Semantic Trajectory Segmentation and Comparison for Robot Demonstration Learning with Different Input Devices in Virtual Reality

Robin Helmert
Medical School OWL
Bielefeld University
Bielefeld, Germany
robin.helmert@uni-bielefeld.de

Kira Loos
Medical School OWL
Bielefeld University
Bielefeld, Germany
kloos@uni-bielefeld.de

Anna-Lisa Vollmer
Medical School OWL
Bielefeld University
Bielefeld, Germany
anna-lisa.vollmer@uni-bielefeld.de

***Abstract*—Training a robot through demonstration requires robust algorithms capable of processing provided trajectories to generate high-quality task executions. However, the success of this process is highly dependent on the quality of the trajectories. Poor-quality trajectories hinder the ability of algorithms to learn and generalize effectively, while high-quality trajectories can improve learning outcomes, reducing the need for overly complex algorithms. In this work, we propose an initial method for comparing sets of semantically annotated trajectories used for robot demonstration. To evaluate the proposed methodologies, we recorded trajectories of 60 participants in a study setup in Virtual Reality where each participant tested one of three different control-visualization compositions. We use classical approaches such as Dynamic Time Warping and Discrete Fréchet-Distance to measure the similarity between segments of recorded trajectories and show that different input devices and visualization combinations affect the resulting trajectory metrics.**

***Index Terms*—human-robot interaction; virtual reality; semantic trajectories; trajectory similarity; input controls**

Fig. 1. The virtual setup for the bread-cutting task, showing the bread and knife used in the experiment.

I. Introduction

In the field of robotic learning from demonstration (LfD), a key challenge is to obtain high-quality trajectories recorded from human demonstrations [1]. These trajectories serve as the foundation for robots to replicate human behavior or execute specific tasks. As the quality of a demonstration trajectory directly influences learning outcomes, a well-executed human demonstration is essential for effective LfD. Demonstrating a task to a robot in a real-world setting relies on precise object and motion tracking, but directly moving the robot by hand is challenging due to the need to counteract motor forces. Moreover, the latter often affects additional joints unintentionally. Virtual reality (VR) provides a powerful solution to these challenges by providing a manipulatable environment with precise tracking [2], [3] and thus ensuring high accuracy in recording human-generated trajectories. However, in VR, different input devices and interfaces offer varying levels of immersion, each with its own advantages and disadvantages for demonstration quality. In addition, while methods such as Dynamic Time Warping, Fréchet Distance, and Least Common Subsequence are commonly used to define differences or similarities between trajectories recorded through demonstrations, they often fall short in capturing nuanced differences in task-specific trajectories, particularly when semantic context is involved.

In order to compare trajectory quality, we analyzed trajectories obtained from a VR study, involving 60 participants, on the task of bread cutting. As shown in Figure 2, we used different input controls including motion capture gloves and a standard controller with various visualizations. The primary objective of this research is to investigate how to create high-quality trajectories created by lay user participants suitable for robot training. To find nuanced differences between recorded trajectories, we propose new approaches by combining existing trajectory handling methods, trajectory segmentation, and other quality assessments. Using the cutting task as a practical example, we address the research question of which control-visualization combination produces the best trajectory quality.

979-8-3503-7893-1/25/$31.00 ©2025 IEEE

II. Background and Related Work

A. Demonstration learning

Learning from demonstration, also known as imitation learning, has been extensively researched [4], [5]. Research has shown that the learning performance is closely tied to the quality of demonstrations [6]–[8], and optical recordings, such as those made with Kinect cameras [9], can degrade trajectory quality. This can be overcome by using virtual reality, where movements can be tracked in millimeter range [2], [3]. Demonstration recordings and trajectories directly reflect actual movements, making it essential to understand the influence of different controls and interfaces. It has already been shown that for common tasks in virtual reality, usability and task load does not differ much for gesture and controller based controls [10]. However, it is not clear how the different controls affect the quality of the resulting trajectories which can hold information like task performance, accuracy, velocities and pose data which all can be useful for training robots. While there is a lot of research that utilizes trajectories [9], [11], [12], methods to evaluate trajectory quality are scarce and, to our knowledge, not well-explored [13]. Moreover, while methods like smoothness and continuity are sometimes used for similarity [14], they do not work for trajectory paths where abrupt movement changes are necessary for success, i.e for the sawing movements during cutting.

