The incorporation of neural networks into this model has led to deep knowledge tracing, which supposedly offers better prediction capabilities because it is able to automatically create richer representations of the student knowledge. It has been shown however, that with some extensions, the regular Bayesian knowledge tracing is able to make predictions on par with deep knowledge tracing [11].

A particular case of knowledge tracing is second language acquisition modelling (SLAM). SLAM was proposed by the makers of Duolingo, a language-learning app, to investigate the prediction of errors by second language learners [12]. Duolingo provided an extensive corpus of language learner data of 6400 learners on three languages, and organised a competition for error prediction. The winner of the competition was a combination of an RNN with a gradient boosted decision tree (GBDT). The authors of the method successfully combined the strengths of the two methods: the RNNs excel at sequential prediction tasks, whereas the GBDT can learn effectively on large, high-dimensional data sets [13].

4.0 OUR APPROACH

We are taking an iterative approach to designing a retention model, using learner data from multiple tasks. Although a single model that could make predictions for any learner on any task would be ideal, we believe this is not a realistic end goal in the near term because of the arduous task of data collection. Rather, we aim to design a model architecture that can be applied to particular tasks as data becomes available. We give an overview of our approach below. First, we describe the Space Fortress task (section 4.1). The Space Fortress task is a fairly complex task, on which we aim to collect retention data from many participants. Next, we describe the F-16 Mission Simulation task (section 4.2). This is very complex task that we will present to a select group of subject matter experts. We use the data and experience gathered on the Space Fortress task to build a somewhat generic model, and then apply the model to the data gathered on the F-16 Mission Simulation task to tune and validate the model for this particular case.

4.1 The Space Fortress task

Space Fortress (SF) is a video game that was specifically developed for the study of the acquisition of complex skills [14]. The game contains many different game elements that call for cognitive and motor skills, and the handling of procedures. The combination of the required skills makes SF a fitting task for retention research, as each of the skills may decay in their own manner. For data collection, we built a freely accessible online adaptive instructional system (see: https://spacefortress.nlr.nl), which first teaches participants the game, and then imposes a retention interval (forced no-play). After the interval, the participant is invited back to perform a retention test. The result of the test is used to select a brief re-training schedule, after which the participant enters a new retention interval. An in-depth description of SF and the adaptive instructional system are available in [15].

At the time of writing, we do not yet have enough data to build and validate an end-to-end model of retention. To work around this, we have begun to incorporate specific assumptions about the data into our research. The assumptions are based on the available skill decay literature [3], [16] and on our own experience working with students through various training methods. The four assumptions are:

1. Skill decay rate increases with age.

2. Skill decay rate decreases with prior experience (e.g., gaming experience in the case of SF).

3. Skill decay increases with the length of the retention interval.

4. Skill decay differs among the sub-skills (e.g., accuracy decays faster than task completion speed).

Starting with the distribution of the data that we obtained from the SF task, we use these assumptions to generate new data points by gaussian processes, and add noise to reduce the homogeneity of the new data. The result is a mainly synthetic data set by which we can start our modelling efforts. As new data comes in, we can (a) formulate new assumptions based on our observations, or (b) reduce the amount of synthetic data.

Using regression techniques, we determine whether some relationship can be found between the performance of participants up to the retention interval, the retention interval length, and the performance on the retention test. So far, the best performing regression technique has been XGBoost [17], although we cannot draw strong conclusions at this point. With more data, regression may indicate which factors contribute the most to the measured decay of complex skills.

Our near-term goal is developing an RNN model that will successfully train and make predictions on each of the two tasks. Here, data pre-processing and feature selection become important considerations. Each task comes with its own performance measures, such as scores with varying ranges (e.g., one to ten for test scores, one to five for questionnaires, minus infinity to plus infinity for video games) and physiological measures (e.g., heart rate variability, saccades and fixations). Furthermore, personal information such as age, sex, education level and work experience may not always be available in the same quantities across tasks. Through our modelling efforts, we aim not only to build a predictive model, but also to advance theory building on how each of these measures should be treated, and to what extent they are important to the prediction of retention.

