PRO-SKILL: A Platform for Multi-Modal Capturing and Automated Assessment of Manual Craftsmanship Skills using Trajectory-Based Probabilistic Modeling

JASPER VAN DER AUWERA, KU Leuven - Flanders Make, Belgium
DANNY LEEN*, GUSTAVO ALBERTO ROVELO RUIZ, and LARS GIELEN, Hasselt University - Flanders Make, Belgium
WILM DECRÉ and ERWIN AERTBELIËN, KU Leuven - Flanders Make, Belgium
STIJN KINDT, HOANG-LONG CAO, and TOM VERSTRATEN, Vrije Universiteit Brussel - Flanders Make, Belgium
HERMAN BRUYNINCKX, KU Leuven - Flanders Make, Belgium
RAF RAMAKERS, Hasselt University - Flanders Make, Belgium

Pro-Skill is a novel system for capturing, modeling, and evaluating the performance of motor skills in manual craftsmanship activities. The system uses a multi-modal capturing technique that combines motion tracking, tool tracking, and other sensor data trajectories (e.g., IMU, force sensing) to record performances of craftsmanship tasks. Using Trajectory-parameterized Probabilistic Principal Component Analysis (TraPPCA), Pro-Skill learns a probabilistic expert model that captures both the core patterns and natural variability from a limited number of expert demonstrations. A trainee's performance can then be compared with the TraPPCA model to assess how closely their execution aligns with that of a single reference expert, providing a basis for personalized feedback. We validate Pro-Skill's automated assessments against human expert judgments in a woodworking task involving a handsaw, finding almost-perfect agreement on overall performance. These results show that Pro-Skill can provide valid assessments of the handsaw task from a small set of expert demonstrations. In addition, two proof-of-concept demonstrations illustrate that Pro-Skill's capturing, modeling, and assessment pipeline can be adapted to different sensor configurations and to other manual tasks with a nominally monotonic structure.

CCS Concepts: • **Human-centered computing → Interactive systems and tools**.

Additional Key Words and Phrases: Craftsmanship, Skill Assessment, Multi-Modal Sensing

* Corresponding author

Authors' Contact Information: Jasper Van der Auwera, KU Leuven - Flanders Make, Department of Mechanical Engineering, Leuven, Belgium, jasper.vanderauwera@kuleuven.be; Danny Leen (corresponding author), Hasselt University - Flanders Make, Digital Future Lab, Diepenbeek, Belgium, danny.leen@uhasselt.be; Gustavo Alberto Rovelo Ruiz, Hasselt University - Flanders Make, Digital Future Lab, Diepenbeek, Belgium, gustavo.roveloruiz@uhasselt.be; Lars Gielen, Hasselt University - Flanders Make, Digital Future Lab, Diepenbeek, Belgium, lars.gielen@uhasselt.be; Wilm Decré, KU Leuven - Flanders Make, Department of Mechanical Engineering, Leuven, Belgium, wilm.decre@kuleuven.be; Erwin Aertbeliën, KU Leuven - Flanders Make, Department of Mechanical Engineering, Leuven, Belgium, erwin.aertbelien@kuleuven.be; Stijn Kindt, Vrije Universiteit Brussel - Flanders Make, Brubotics, Brussels, Belgium, stijn.kindt@vub.be; Hoang-Long Cao, Vrije Universiteit Brussel - Flanders Make, Brubotics, Brussels, Belgium, hoang.long.cao@vub.be; Tom Verstraten, Vrije Universiteit Brussel - Flanders Make, Brubotics, Brussels, Belgium, tom.verstraten@vub.be; Herman Bruyninckx, KU Leuven - Flanders Make, Department of Mechanical Engineering, Leuven, Belgium and TU Eindhoven, Department of Mechanical Engineering, Eindhoven, The Netherlands, herman.bruyninckx@kuleuven.be; Raf Ramakers, Hasselt University - Flanders Make, Digital Future Lab, Diepenbeek, Belgium, raf.ramakers@uhasselt.be.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
© 2026 Copyright held by the owner/author(s).
ACM 2474-9567/2026/12-ART257
https://doi.org/10.1145/3857928