B. Trajectory similarity measures

Since the similarity of trajectories can be used to differentiate between recordings, effective similarity measures are essential. Basic trajectory similarity measures are well-established, with various methods focusing on different aspects. For comparing the direct metric distance between two trajectories of the same length the *Lock-step Euclidean Distance* (LSED) is the most straightforward method [15]. As it compares point by point for equal length trajectories, it does not account for shape similarity and is highly sensitive to velocity variations. Therefore it is not suitable for our approach. This also counts for the *Least Common Subsequence* (LCSS) [12]. The LCSS is also a measure of similarity, although it focuses on finding the amount pairs with similar points. Two points are similar when their distance falls below a defined threshold. It does not rate the similarity as a whole and gives bad results when points are for example partially clustered. The *Edit Distance on Real Sequences* [16] also calculates similarity in a form of, given a defined threshold distance, how many points of the trajectory would needed to be edited such that all points are below the threshold. It shares the same limitations as LCSS and is therefore unsuitable for quality assessment.

To overcome velocity and clustering problems *Dynamic Time Warping* (DTW) can be used. [17] [18]. It can handle different speeds and timings but cannot be used as a metric since it does not satisfy triangle inequality [15]. Finally, there is the *Fréchet distance* (FD) and its discrete approximation (DFD), to measure similarity in form of a distance between two trajectories [19] [20]. It is often called the dog-leash distance. If two trajectories would be a dog and the owner, the distance given here measures the minimal length of the leash to make that walk possible. For our analysis we use the DFD and DTW, since they give an estimate about the shape similarity of our trajectories; particularly of the segmented ones.

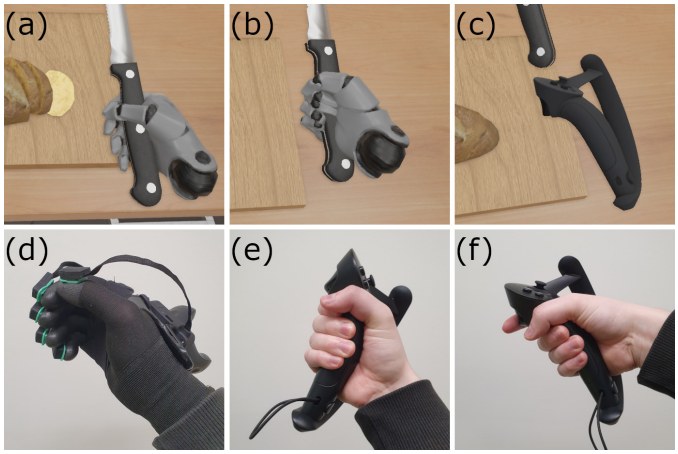

Fig. 2. The different input and visualization combinations and their virtual and physical looks when holding the knife. (a) and (d) depict the Manus gloves and their visualization, (b) and (e) show one input medium with a Valve Index controller and its visualization, and (c) and (f) show the other visualization of the Valve Index Controller.

C. Semantic trajectories

It is not enough to measure the trajectory as a whole. Since tasks often include subtasks with varying levels of importance, it is reasonable to divide trajectories into sub-trajectories. For instance, movements such as headset adjustments should not be included, as they do not contribute to the task. Since our trajectories additionally have annotations like events or stops, they are called semantic trajectories [21]. We can use those semantic events as segmentation candidates to automate the segmentation process.

While research has compared automated and manual annotation for creating semantic trajectories [22] or segmenting those [23]–[25], assessing their importance and quality, and identifying which input mediums yield the best results remain underexplored. Through our study and the resulting trajectories presented in this paper, we contribute to the ongoing efforts to address these questions.

III. Study Design

In our study, participants were placed in a virtual world where they performed everyday tasks in a physics-based scenario designed to be as realistic as possible. One of their tasks was to cut bread with a knife, as shown in Figure 1.

A. Participants

We recruited 60 participants (32 male, 27 female, 1 diverse), all right-handed, mostly students aged 18 to 34 from social sciences, engineering, or education. Most of them were inexperienced with VR and its controls. All subjects provided informed

consent before participating, and the study was conducted in accordance with the Declaration of Helsinki, with protocol approval from Bielefeld University's Ethics Committee.

