Optimizing a Scaffold to Guide Motor Skill Learning

Dennis Heitkamp,[1] **Kathrin Krieger,**[1] **Jason Friedman,**[2] **Alexandra Moringen** [1]

[1]Citec - Bielefeld University, Inspiration 1, 33619 Bielefeld, Germany
[2]Department of Physical Therapy - Tel Aviv University, Tel Aviv-Yafo, Israel
abarch,dheitkamp,kkrieger@techfak.uni-bielefeld.de

Abstract

When learning a new motor skill, such as a particular tennis shot or a swimming stroke, an expert *scaffolds* the novice throughout the learning process. One of the ways to do this efficiently is by giving different types of feedback to the learner based on their current performance. Following this idea, in order to improve the efficiency of the learning process when an expert is not available, we propose to explore reinforcement learning as a method for building a very simple intelligent tutoring system in the domain of motor skill learning. Our optimization approach builds on the main idea that the policy is rewarded based on how quickly the learner learns the skill. The paper presents promising preliminary results from policy optimization based on 18 training sessions, each consisting of multiple trials.

Introduction

The learning process is commonly accompanied by an expert, who scaffolds the learner (Gonulal and Loewen 2018). Scaffolding has been defined as a process that enables a child or novice to solve a problem, carry out a task or achieve a goal which would be beyond their unassisted efforts (Wood, Bruner, and Ross 1976). According to Zydney (2012), scaffolding provides a temporary structure or support to assist a learner in a task and can be gradually reduced and eventually removed altogether once the learner demonstrates the ability to perform the task independently (Pea 2004). In order to determine the adjustable level of support that meets the learners needs at a particular time, the scaffolding process involves an ongoing diagnosis of a learners proficiency in the task. Previous studies have employed RL in the educational domain (see (Singla et al. 2021) for an overview), however until now very little work has been done in computational scaffolding of motor learning, such as in sports or learning to play a musical instrument (Moringen et al. 2021). Inspired by the above-mentioned characterisation of scaffolding is our two-fold approach. First, we use reinforcement learning (RL) to optimize a policy that provides the level of guidance to the learner based on the observation of their skill level. Second, the policy is rewarded based on how fast the learner is improving - the faster the improvement, the greater the reward.

In many sports, one does not see the hand/arm while it performs the target movement, such as swimming strokes or tennis shots. Nevertheless, such complex movement have to be learned and performed correctly to be effective. Inspired by this example, in our work the study participants learn to perform a simple haptic task: rotation of a knob to a target angle without being able to see it. During the experiment the study participants perform two types of activities, they learn the task and they are being tested. While during the test phase they perform the rotation without seeing, the learning is accompanied by different types of abstract visual feedback described in detailed in the next section.

Our previous study (Heitkamp, Krieger, and Moringen 2022) compared different types of visual guidance that were tested in different groups of study participants, one type of guidance per group. In the above-mentioned study neither the performance of the learners nor the efficiency of learning was taken into consideration while guiding the study participants. No significant advantage could be found among the groups. In this work we show an approach where we aim to optimize visual guidance type based on the efficiency of learning and the performance error. To this end, a set of visual guidance levels represents actions of the policy, while the performance errors calculated after the completion of the rotation task during test represent the states of the policy. The policy is rewarded based on the improvement of the learner from one test trial to the other. At the end of the optimization process, a trained policy should provide, given current performance error of the learner, the type of guidance which results in the fastest improvement (see Figure 1).

We have selected the task, haptic rotation to a target angle, because our previous work showed that without feedback study participants could not perform it correctly (Krieger, Moringen, and Ritter 2019; Krieger et al. 2018a). We found that it is particularly difficult for study participants to perform the rotation with a cylinder (better performance was demonstrated for shapes with edges and vertices, than for "featureless" round shapes). We have therefore selected this shape for the current experiment[1].

Virtual Reality (VR) has previously been used in the education domain and medical training, (e.g. Hsieh and Lee

[1]The illustration of the VR scene and the visual guidance is presented at the following link: https://youtu.be/JlXTxtNCqcw

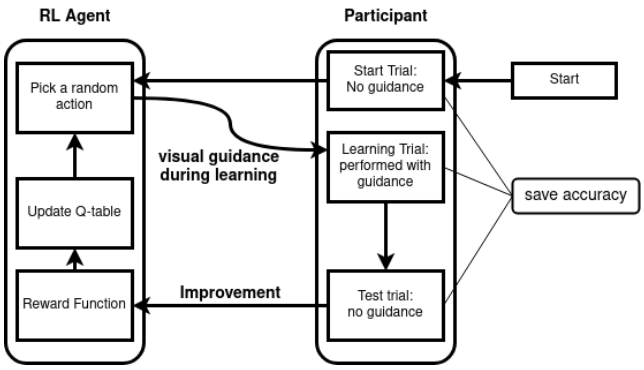

Figure 1: An outline of interaction between the learner and the scaffold represented by an RL policy during training (exploration only).