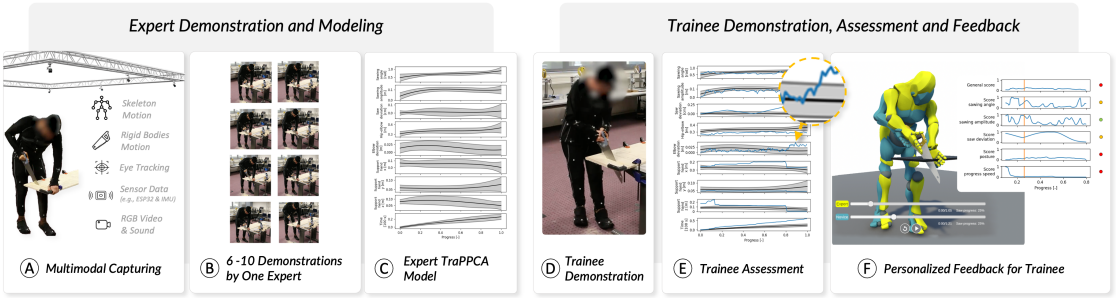


Fig. 1. Overview of the Pro-Skill workflow. (a) Multi-modal capturing records skeleton motion, rigid body motion, eye gaze, sensor data, and video/sound. (b) A single expert provides 6–10 demonstrations of the craftsmanship task. (c) These demonstrations are used to construct an expert skill model using TraPPCA. (d) Trainee task execution is subsequently captured. (e) Performances are automatically assessed against the expert model across multiple parameters (e.g., angle, amplitude, posture, progress speed, deviation). (f) Pro-Skill provides personalized feedback to the trainee.

ACM Reference Format:
Jasper Van der Auwera, Danny Leen, Gustavo Alberto Rovelo Ruiz, Lars Gielen, Wilm Decré, Erwin Aertbeliën, Stijn Kindt, Hoang-Long Cao, Tom Verstraten, Herman Bruyninckx, and Raf Ramakers. 2026. PRO-SKILL: A Platform for Multi-Modal Capturing and Automated Assessment of Manual Craftsmanship Skills using Trajectory-Based Probabilistic Modeling. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 10, 4, Article 257 (December 2026), 30 pages. https://doi.org/10.1145/3857928

1 Introduction

Manual craftsmanship skills such as cutting, joining, and finishing in woodworking, brazing and welding in metalworking are often passed on through apprentice-based training, where the apprentice learns from the expert over an extended period involving observation, discussion, guided practice, repetition, and gradual improvements. Traditional apprentice-based learning is time-consuming and does not scale well because it depends heavily on the availability of skilled experts. An effective apprenticeship also requires the expert to reflect consciously on their own processes, when and how actions are performed. Yet, because many actions are embodied, experienced craftspeople may be unaware of the exact actions they take or struggle to articulate them, as their routines have become automatic and deeply internalized. As Polanyi states, ”we know more than we can tell” [39].

Pye [40] describes manual craftsmanship in terms of *workmanship of risk*, where quality is not predetermined, but remains “at risk” throughout the act of making, controlled by moment-to-moment judgment, dexterity, and care. For example, a minor lapse, such as a slipping chisel, can erase hours of work. Mastering craftsmanship skills requires embodied knowledge gained and refined through hands-on experience. In contrast to procedural knowledge, which involves discrete instructions, transferring embodied knowledge is challenging because it is often difficult to visualize, verbalize, or formalize in a step-by-step plan [33, 49]. For example, a welding tutorial can list the steps, yet successful execution hinges on precise control across multiple modalities: maintaining the correct torch angle, travel speed, and arc length while continuously adapting to visual and auditory cues and tactile feedback. A crucial aspect of mastering embodied knowledge is the expert’s feedback on the trainee’s performance. An expert not only evaluates the resulting outcome, but also observes how the work is done, providing guidance on technique. This distinction is well captured by the classic separation between *Knowledge of Results* (KR) and *Knowledge of Performance* (KP). KR concerns the outcome, while KP concerns the execution (e.g., saw angle, wrist posture, pace) [46]. From a motor learning perspective, providing KP-oriented feedback to refine execution is particularly valuable during the associative stage of the Fitts and Posner model [14], where the learner has grasped what the task entails, but must improve how it is executed.

To address these challenges, HCI and adjacent fields have explored technology-assisted skill learning, ranging from human skill capturing [16] and tools for improving motor skills [54] to systems for automated assessment of performances [3, 50, 55]. These systems allow for frequent and detailed monitoring of performances and improvements over time. Despite their potential, the current generation of skill-assessment systems has significant limitations. First, systems that model skills through explicit rules (rule-based approaches) require significant time investment in manual modeling to adapt to new skills [18]. Second, systems that assess performances by comparing them to previously captured executions (template-based approaches) often require hundreds of demonstrations, making them time-consuming to set up and train [31], or do not consider expert variability, making them unable to handle valid variations in execution styles [55].