B. Input Controls

For the study, participants were randomly assigned to one of the three control-visualization combinations shown in Figure 2. The first used Manus Quantum XR motion capture gloves (M) to track the precise movements of the user's hands and fingers [26]. Users could grab objects in various ways, such as using pinch gestures with two or more fingers or performing fist grabs. The second combination used a valve index controller[1]. Using touch and force sensors, the user's hand was simulated in VR (H). By applying force to the controller users could grab and hold virtual objects. The third and last control used the same controller but in VR the controller itself was shown (C). In this case, grasping was performed by pressing the controller's trigger button, as is standard in state-of-the-art VR controllers. We selected these input mediums and their respective controls to balance state-of-the-art techniques, such as those used in gaming, with natural interactions, aiming to assess their impact on teaching robots tasks.

C. Task Description

In the study, users were tasked with completing some common kitchen activities, among them, the task of cutting bread. Here, a bread was placed on a table in the virtual scene and a knife was positioned beside it. The participants should grasp the knife and cut the bread exactly three times. No automated task completion mechanism was provided. The participants were told that they should let us know when they thought that they completed the task successfully.

IV. Results and Analysis

This paper focuses on the trajectories captured during the study. The extracted trajectories for this task concerned real hand movements in VR, including reaching for and grasping the knife, holding the knife while moving it to the bread, cutting the bread three times, and finally placing the knife back down.

A. Trajectory Segmentation

As the complete task of cutting bread with a knife consists of multiple parts, each with its own potential quality, a single measurement for the entire trajectory would not be suitable. Instead, the trajectory was divided into segments, each evaluated individually. For example, the segment where the hand approaches the knife focuses on a different aspect than the segment involving the cutting action. The first focuses on the correct endpoint where the latter focuses on the movements during the actions. Therefore, segmentation points were chosen based on changes in the subtasks. In the cutting scenario these are for example the points where the knife is grasped and finally placed back down. In general, the one optimal trajectory may not exist, and there are often multiple solutions of similar high quality. But there is an optimal trajectory for this task if we consider the trajectory of the semantics. The task description infers that the bread has to be cut three times. Therefore the optimal trajectory considering only the semantic aspects would start with no action, followed by a grasp of the knife which in turn is followed by three cuts and afterward the placing down of the knife leaving with and empty action set.

[1]https://www.valvesoftware.com/en/index/controllers

B. Semantic Trajectory

The semantic approach focuses solely on the actions within the trajectory, disregarding its spatial aspects. With this the duration of actions, the position where they were happening and additionally their correct order and amount can be confirmed. For example, in a task where the user cuts the bread exactly once, the semantic order of events would be represented as

$$(\{\}, \{grasp\}, \{grasp, cut\}, \{grasp\}, \{\})$$

indicating that the hand was empty at the beginning, then grasped an object, presumably the knife, cut the bread one time and then released the grasp such the hand was empty again. At the same time this is the optimal semantic trajectory as it contains only the necessary actions. For the study task, the optimal trajectory for three cuts could be used as a filter to remove all trajectories that have irrelevant or erroneous actions. However, instead we calculate the Levenshtein distance [27] to measure the semantic distance between two given trajectories. If the distance is zero, the two semantic trajectories are equal and if not the result will indicate how many insertions, edits, or deletions are necessary to align them. We calculated the Levenshtein distance for the whole trajectory, and all other distances only for the cutting part since what happened before and after that part might vary heavily. The cutting part starts with the knife grasp before the first cut and ends after the last cut. A cut was recorded if the knife entered the bread on one side. A cut was considered successful if the knife exited the bread from the opposite side. Unsuccessful cuts occurred when the knife entered the bread from the top, initiated the cutting action, but exited from the same side without cutting through. In real-world scenarios, this would be considered a failure, and the same criteria was applied in the virtual setting too. This ensures that all subtrajectories are compared only to similar subtrajectories considering the cutting subtask which starts by grasping the knife and ends with the last cut. We then analyzed the resulting trajectories with different metrics in regards to different aspects of their attributes, like pairwise average DFD, pairwise average DTW, velocities and Levenshtein distances. For the latter, each cutting action, independent of its success, was counted as it was recorded. The results are shown in Table I. We also calculated significances regarding velocity and Levenshtein distance. A Kruskal-Wallis test revealed significant differences for both measurements. Post-hoc pairwise Mann-Whitney U tests with Holm correction indicated that Manus users exhibited lower average velocities than users with the controller and hand visualization ($p < 0.05$). Additionally, when considering the average Levensthein distance, Manus