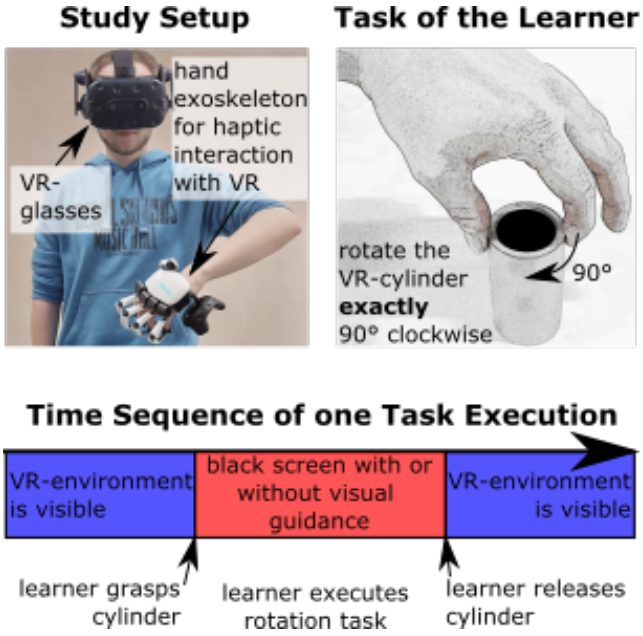

Figure 2: (left) Visualization of the hardware setup, with a study participant wearing the VR glasses and the Dexmo exoskeleton; (right) the task that had to be performed (lower) time sequence of performing a single task

2018). The study participants perform the task in VR, and this allows us to easily switch between different types of guidance as well as the different types of task settings (see Figure 2 for an illustration of hardware and the time sequence of the experiment). They wear an HTC Vive and a Dexmo hand exoskeleton. During practice with an HTC Vive, they are provided with different levels of visual guidance. The sensation that they hold an object is provided by the Dexmo exoskeleton, which also tracks the hand pose. During the test, the study participants are also not provided with visual feedback.

The long-term goal of our research is to use the above setting to optimize and explore guidance to accompany learning a motor task. While this experiment is dedicated to optimizing the choice of guidance types, depending on the skill level (one type of guidance is chosen for each trial), there are many possible optimization problems that can be addressed to improve guidance. e.g. how to optimize guidance while the learner is carrying out the task? What is the optimal sensory modality (audio, vibration, vision) or their combination that should be employed for guidance of learning a motor task? How should a policy trained on multiple learners be adapted to suit individual needs?

To sum up, the main results and contributions of this paper are as following: 1) We proposed and implemented a motor skill learning paradigm in a VR setting with an exoskeleton, integrating a policy trained by reinforcement learning such that the policy is rewarded by a quick improvement of the learner 2) We implemented and tested different types of visual guidance, 3) In the resulting policy, the highest level of feedback prevailed. Multiple simplifications with respect to the chosen modeling approach have been made to deal with the small amount of data.

Methods

Scenario and Task

The experiment takes place in virtual reality. The participants wear a head-mounted display (HTC Vive Pro) and a hand exoskeleton Dexmo (Gu et al. 2016) on the left hand (see Figure 2, left). Participants' hand movements are tracked by an HTC Vive tracker, while finger movements are measured by the Dexmo. The data is transferred into the virtual world and displayed there as a virtual hand that moves according to the trajectory of the real hand. When interacting with virtual objects, when the hand touches the objects and a collision occurs, Dexmo applies force to the fingertips counteracting the collision. This enables the participants to feel as if they are holding a rigid object and allows almost natural interaction with virtual objects. The virtual world is implemented with Unity. The scene contains the cylinder which should be rotated by the subject, the virtual representation of the exoskeleton and the progress bar.

The participants see a virtual cylinder and their task is to grasp the cylinder with the left hand and turn it clockwise exactly 90 degrees (see Figure 2, right). When the cylinder is grasped, the study participants see a black screen only and perform the rotation task, similar to how the participant were *blindfolded* in the previous studies. During the trials during which the participant learns the task, the system provides visual feedback to them that accompanies the task execution. After the rotary knob is released, the black image disappears and the scene is visible to the participant (see Fig. 2 bottom row).