4.2 The F-16 Mission Simulation task

Due to the abundance of data generated by modern flight training facilities, a more thorough and objective analysis of pilots' performance has become possible. As a result, the simulated scenarios reflect the skills of the fighter pilots from several perspectives, such as tactical, procedural, and basic flying.

4.2.1 Setup

The F-16 mission simulation consists of multiple relatively short scenarios (5-10 minutes) which we classify as low-, medium-, and high intensity, depending on the specified assignment. We regulate the intensity of the task and the workload of the fighter pilot by adjusting the speed at which events take place and the number of virtual adverse entities during the combat operation. The experiment is performed in the F-16 PC-based *Softpit* Simulator which is a facility of the Royal Netherlands Aerospace Centre.

The participants involved in the experiment are former Royal Netherlands Air Force F-16 pilots. As such, in contrast to operational pilots, there will be no interference in skill development caused by operational flights. Throughout the experiments we focus on the personal training needs for each pilot instead of a generic and fixed curriculum. Because we are aiming to develop a pilot specific model, also task-unrelated variables might hold information with regards to the prediction of performance. For example, flight experience and weapon instructor capabilities are expected to be of importance, as experience relevant to the task is known to reduce skill decay both in literature [16] as in practise. More influencing factors added to the model may improve the individual predicting capabilities.

4.2.2 Assignments

For all scenarios a division in assignments is defined to create an exercise- and competence specific training model. The fighter pilot has to execute multiple intercept manoeuvres to successfully fly the scenarios. This is a relatively common, but also complex operational procedure, which requires situational awareness to oversee the airspace and anticipation of the actions of the opponents.

However, all five scenarios are flown as single ship missions, which is not realistic but sufficient for the purpose of or study and not harmful to non-active pilots. The level of intensity depends on the number of opponents involved and the speed by which the opponents attack.

4.2.3 Performance Measurements

For each assignment the pilot requires skills which are rated according to a specified set of performance metrics, depending on the complexity of the assignment. Whereas basic flying is measured through simple telemetric indicators (e.g., deviation from the expected flightpath), tactical skills require more specific performance indicators (e.g., relative position to the adversary). To retrieve these performance measurements the Performance Effectiveness Tracking System (PETS) is used. This software has been developed by the Air Force Research Laboratory (AFRL) since 2001 and provides real-time performance assessment for multiple entities within a simulation and assesses fighter pilots' skills in relation to Mission Essential Competences (MEC) [18], [19]. The capability of PETS to quantify human performance makes it suitable for this research. The metrics will be adjusted to meet the specific requirement of the training.

5.0 DISCUSSION

Developing a predictive performance model based on a new data set has the advantage that you control the measures that you expect will be needed to generate a robust model. The drawback is that data may not come as smooth and fast as expected. The Space Fortress online game has generated much less data than expected and a considerably longer time will be needed before a sizeable dataset containing retention data over an extensive period of time will be gathered. At time of writing, we therefore expanded the current data set with augmented data using gaussian processes while injecting certain skill fading assumptions regarding age, previous experience, and performance accuracy vs speed. Initial results revealed the best fitting techniques to model this 'known' dataset. Although promising, it remains to be seen whether these findings will be useful and valid under more realistic and complex skillsets and test conditions. More conclusive results will therefore have to wait for a more mature retention model that has been validated in the more realistic F-16 training set up.

REFERENCES

[1] E. L. Fjærbu, G. K. Svendsen, and J. Van der Pal, “Feasibility of Performance-Based Training Programs for Combat Aircraft Pilots,” presented at the MSG-177 Symposium on “Towards On-Demand Personalized Training and Decision Support,” 2020.

[2] W. Arthur Jr, W. Bennett Jr, P. L. Stanush, and T. L. McNelly, “Factors That Influence Skill Decay and Retention: A Quantitative Review and Analysis,” Human Performance, vol. 11, no. 1, pp. 57–101, Mar. 1998, doi: 10.1207/s15327043hup1101_3.

