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 kicking a ball or playing the piano, an expert may use scaffolding to teach the novice - that is, initially focusing on a particular facet of the task, such as first playing the piano with only one hand. To complement a human expert and to improve the efficiency of the learning process when an expert is not available, we propose to use reinforcement learning to build an intelligent tutoring system in the domain of motor skill learning, based on the notion of scaffolding. Our optimization approach builds on the main idea that the policy is rewarded based on how quickly the learner learns the skill.

To demonstrate our approach, we reduce the complexity of the motor task to be learned. The study participants learn to rotate a knob to a target angle in a virtual reality setting. During practice, they are provided with visual guidance according to a policy. During the test, they have to rely on haptic feedback only. The haptic feedback and hand tracking is performed by the Dexmo exoskeleton. The paper presents preliminary promising results from policy optimization based on 18 training sessions, each consisting of multiple trials. Early evaluation shows a noisy increase of the average reward with the number of sessions, indicating a rising efficiency in learning.

Introduction

Scaffolding is central to teaching and learning (Gonulal and Loewen 2018). It 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 his 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 can carry out the performance on his or her own (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. Inspired by this definition, the main idea of this paper is to use reinforcement learning (RL) to optimize a policy that provides the level of guidance to the learner based on their skill level. The policy is rewarded based on how fast the learner is improving - the faster the improvement, the greater the reward. Previous studies have employed RL in the educational domain (Singla et al. 2021), however until now very little work has been done in computational scaffolding of motor learning, such as sports or learning to play a musical instrument (Moringen et al. 2021). We have selected the target 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; Moringen et al. 2017). We found that it is particularly difficult for study participants to perform the rotation with a cylinder. We have therefore selected this shape for the current experiment[1].

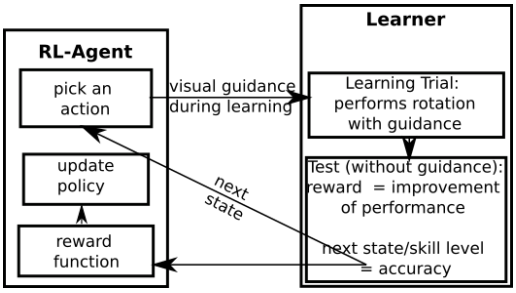

Figure 1: An outline of interaction between the learner and the scaffold represented by an RL policy.

The core of our approach is to develop an experimental framework which can be tested with a small amount of data. In our first experiment, we strongly constrain both the complexity of the motor task to be learned as well as the policy modeling approach. For the same data-efficiency reasons, we have decided to use presumably the most effective guidance modality through visual feedback. To sum up, given the state, the skill level of the learner calculated based on their errors, the policy is optimized to provide the type of visual feedback to foster the quickest reduction of error during the execution of the motor task (see Figure 1).

Virtual Reality (VR) has previously been used in the education domain and medical training, e.g. (Hsieh and Lee 2018). It is also advantageous for the current experiment

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

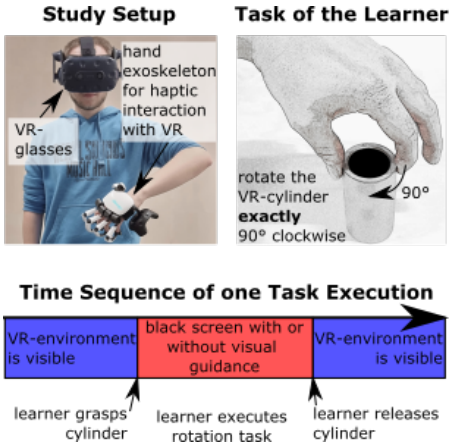

Figure 2: (left) Visualization of the hardware setup, with a study participant weearing the VR glasses and the Dexmo exoskeleton; (right) the task that had to be performed (lower) time sequence of performing a single task

and in the long run. It allows us to easily switch between different types of guidance as well as the different types of task settings. The test subjects perform the task in the VR scene where they experience the visual feedback while they either learn to perform the target task, or their skill is tested (see Figure 2 for illustration of hardware). 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 contributions of this paper are as following: 1) we proposed and implemented a motor skill leaning 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) our approach showed improvement of reward with very limited data.

Methods

Scenario and Task

The experiment takes place in virtual reality. The participants wear a head-mounted display (HTC Vive Pro) and wear a Dexmo (Gu et al. 2016) exoskeleton for the left hand (see Fig 2, left). Participants' hand movements are tracked by an HTC Vive tracker, while finger movements are measured by Dexmo. The data is transferred into the virtual world and displayed there as a virtual hand that moves similar to the real hand. When interacting with virtual objects, e.g. collisions, Dexmo produces force feedback on the fingertips. This allows almost natural interaction with virtual objects.