Scaffolding with different levels of visual feedback

In our experiment we used four types of visual feedback $L_0 - L_4$. In level L_0 no feedback is given during the rotation. Figure 3 illustrates levels $L_1 - L_3$, characterized by an increasing amount of visual guidance. In L_1 presented in the bottom row, a progress bar shows the current rotation angle, while the color of the progress bar remains the same

regardless of the angle. In L_2 (middle row), the color of the progress bar changes from red to green when the participant is close to the 90 degrees goal. In L_3 (top row), additionally the German word "Perfekt" (en:"well done") appears when close to the goal angle. In all feedback levels, the final rotation angle was displayed to the subject after the cylinder was released. In the long-run, this design can be employed to train a policy that, with a growing skill of the learner, fades out support and provides less and less visual guidance (see Discussion). However, in the current work we do not model a long-term performance gain, but evaluate the policy only with respect to a training session.

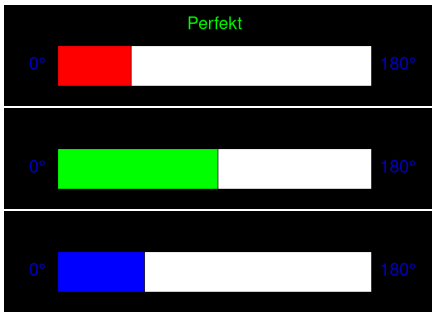

Figure 3: As visual guidance, the participant sees a progress bar, which indicates how far the object was turned. Visualisation of three feedback levels: L_1 illustrates just the progress (bottom row), L_2 also changes color closer to the goal (middle row), and L_3 also indicates when the goal has been reached (top row).

Participants

A total of 10 subjects participated. 7 subjects had experience in VR, 2 had often worked with VR and one subject had no experience with VR. 7 of the 10 subjects had already worked with the setup, while the other 3 subjects had never worked with it. 7 subjects were male and 3 were female. Age ranged from 21 to 26. All participants are university students. All participants had no impairments and were right-handed. This experiment received ethical approval from the ethics commission of Bielefeld University, and participants signed an informed consent form before starting the experiment.

RL policy training

Our goal is to train a policy that picks the level of visual guidance ($L_0 - L_3$) that improves performance of the learner the quickest. Each participant makes multiple turns during an experimental session. In our setting, three turns (baseline, visually guided and test) are called a *trial*. In the **first turn** of a trial the accuracy is measured (the subject does not get feedback). The policy outputs an action (corresponding to the feedback levels $L_0 - L_3$), given the accuracy measurement. Note, this does not happen during policy training. While the policy is trained, an action is selected randomly to enable the policy to explore. The selected visual feedback level guides the learner during the **second turn**. Finally, in the **third turn**, the participant is tested, so they do not receive any feedback. To save time, the third turn of a trial is used as the first turn of the next trial. The participants receive in alternating order feedback and no feedback during the experiment and the RL agent learns based on the trials which consists of three turns, by comparing the improvement from the first to the third turn. The procedure is visualised in pseudo code in Figure 1.

The agent is trained with Q-learning where Q-values $Q(s, a)$ are the expected cumulative rewards when taking action a in state s and following a policy afterwards. When applying Q-learning the following equation is used to update the entries in the q-table:

$$\begin{aligned} Q^{\text{new}}(s_t, a_t) \leftarrow (1-\mu)Q(s_t, a_t) \\ +\mu(R(s_t, s_{t+1}) + \gamma \max_a Q(s_{t+1}, a)), \end{aligned} \tag{1}$$

where a discounting factor $\gamma = 0.95$ is used, and the extent of which the q-values change is determined by the factor $\mu = 0.1$. The q-table is initialised with zeros and default action per state in the policy is also initialised with zero. Because mostly new states were reached during the training, an action was selected randomly.

State Space A state $s \in S \subset \mathbb{R}$ is a scalar value corresponding to the performance error of the participant. It is calculated as the difference between the angle θ captured after an unguided turn and the target angle of 90 degrees: $s = 90 - \theta$.

Action Space The goal of policy training is to learn to provide the kind of visual guidance that results in the quickest improvement, given the learner performance. An action $a \in A := \{0, 1, 2, 3\}$, with actions corresponding to the levels of visual guidance $L_0 - L_3$.

Reward function The reward is calculated as the difference between the first performance error (before the visually guided turn) and the third performance error of a trial $|s_t| - |s_{t+2}|$ (after the visually guided turn).