[3] F. Sense, T. S. Jastrzembski, M. C. Mozer, M. Krusmark, and H. van Rijn, “Perspectives on Computational Models of Learning and Forgetting,” in Proceedings of the ICCM 2019: 17th International Conference on Cognitive Modeling, 2019, pp. 216–221.

[4] T. S. Jastrzembski et al., “Personalizing Training to Acquire and Sustain Competence Through Use of a Cognitive Model,” in Augmented Cognition. Enhancing Cognition and Behavior in Complex Human Environments, Cham, 2017, pp. 148–161, doi: 10.1007/978-3-319-58625-0_10.

[5] J. M. J. Murre and J. Dros, “Replication and Analysis of Ebbinghaus’ Forgetting Curve,” PLOS ONE, vol. 10, no. 7, p. e0120644, Jul. 2015, doi: 10.1371/journal.pone.0120644.

[6] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “Statistical and Machine Learning forecasting methods: Concerns and ways forward,” PLOS ONE, vol. 13, no. 3, p. e0194889, Mar. 2018, doi: 10.1371/journal.pone.0194889.

[7] G. Papacharalampous, H. Tyralis, and D. Koutsoyiannis, “Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes,” p. 45, 2019.

[8] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, “Recurrent Neural Networks for Multivariate Time Series with Missing Values,” Sci Rep, vol. 8, no. 1, pp. 1–12, Apr. 2018, doi: 10.1038/s41598-018-24271-9.

[9] S. Smyl, “A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting,” International Journal of Forecasting, vol. 36, no. 1, pp. 75–85, Jan. 2020, doi: 10.1016/j.ijforecast.2019.03.017.

[10] T. S. Jastrzembski, K. A. Gluck, and G. Gunzelmann, “Knowledge Tracing and Prediction of Future Trainee Performance,” Air Force Research Laboratory, AFRL-HE-AZ-TP-2007-0001, 2006.

[11] M. Khajah, R. V. Lindsey, and M. C. Mozer, “How deep is knowledge tracing?,” arXiv:1604.02416 [cs], Jun. 2016, Accessed: Aug. 28, 2020. [Online]. Available: http://arxiv.org/abs/1604.02416.

[12] B. Settles, C. Brust, E. Gustafson, M. Hagiwara, and N. Madnani, “Second Language Acquisition Modeling,” in Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, New Orleans, Louisiana, 2018, pp. 56–65, doi: 10.18653/v1/W18-0506.

[13] G. Ke et al., “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 3146–3154.

[14] A. Mané and E. Donchin, “The space fortress game,” Acta Psychologica, vol. 71, no. 1–3, pp. 17–22, Aug. 1989, doi: 10.1016/0001-6918(89)90003-6.

[15] J. van der Pal and A. Toubman, “An Adaptive Instructional System for the Retention of Complex Skills,” in Adaptive Instructional Systems, Cham, 2020, pp. 411–421, doi: 10.1007/978-3-030-50788-6_30.

[16] J. I. D. Vlasblom, H. J. M. Pennings, J. van der Pal, and E. A. P. B. Oprins, “Competence retention in safety-critical professions: A systematic literature review,” Educational Research Review, vol. 30, p. 100330, Jun. 2020, doi: 10.1016/j.edurev.2020.100330.

[17] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[18] E. A. Watz, B. T. Schreiber, L. Keck, J. M. McCall, and W. Bennett Jr, “Performance Measurement Challenges in Distributed Mission Operations Environments,” Simulation Technology Newsletter, no.

03F-SIW-022, 2003.

[19] E. A. Watz, L. Keck, and B. T. Schreiber, "Using PETS software to capture complex objective measurement data from Distributed Mission Operations (DMO) environments," Proceedings of the 2004 Spring Simulation Interoperability Workshop, no. 04S-SIW-143, 2004.