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. When the cylinder is grasped, a black image appears that blocks the sight of the cylinder, analogous to how the participant was blindfolded in previous studies (see Fig 2, right and lower). The participant can only see the visual feedback provided by the system. After the rotary knob is released, the black image disappears and the scene is visible to the participant.

The experiment within the virtual world is implemented with Unity. The scene contains the cylinder which should be rotated by the subjects, the virtual representation of the exoskeleton and the progress bar.

Scaffolding feedback levels

Four feedback levels were created (see Figure 3) from 0-3, with increasing levels of feedback resp. The third level contains the maximal amount of information. A progress bar shows the current rotation angle of the cylinder. The color changes from red to green when the local rotation approaches the desired 90 degree goal. When reaching close to the goal, the German word "Perfekt" (meaning "well done") appears. In the second feedback level the "Perfekt" text is not shown. In the first level the color of the progress bar stays the same. On the level 0 the progress bar is not shown to the subject, so no feedback is given during the rotation of the cylinder. In all feedback levels, the final rotation angle was displayed to the subject after the cylinder was released. With this design we envision in the long-run to allow for a policy to fade out support, the better the learners performed (see Discussion).

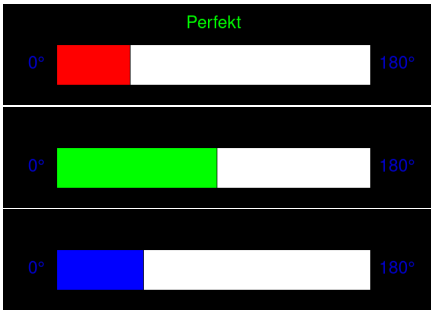

Figure 3: Visualisation of the feedback level 3 to 1. The subject sees a progress bar, which indicates how far the object was turned. In the level 0 (not shown), no feedback is given on the rotation.

Participants

In classic RL, the environment is the world in which the agents can make steps and experience the consequences of their actions. The environment in this case is the combination of all subjects who participated.

A total number 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 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. The 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.

The RL Agent

In this research, the RL agent tries to learn what kind of feedback level a subject needs to improve the performance on turning the cylinder.

A participant makes multiple turns during the experiment. Three turns are called a *trial*. In the first turn of a trial the accuracy is measured (the subject does not get feedback). Then the agent predicts an action (feedback level) which is presented to the subject in the second turn. In the third turn, the subject again receives no feedback. To save time, the third turn of a trial is used as the first turn of the next trial. The only exception to this rule is the very first turn of a training session. This is only measured to get the improvement and is not compared to previous trials. So 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 agent is trained with Q-learning where Q-values $Q(s_i, a)$ are the expected cumulative rewards when taking action a in state s_i and following a policy afterwards. When applying Q-learning the following equation is used to update the entries in the q-table:

$$\begin{aligned} Q_m(s_i, a) \leftarrow (1 - \mu_u) Q_{m-1}(s_i, a) \\ + \mu_u (R(s_i, s_j) + \gamma * \max_{a'} Q_m(s_j, a')) \end{aligned} \tag{1}$$

where m is the current state if the q-table. To reduce the effect of future steps, a discounting factor $\gamma = 0.95$ is used. The extent of which the q-values change is determined by the factor $\mu_u = 0.1$. The q-table is initialised with zero values and default action per state in the policy is initialised with zero. During the training when the agent enters a new state, an action is selected randomly. The random action then determines the feedback level of the next turn.

State Space The state space is a scalar value. The state represents the last accuracy of the subject, i.e. the distance to the goal angle of 90 degrees. It can be positive or negative. The state is used in the reward function where difference between two states is calculated.

Action Space The agent learns to give the right kind of feedback at the right time. Therefore the *action space* is also a scalar which represents the action index (0,1,2,3). *Action 0* is giving no feedback, while *action 3* is giving the most amount of information in the feedback. The detailed actions are already described in Section *Scaffolding feedback levels*.

Reward function The reward function is defined such that feedback that leads to improvement is rewarded and feedback that leads to deterioration is penalized. The reward is the difference between the rounded first angle and the rounded third angle of a trial (see eq. 2). When the difference is positive (so the difference on the third turn is less than in the first turn) it is counted as an improvement. If the difference is negative, it is a deterioration.

This reward function was used because multiple feedback levels may result in an improvement, but the amount of improvement may differ between them. So if an equal amount of reward is given for any improvement, we cannot distinguish between the efficacy of the feedback levels. So the effect of the feedback level is included in the reward function. Lastly, when a participant reaches the desired goal of $accuracy = 0$ two times in a row, reward 1 is given. We chose reward 1 because in this special case exists no improvement that could be rewarded. The effect of this amount is similar to small steps towards 90 degrees.