This reward function was used based on the results of our study (Heitkamp, Krieger, and Moringen 2022). It showed that all visual guidance levels result in an improvement. However, on one hand their impact may be different depending on the progress of the skill training. On the other hand, if an equal amount of reward is given for any improvement, given a particular state, we cannot easily distinguish between the efficacy of different feedback levels. So the effect of the feedback level is included in the reward function. Finally, when a participant reaches the desired goal of 90 degrees two times in a row, a reward of 1 is given. We chose reward 1 because in this special case there exists no improvement that could be rewarded. The effect of this amount is similar to small steps towards 90 degrees. Altogether, we get:

$$R(s_t, s_{t+2}) = \begin{cases} 1 & s_t, s_{t+2} = 0 \\ |s_t| - |s_{t+2}|, & \text{otherwise.} \end{cases} \tag{2}$$

Therefore, for a given state, the guidance level should be selected that yields the quickest improvement.

Algorithm 1: run training session

```
initialise turn ← 0;
while experiment runs do
    if turn == 0 then
        Display no feedback;
        Get participants accuracy;
    else if turn % 2 == 0 then
        Display no feedback;
        Get participants accuracy;
        Calculate reward;
        Update q-table;
    else
        Display random feedback;
    turn ← turn + 1;
end
```

Results

RL Policy

After the policy is trained, the policy Π^* can be derived from the q table. For each state s, the action a with the highest q-value is chosen. Figure 4 illustrates the distribution of states as they occurred during training the policy. It shows that the states occur approximately according to a normal distribution, with most values in the interval $[-25, 25]$ with a mean of -1.0 and a standard deviation of 11.9

A chi-squared test is used to identify if certain actions occur significantly more often than others in the trained policy. The null-hypothesis of the test is that the actions are identically distributed

$$H_0 : F_0 = F_1 = F_2 = F_3$$

where F_i is the distribution function of the ith action. The results are $\chi^2(3, 60) = 10.13$, $p = 0.02$. With $p < 0.05$ the H_0 can be rejected. To find out which actions differ significantly in appearance, a posthoc test is performed. The corrected alpha value (Bonferroni correction) is $\alpha_{Bonferroni} = \alpha/k = 0.05/6 = 0.008$, where k is the number of pairwise repeated tests. Action 3 will be selected significantly more often than action 0 ($\chi^2(1, 34) = 7.53, p = 0.006$) by the trained policy.

To visualize the policy, we employ a window function (Equation 3, with $d = 3$) that counts how often an action appears also in the window around that state. The results are shown in Figure 5. The output of the window function is then put into a Savitzky-Golay filter to smooth the curve for a better visualization.

$$W(s) = \sum_{i=s-d}^{s+d} \mathbf{1}(\Pi^*(i), \Pi^*(s)) \tag{3}$$

$$\mathbf{1}(a, b) = \begin{cases} 1 & \text{if a = b} \\ 0 & \text{otherwise} \end{cases}$$

Data points that lie in the two areas where the purple line (or "missing" action) has a high value are data points that have not been reached very often in the training. Therefore, no specific expression can be made, whereas in the state space from approx. -25 to +25, action 3 is being selected most often.

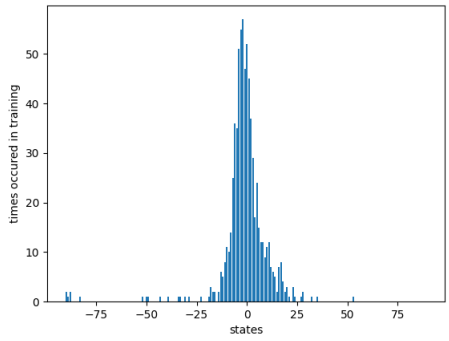

Figure 4: Total number of times a state occurred in training.

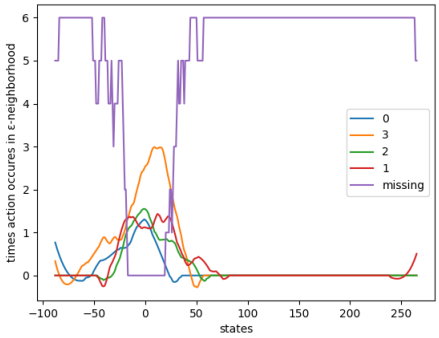

Figure 5: Visualization of possible regions in the policy where actions are likely to be chosen. A window function (with size d=3 in both directions) counts how often an action is predicted by policy for each state. In the interval where the states often occurred in the training (see 4) it is most likely to find action 3. All other actions evenly often present. (Values below 0 appear due to the Savitzky-Golay filter.)

Discussion

In this work, we developed a paradigm in which a learner and a RL policy learn in parallel, while the policy is being optimized with the goal to improve the performance of the learner as quickly as possible. Once converged, we envision such a policy to play a role of a scaffold, and improve learner's training.