Recent advancements in automated performance assessment have adopted deep-learning models [26, 51, 59, 61, 62, 64, 65] to extract relevant features from specific sensor modalities (e.g., video recordings, skeleton data) and map them to discrete grades, ranks, or continuous scores [63]. Video-based approaches are the most widely studied, but several limitations remain [63]. First, deep-learning models are typically data-hungry, requiring large datasets to generalize effectively. Although transfer learning is commonly used to mitigate this limitation by leveraging pre-trained backbones (e.g., ResNet [19], C3D [53], I3D [9], VST [32]) from large-scale image or action recognition datasets, domain shifts can limit transferability, requiring computationally intensive fine-tuning on substantial domain-specific datasets. Second, providing interpretable and actionable Knowledge of Performance (KP) feedback remains challenging, with limited explainability undermining user trust. Third, most approaches rely on a single sensor modality, potentially overlooking relevant information from complementary sensors. This highlights an opportunity for multi-modal approaches combining diverse sensor modalities, which could provide a more comprehensive understanding of motor skill performance and enable more detailed and targeted assessments.

In this work, we present Pro-Skill, a system for automated, post-hoc assessment of manual craftsmanship skills that focuses on Knowledge of Performance (KP). To our knowledge, Pro-Skill is the first skill-assessment platform to create an expert model of complex manual craftsmanship skill variability from fewer than 10 demonstrations, making the ad-hoc creation of expert models practical in training contexts. Our approach offers an extensible multi-modal solution for capturing all aspects that are important in a craftsmanship skill, including body posture tracking, eye tracking, tracking of tools and workpieces, sound, RGB video, and exerted force (Figure 1a). After capturing 6-10 task executions by an expert (Figure 1b) that reflect the natural variations in their execution, Pro-Skill semi-automatically builds a model capable of assessing future performances of the same task (Figure 1c). Upon a trainee's completion of the same task (Figure 1d), Pro-Skill compares the execution with the expert model (Figure 1e), and the trainee receives personalized feedback, including a detailed performance analysis of the task and parameter-specific analyses that offer insight on how to improve performance. Rather than providing continuous real-time corrections, which can promote learner dependency and are cautioned against by the Guidance Hypothesis in motor learning research [45, 47], Pro-Skill delivers feedback post-hoc, giving trainees space to execute the task uninterrupted and reflect on their performance afterwards (Figure 1f).

In this paper, we contribute (1) an extensible, multi-modal sensing platform that instruments real-world craftsmanship activity, combining body-posture, tool, and workpiece tracking with complementary modalities such as eye gaze, exerted force, audio, and video. Built on the Lab Streaming Layer [1] (LSL) standard, it time-synchronizes these streams and can be reconfigured or extended with additional sensors to suit the requirements of specific skills and tools. (2) A novel approach for modeling performances and natural variations of craftsmanship skills using Trajectory parameterized Probabilistic Principal Component Analysis (TraPPCA) [1]. TraPPCA is a learning from demonstration approach adopted from the robotics community, presented as a method for learning human and robot motions from a limited number of human demonstrations [1]. We adopted and further refined this method for modeling craftsmanship skills. As TraPPCA is intended for modeling trajectories, our

[1] https://labstreaminglayer.org

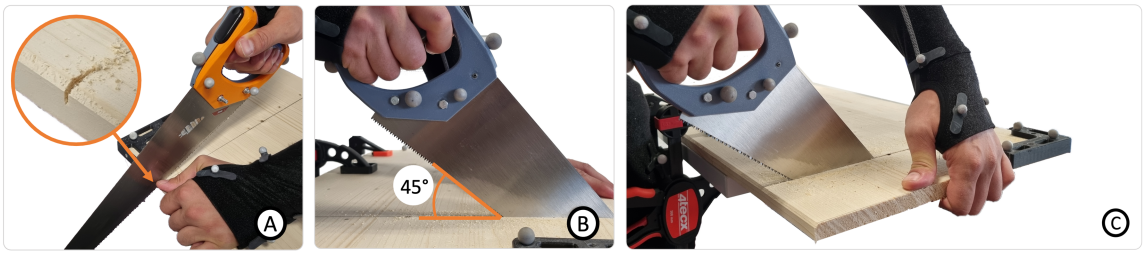


Fig. 2. Making a straight crosscut with a handsaw. (a) Starting the cut: the expert establishes a groove using short, controlled strokes, often guided by a finger of the non-dominant hand. (b) Main cutting phase: the 'full' saw length is used in rhythmic back-and-forth strokes at an angle of roughly 45°, while the non-dominant hand stabilizes the board. (c) Completing the cut: the offcut is supported to prevent snapping, ensuring a clean finish.