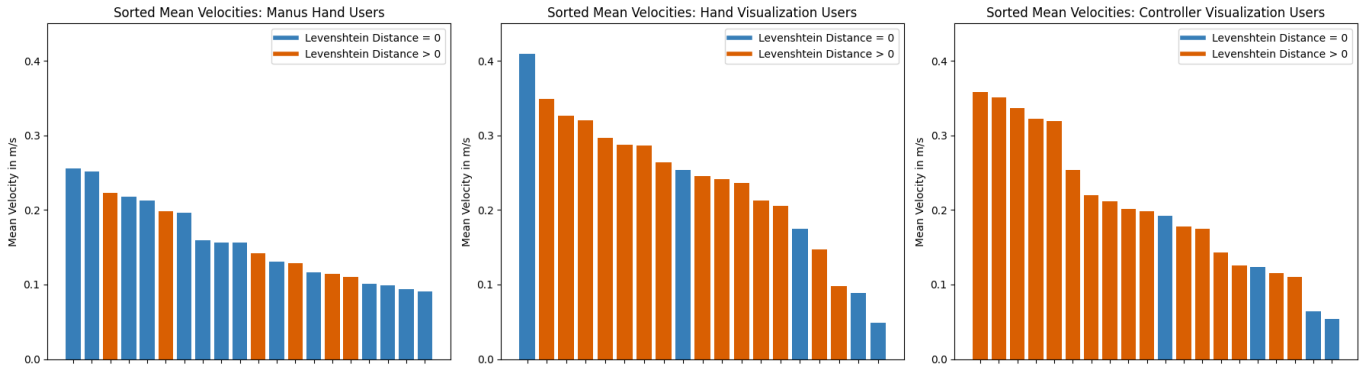

Fig. 3. Average velocities per user. Sorted for better visualization. Blue bars indicate a semantically perfect cutting execution and thus a Levenshtein distance of zero, and orange bars indicate a distance of greater than zero.

hands significantly outperformed both other controls ($p < 0.05$ for both). It shows that for Manus gloves, the amount of trajectories that fit the perfect semantic trajectory, i.e. have a Levenshtein distance of zero, is higher than for the other controls. Also the mean and variance of the Levenshtein distance is also lower for the Manus Hands. For a clearer evaluation of the velocity values, the sorted average velocities per user are shown in Figure 3. Here, blue bars denote a Levenshtein distance of zero, and orange ones indicate a distance of greater zero.

Metric	Manus(M)	Hand(H)	Controller(C)
#Levenshtein = 0	14	5	4
#Levenshtein > 0	6	14	16
Levenshtein Mean	0.7^{HC}	3.68	4.65
Levenshtein Variance	1.31	19.8	53.33
Velocity Mean	0.16^{H}	0.24	0.20
Velocity Variance	0.0033	0.0048	0.0051
Discrete Frechet Dist.	0.22	0.66	0.55
Dynamic Timewarping Dist.	44.96	139.87	82.48
#Levenshtein = 0*	14	5	4
#Levenshtein > 0*	5	10	14
Velocity Mean*	0.15	0.26	0.19
Velocity Variance*	0.0032	0 0.0042	0.0048
Discrete Frechet Dist.*	0.22	0.48	0.51
Dynamic Timewarping Dist.*	42.86	90.67	66.69

TABLE I
CHARACTERISTICS OF THE CUTTING TRAJECTORIES OF THE USER GROUPS, EXCLUDING SUBTRAJECTORIES BEFORE AND AFTER. GREY HIGHLIGHTS INDICATE SEMANTIC ANALYSES, SUPERSCRIPTS DENOTE SIGNIFICANT DIFFERENCES, AND ASTERISKS MARK TRAJECTORIES LIMITED TO THE FIRST THREE CUTS, EXCLUDING THOSE WITH FEWER THAN THREE.