In the current work the policy's state is the learners' skill level. The quicker the learner improves on the task, the higher the reward the policy receives. We expected to see a fading out effect in the policy: the closer the learner gets to the perfect performance, the less feedback they receive from the policy. When looking at the visualization of the current policy (see Figure 5), we observe that action 3 (corresponding to the highest level of feedback) shows up most

of the time in the interval $[-25, 25]$. The findings of Douglas and Kirkpatrick (1999) support these results, where they suggest that more information (or feedback) leads to better outcomes. In particular, signaling that the exact target rotation has been reached by displaying "well done", seems to be important.

In our previous experiments, in which blindfolded study participants rotated a knob to a target angle without any guidance, we found a bias for rotation further than the target (Krieger et al. 2018b). Approximately 700 data points were generated during the current experiment. The average rotation error over all trials is 0.26, which may be explained by measurement noise, or the visual guidance that accompanied the learning of the task. It is unlikely that with the current number of samples we have achieved a policy convergence. A larger data sample will be needed to get to the point in which the policy converges. Ultimately, the policy has to include not only exploration, but also exploitation. To this end, the experiment will be repeated with more participants, more information included into the state (such as e.g. velocity during rotation), and a more general model, such as e.g. deep Q-network. Another research thread will be dedicated to optimizing the type of feedback that is given at each point in time during the execution of the motor task. Here the focus will be on a more fine-grained optimization of feedback, which will then automatically generate an optimal visual feedback, instead of optimizing among manually designed feedback modes (such as 0-3 used in this experiment).

Acknowledgements

This work is supported by a Grant from the GIF, the German-Israeli Foundation for Scientific Research and Development.

References

Douglas, S. A.; and Kirkpatrick, A. E. 1999. Model and representation: the effect of visual feedback on human performance in a color picker interface. *ACM Transactions on Graphics (TOG)*, 18(2): 96–127.

Gonulal, T.; and Loewen, S. 2018. Scaffolding technique. *The TESOL encyclopedia of English language teaching*, 1–5.

Gu, X.; Zhang, Y.; Sun, W.; Bian, Y.; Zhou, D.; and Kristensson, P. O. 2016. Dexmo: An inexpensive and lightweight mechanical exoskeleton for motion capture and force feedback in VR. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 1991–1995.

Heitkamp, D.; Krieger, K.; and Moringen, A. 2022. Analyzing hand posture during scaffolded learning of haptic rotation in VR. In *submitted to Eurohaptics*.

Hsieh, M.; and Lee, J. 2018. Preliminary study of VR and AR applications in medical and healthcare education. *J Nurs Health Stud*, 3(1): 1.

Krieger, K.; Moringen, A.; Kappers, A.; and Ritter, H. 2018a. Influence of Shape Elements on Performance During Haptic Rotation. In *Haptics: Science, Technology, and Applications*, 125–137. ISBN 978-3-319-93444-0.

Krieger, K.; Moringen, A.; Kappers, A. M.; and Ritter, H. 2018b. Influence of shape elements on performance during haptic rotation. In *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*, 125–137. Springer.

Krieger, K.; Moringen, A.; and Ritter, H. J. 2019. Number of Fingers and Grasping Orientation Influence Human Performance During Haptic Rotation. In *2019 IEEE World Haptics Conference, WHC 2019, Tokyo, Japan, July 9-12, 2019*, 79–84. IEEE.

Moringen, A.; Ruettgers, S.; Zintgraf, L.; Friedman, J.; and Ritter, H. 2021. Optimizing piano practice with a utility-based scaffold. Technical report, https://arxiv.org/pdf/2106.12937.pdf, Universitt Bielefeld.

Pea, R. D. 2004. The Social and Technological Dimensions of Scaffolding and Related Theoretical Concepts for Learning, Education, and Human Activity. *Journal of the Learning Sciences*, 13(3): 423–451.

Singla, A.; Rafferty, A. N.; Radanovic, G.; and Heffernan, N. T. 2021. Reinforcement Learning for Education: Opportunities and Challenges. *arXiv:2107.08828 [cs]*. ArXiv: 2107.08828.

Wood, D.; Bruner, J. S.; and Ross, G. 1976. The Role of Tutoring in Problem Solving. *Journal of Child Psychology and Psychiatry*, 17(2): 89–100.

Zydney, J. M. 2012. Scaffolding. In Seel, N. M., ed., *Encyclopedia of the Sciences of Learning*, 2913–2916. Boston, MA: Springer US. ISBN 978-1-4419-1428-6.