approach is designed for and expected to suit tasks with inherent monotonic progress, such as time-based or spatial progress. (3) Performance assessment methods for a detailed comparison between new task executions and the learned TraPPCA model. This assessment provides users with feedback on their performance relative to the expert's performance throughout the task, as well as parameter-specific feedback, for example, on body posture or execution speed. To evaluate the validity of our approach, we compare Pro-Skill's assessment of a manual woodworking task involving a handsaw against expert human judgments. We also illustrate the system's reconfigurability through proof-of-concept demonstrations of two additional tasks.

2 Walkthrough: Cutting a Wooden Board With a Handsaw

We illustrate the workings of Pro-Skill through the example of a basic woodworking skill: making a straight crosscut in a wooden board (40 cm wide, 15.7 inches; 1.8 cm thick, 0.71 inches) using a traditional handsaw. Although seemingly straightforward, executing this task effectively requires a skillful performance, involving fine motor skills and eye-hand coordination.

The process begins with starting the cut, where experts establish a shallow groove (scoring) by making a few short, controlled strokes. To guide the saw accurately, many practitioners position the fingers of their non-dominant hand close to the traced line for support (Figure 2a). Once the scoring is established, the main cutting phase begins. During this phase, the full length of the saw is used in controlled, rhythmic back-and-forth motions, with additional force typically applied during the push stroke. These repetitive movements should be performed at a steady pace, with the saw held at approximately 45° to the workpiece (Figure 2b). The non-dominant hand now pushes on top of the board at a few centimeters distance from the cutting line to provide stability. Throughout the cut, it is crucial to visually monitor the intersection of the saw teeth and the board to maintain alignment with the traced line. Experts continuously adjust the angle of the saw to correct deviations in real-time. As the cut nears completion, the non-dominant hand supports the offcut to prevent it from snapping off prematurely, ensuring a clean finish (Figure 2c). Throughout this process, it is important that the elbow is held close to the hip, and the user should look straight down the saw to avoid unintentional tilting, which can lead to deviation from the desired path.

In traditional training settings, mastering this skill typically involves guidance from an experienced woodworker. The instructor demonstrates the technique and later closely observes each trainee's execution, evaluates the outcomes, and provides personalized feedback. While effective, this approach is challenging to scale to larger groups, as it demands substantial time and attention from the instructor.

The following walkthrough demonstrates how Pro-Skill automates skill assessment for training in this woodworking skill and offers feedback for users to improve their performance:



Fig. 3. Step 1: Capturing expert demonstrations. Pro-Skill records the full-body pose and the pose of the saw and the board using an optical tracking system. A single expert performs 6–10 demonstrations, from which key skill parameters are extracted.

Step 1. Capturing Expert Demonstrations. To capture this skill, Pro-Skill's capturing solution is first configured to record the full body pose, as well as the pose of the saw and the board using an optical tracking system (Figure 3). The expert carpenter then records 6-10 performances of the task. These recorded demonstrations capture the natural variations in the sawing technique for this task. From this raw recording of human behavior, skill-specific parameters are extracted. For the handsaw, these parameters include the angle between the saw and the workpiece (saw angle), the length of the saw used during cutting (saw amplitude), the distance between hip and elbow, the stability of the elbow (elbow deviation), the deviation of the cut from the traced line, the position of the non-dominant hand relative to the traced line, and the speed of the sawing progress.

Step 2. Creating an Expert Model for the Training Task. In this second step, Pro-Skill creates a probabilistic skill model using the extracted parameters. This model creation takes only a minute, during which Pro-Skill leverages data from all 6-10 demonstrations of the expert, capturing the variability of all parameters and relations between parameters. If all the expert's demonstrations are very consistent with respect to the saw angle, the model reflects this with minimal allowed variability for this parameter. To align different demonstrations, TraPPCA requires a progress metric. In this training task, we use spatial progress, the advancement of the saw along the cutting path, rather than time, as the cutting speed profile over the execution may vary between the expert and trainee. This alignment ensures that parameters are compared meaningfully across performances, regardless of tempo variations. The resulting model captures the evolution and natural variability of all parameters along the cutting trajectory. As shown in Figure 4, for instance, the saw amplitude is lower at the beginning of the cut, reflecting the short, controlled strokes used for scoring the wood in the initial phase of the task.