V. DISCUSSION AND FUTURE WORK

Considering the presented results of Table I, it shows that the Manus gloves have in comparison to the other input modality the least errors regarding the semantic trajectories. Additionally, the velocities and distances between the recorded trajectories are the lowest for Manus gloves. Figure 3 enhances the argument of the velocities. Lower velocities appear to correlate with more accurate executions of semantic trajectories. This indicates that selecting appropriate inputs and visualizations affects the quality of trajectories, a factor often overlooked. However, the general relationship between these measures and other not yet considered measures remains unclear. Also for generalization in more varied tasks, the relevant qualities per subtask need to be evaluated. However, since often subtasks are similar, already known quality measurements can then be reused and weighted according to the relevance of the subtask. For this study, we focused exclusively on the knife-cutting task, in particular the grasping and cutting actions was relevant in our case. The focused cutting subtask was path-oriented, meaning that the method of execution was critical. We accounted for this by including partially completed cuts in our analysis. However, for tasks where execution methods are crucial, more robust measurement techniques are needed. We only used two standardized similarity measures which each has its own weaknesses. For example, a key limitation of the DFD is its susceptibility to outliers. The DFD calculates the largest point-wise smallest distance, meaning that a single distant point in one trajectory can dominate the value, rendering the rest of the trajectory irrelevant. On the other hand, the DTW algorithm provides a better similarity measure by minimizing the sum of point-wise distances. However, it is not a metric, as it fails to satisfy the triangle inequality. Therefore, our next steps involve extending and improving these measurements while analyzing the other tasks conducted in the study. We will further explore trajectory features such as acceleration, pauses, and pace, as demonstrated by Vollmer et al. [28]. Subsequently, we aim to develop a general segmentation framework for more complex tasks, where path-oriented subtasks alternate with goal-oriented ones. Additionally, we plan to identify differences in recordings and, by combining the results, investigate optimal environments and input modality choices for generating high-quality trajectories. These trajectories, produced by both lay users and experts, will be applicable to various purposes, including robot training. Finally, we plan to annotate and publish all recorded trajectories to enable other research groups to utilize this data for training purposes.

REFERENCES

[1] D. Koert, G. Maeda, R. Lioutikov, G. Neumann, and J. Peters, "Demonstration based trajectory optimization for generalizable robot motions," in *16th IEEE-RAS International Conference on Humanoid Robots, Humanoids 2016, Cancun, Mexico, November 15-17, 2016*, pp. 515–522, IEEE, 2016.

[2] D. C. Niehorster, L. Li, and M. Lappe, "The accuracy and precision of position and orientation tracking in the htc vive virtual reality system for scientific research," *i-Perception*, vol. 8, no. 3, p. 2041669517708205, 2017.

[3] S. P. Sitole, A. K. LaPre, and F. C. Sup, "Application and evaluation of lighthouse technology for precision motion capture," *IEEE Sensors Journal*, vol. 20, no. 15, pp. 8576–8585, 2020.

[4] C. G. Atkeson and S. Schaal, "Robot learning from demonstration," in *Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997), Nashville, Tennessee, USA, July 8-12, 1997* (D. H. Fisher, ed.), pp. 12–20, Morgan Kaufmann, 1997.

[5] S. Schaal, "Is imitation learning the route to humanoid robots?," *Trends in cognitive sciences*, vol. 3, no. 6, pp. 233–242, 1999.

[6] B. D. Argall, S. Chernova, M. M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics Auton. Syst.*, vol. 57, no. 5, pp. 469–483, 2009.

[7] M. Sakr, Z. J. Li, H. F. M. V. der Loos, D. Kulic, and E. A. Croft, "Quantifying demonstration quality for robot learning and generalization," *IEEE Robotics Autom. Lett.*, vol. 7, no. 4, pp. 9659–9666, 2022.

[8] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation," in *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pp. 1–8, IEEE, 2018.

[9] D. A. Duque, F. A. Prieto, and J. G. Hoyos, "Trajectory generation for robotic assembly operations using learning by demonstration," *Robotics and Computer-Integrated Manufacturing*, vol. 57, pp. 292–302, 2019.