The final formula of the reward function has the following form:

$$R(x, y) = \begin{cases} 1 & \text{x = y = 0} \\ |x| - |y| & \text{otherwise} \end{cases} \tag{2}$$

where y is current state and x is the previous state. The absolute values of the states are important, not the sign. E.g. a participant who reaches first an accuracy of 70 and then an accuracy of -50 would be interpreted as an improvement of 20.

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

Policy of RL agent

After the agent was trained, the optimal policy Π_* can be derived from the q table with the following formula:

$$\Pi_*(s) = \underset{a' \in \mathcal{A}}{\operatorname{argmax}} Q(s, a') \tag{3}$$

For each state s is the action a' chosen, which has the highest q-value. In Figure 4 the total number of times a state occurred while training the agent is shown. It shown that the

states do not occur evenly, but rather are approximately normally distributed, with most values between -25 and 25, and a mean of -1.0 and std of 11.9

A chi-squared test is used to identify if certain actions occur significantly more often as others. The null-hypothesis of the test is 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 occurs significantly more often than action 0 ($\chi^2(1, 34) = 7.53, p = 0.006$).

To visualize the policy, we employ a window function (Equation 4, 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.

$$W(s) = \sum_{i=s-d}^{s+d} \mathbf{1}(\Pi_*(i), \Pi_*(s)) \tag{4}$$

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

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 seems to be most selected.

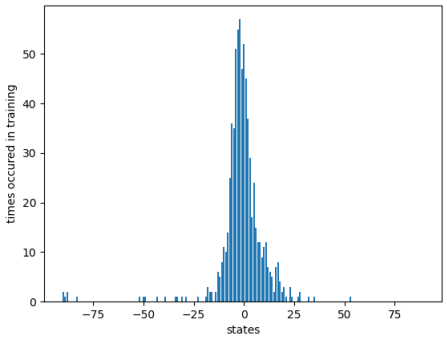

Figure 4: Total number of times a state occurred in training.

Discussion

In this work, we tested a paradigm in which a learner and a RL policy learn in parallel, while maintaining a reciprocal relationship with each other. Once converged, we envision such a policy to play a role of a scaffold, and improve learner's training.

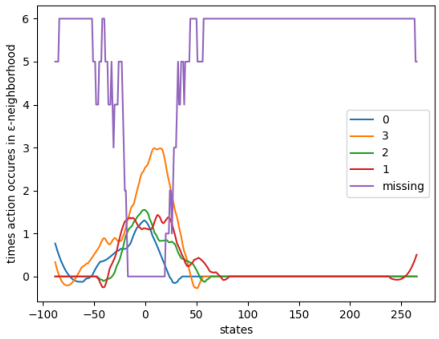

Figure 5: Visualization of possible regions in the policy where actions are likely to be chosen. A window function (with size d=3 in both directions) counts how often an action is predicted by policy for each state. In the interval where the states often occurred in the training (see 4) it is most likely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

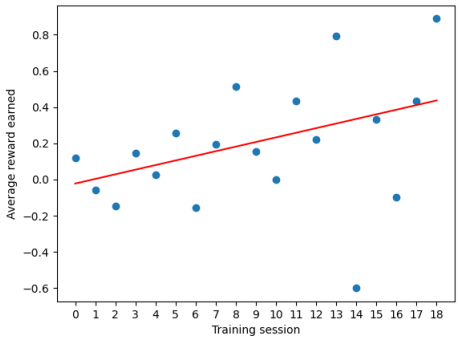

Figure 6: Average reward gained per session. One outlier was removed.

In the current work the policy's state is the learners' skill level, it is rewarded for providing visual guidance that results in the most effective learning by the human learner. 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 order to achieve the fading out effect that is one of the main principles of scaffolding, we will penalize the usage of the action 3, when the learners get close to an optimal performance.

In our previous experiments, in which blindfolded study participants rotated a knob to a target angle, we found a bias for rotation further than the target (Krieger et al. 2018b).

Approximately 700 data points were generated during the current experiment. And the average rotation error over all trials is 0.26, which can be explained by measurement noise. This may be due to the visual guidance that accompanied the learning of the task. The increase in the average reward with the increasing number of training sessions may also explain the improvement in the average error. Although we can observe an increasing trend of the average reward, we have not yet achieved policy convergence with the current number of trials. A larger data sample will be needed to get to the point in which the policy converges. 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).

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.

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.; Krieger, K.; Kõiva, R.; and Ritter, H. 2017. Haptic Interface Twister https://ni.www.techfak.uni-bielefeld.de/node/3573. URL.

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.