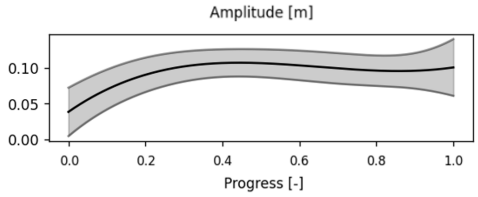


Fig. 4. Step 2: Creating an Expert Model for the Training Task. Pro-Skill uses TraPPCA to build a probabilistic model from 6–10 expert demonstrations. Aligned by progress, the model captures both patterns and natural variability of key parameters (e.g., shorter strokes at the start).

Step 3. Recording and Assessing Trainee Executions. Trainees subsequently perform the sawing task and are recorded using the same capturing system as in step 1. After each performance, Pro-Skill compares the new execution with the expert model. As shown in Figure 5, Pro-Skill allows for precise analysis of the performance over the spatial progress for each parameter and clearly shows when a parameter is outside the valid range compared to the expert executions. During playback of the recording, animated icons show the user when a parameter significantly diverges from the expert model and provide a first step towards corrective coaching based on how the execution deviates from the expert model (e.g., too shallow saw angle, excessive deviation to the right of the traced line, or insufficient progress speed) (Figure 5c). This feedback enables trainees to pinpoint which aspects of their performance deviate from the expert model and require improvement at what stage in the task, with simple but concrete recommendations indicating how the performance can be improved.

3 Related Work

Our work builds on four areas of prior research: capturing skills, modeling performances, assessing motor skills, and technology-assisted learning.

3.1 Skill Capturing

Capturing manual craftsmanship skills is challenging because it involves fine-grained (embodied) movements. Optical motion capture provides sub-millimeter precision but is costly, cumbersome, and non-portable. Depth cameras (e.g., Kinect, RealSense) offer low-cost, markerless tracking at reduced accuracy, and have been widely applied in rehabilitation [21], sports [36], and classroom-scale studies [2]. Professional inertial systems such as Xsens enable portable full-body pose estimation [43], while recent HCI research has shown that consumer-grade inertial measurement units (IMU) embedded in smartwatches and mobile devices can provide cost-effective, mobile motion capture [12, 35]. For fine hand articulation, glove-based devices such as StretchSense [2] and Manus[3] capture finger motion. Other work has explored instrumenting the arm or wrist with systems such as DiscoBand [11], or exploiting inertial, acoustic, and electromagnetic signals available on commodity smartwatches [27, 28]. Other modalities extend beyond motion, such as eye tracking, which can differentiate novices from experts in woodworking [60], using commercial head-worn devices such as Tobii Pro [4] or Pupil Labs[5]. More recently, Project Aria combines gaze with spatial mapping to capture skilled human activity [13, 16]. Gong et al. combine wearables (e.g., eye tracking and smartwatches) with instrumented tools and environmental sensors to capture making activities [15], enabling only coarse skill assessment (novice vs. expert) rather than actionable feedback on task execution. In industrial tasks, sensors are mounted directly on tools, for example, IMUs on a welding torch, and have proven effective for skill capture and evaluation [30]. While other approaches infer human–tool interactions without direct instrumentation [29]. Pro-Skill does not commit to one sensing modality. Instead, it offers an extensible framework for integrating the most practical or precise sensors for a given skill.

3.2 Motor Skill Assessment

Computational approaches for assessing motor skills are generally categorized as *rule-based* or *template-based*. Rule-based approaches define execution constraints manually using explicit rules that describe correct execution [22], which enables structured and interpretable assessments of learner executions. However, authoring the rules is labor-intensive, requiring domain experts to identify key execution features and encode them into formal constraints. Additionally, rule-based methods often struggle to accommodate individual differences and execution variability, limiting their generalizability in dynamic environments [18]. Template-based approaches compare observed executions to reference templates, using similarity metrics such as Dynamic Time Warping (DTW) [20, 48] or learned models, including Artificial Neural Networks (ANNs) [24] and Hidden Markov

[2] https://stretchsense.com/ [3] https://www.manus-meta.com/ [4] https://www.tobii.com/products/eye-trackers/wearables/tobii-pro-glasses-3
[5] https://pupil-labs.com/