[10] C. Khundam, V. Vorachart, P. Preeyawongsakul, W. Hosap, and F. Noël, "A comparative study of interaction time and usability of using controllers and hand tracking in virtual reality training," *Informatics*, vol. 8, no. 3, p. 60, 2021.

[11] H. Manh and G. Alaghband, "Scene-lstm: A model for human trajectory prediction," *CoRR*, vol. abs/1808.04018, 2018.

[12] M. Vlachos, D. Gunopulos, and G. Kollios, "Discovering similar multidimensional trajectories," in *Proceedings of the 18th International Conference on Data Engineering, San Jose, CA, USA, February 26 - March 1, 2002* (R. Agrawal and K. R. Dittrich, eds.), pp. 673–684, IEEE Computer Society, 2002.

[13] S. Xu, Y. Ou, J. Duan, X. Wu, W. Feng, and M. Liu, "Robot trajectory tracking control using learning from demonstration method," *Neurocomputing*, vol. 338, pp. 249–261, 2019.

[14] T. B. Tuli, M. Manns, and S. Zeller, "Human motion quality and accuracy measuring method for human-robot physical interactions," *Intell. Serv. Robotics*, vol. 15, no. 4, pp. 503–512, 2022.

[15] Y. Tao, A. Both, R. I. Silveira, K. Buchin, S. Sijben, R. S. Purves, P. Laube, D. Peng, K. Toohey, and M. Duckham, "A comparative analysis of trajectory similarity measures," *GIScience & Remote Sensing*, vol. 58, no. 5, pp. 643–669, 2021.

[16] L. Chen, M. T. Özsu, and V. Oria, "Robust and fast similarity search for moving object trajectories," in *Proceedings of the ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, USA, June 14-16, 2005* (F. Özcan, ed.), pp. 491–502, ACM, 2005.

[17] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Knowledge Discovery in Databases: Papers from the 1994 AAAI Workshop, Seattle, Washington, USA, July 1994. Technical Report WS-94-03* (U. M. Fayyad and R. Uthurusamy, eds.), pp. 359–370, AAAI Press, 1994.

[18] L. R. Rabiner and B. Juang, *Fundamentals of speech recognition*. Prentice Hall signal processing series, Prentice Hall, 1993.

[19] H. Alt and M. Godau, "Computing the fréchet distance between two polygonal curves," *Int. J. Comput. Geom. Appl.*, vol. 5, pp. 75–91, 1995.

[20] T. Eiter and H. Mannila, "Computing discrete fréchet distance," 1994.

[21] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. F. de Macêdo, F. Porto, and C. Vangenot, "A conceptual view on trajectories," *Data Knowl. Eng.*, vol. 65, no. 1, pp. 126–146, 2008.

[22] J. Salfity, S. Wanna, M. Choi, and M. Pryor, "Temporal and semantic evaluation metrics for foundation models in post-hoc analysis of robotic sub-tasks," *CoRR*, vol. abs/2403.17238, 2024.

[23] M. Liu, G. He, and Y. Long, "A semantics-based trajectory segmentation simplification method," *Journal of Geovisualization and Spatial Analysis*, vol. 5, pp. 1–15, 2021.

[24] L. Santos, K. Khoshhal, and J. Dias, "Trajectory-based human action segmentation," *Pattern Recognit.*, vol. 48, no. 2, pp. 568–579, 2015.

[25] S. H. Lee, I. H. Suh, S. Calinon, and R. Johansson, "Autonomous framework for segmenting robot trajectories of manipulation task," *Auton. Robots*, vol. 38, no. 2, pp. 107–141, 2015.

[26] MANUS, "Quantum xr metagloves — precise finger tracking for xr." https://www.manus-meta.com/products/quantum-xr-metagloves, 2024. Accessed: 2024-11-27.

[27] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Proceedings of the Soviet physics doklady*, 1966.

[28] A.-L. Vollmer, M. Mühlig, J. J. Steil, K. Pitsch, J. Fritsch, K. J. Rohlfing, and B. Wrede, "Robots show us how to teach them: Feedback from robots shapes tutoring behavior during action learning," *PloS one*, vol. 9, no. 3, p. e91349, 2014.