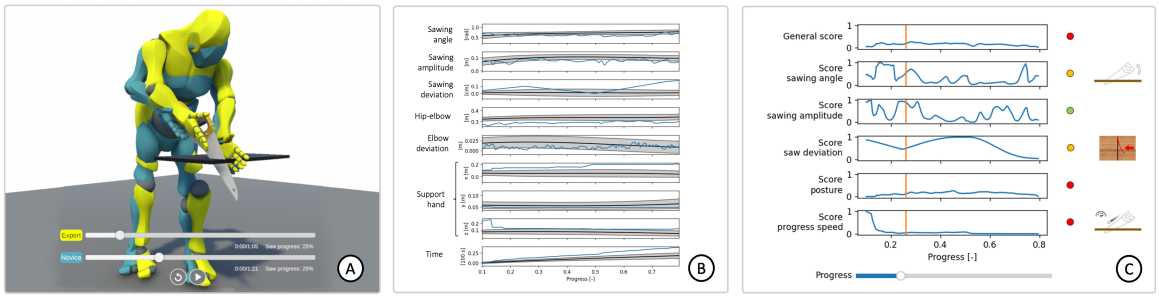


Fig. 5. Step 3: Assessing trainee executions. (a) Trainee and expert demonstrations are visualized in an overlay. (b) Trainee execution is aligned with the expert model, showing how their parameters evolve and differentiate over progress. (c) Parameter-specific and combined scores (e.g., saw angle, saw amplitude, saw deviation, posture, progress speed) over progress highlight where the trainee performance diverges from the expert model, with animated icons providing a first step towards corrective coaching during playback of the recording.

Models (HMMs) [8]. Machine learning is frequently used for movement recognition and assessment [31]. These approaches offer greater flexibility and automation than rule-based methods, but they often require large datasets and may struggle to adapt to new execution styles without retraining. Instead, Pro-Skill introduces a probabilistic learning-from-demonstration approach called Trajectory parameterized Probabilistic PCA (TraPPCA) from robotics [1] into human skill modeling, which enables Pro-Skill to learn and assess motor skills, including natural expert variability across multiple parameters, from limited demonstrations.

Other learning-from-demonstration approaches in robotics also encode, generalize, and reproduce trajectories from limited demonstrations. Dynamic Movement Primitives (DMP) [23] encode trajectories using stable nonlinear differential equations, but cannot capture (co)variance across demonstrations, making them unable to encode expert variability. Probabilistic Movement Primitives (ProMP) [37], a probabilistic extension of DMP, and TraPPCA can produce probabilistic models with similar characteristics, including an explicit progress dependency, and have previously been compared for exoskeleton-based assistance [7]. Gaussian Mixture Models (GMM) [6] encode demonstrated trajectories by modeling their distribution as a mixture of Gaussians, from which trajectories can be inferred using Gaussian Mixture Regression (GMR). Unlike ProMP and TraPPCA, GMM only includes progress information when introduced as an explicit input.

3.3 Technology-Assisted Skill Learning

Augmented and virtual reality (AR/VR) environments, combined with motion tracking, have been widely explored to support motor skill acquisition. Many systems provide structured guidance by analyzing execution and offering real-time feedback. For example, MotionMA [55] builds expert movement models to generate personalized feedback, while YouMove [3] uses augmented mirror interfaces that progressively reduce on-screen prompts to foster autonomous practice. Remote collaboration systems, such as LoKI [52], extend this principle to telepresence, enabling trainers to demonstrate skills in real time. In rehabilitation, systems like Physio@Home [50] guide patients through adaptive motion exercises, prioritizing therapeutic outcomes rather than task-specific skill development. Similarly, sports training platforms have applied AR/VR to improve domain-specific skills such as golf [34], snowboarding [38], or billiards [17]. Surgical training has also benefited from VR simulation, with studies showing improvements in laparoscopic proficiency [5, 25]. Visual guidance and physical constraints can also support dimensional accuracy in manual crafting tasks [42]. Unlike these systems, which emphasize real-time guidance, Pro-Skill focuses on detailed post-execution analysis. This offline approach allows uninterrupted practice and supports reflective learning, enabling a deeper understanding of performance trends (KP).

4 Pro-Skill System: Multi-Modal Skill Capturing

We now introduce the Pro-Skill system, beginning with its extensible multi-modal capturing approach for recording the diverse signals involved in manual craftsmanship tasks. Mastering manual skills requires a broad range of motor abilities: maintaining body posture for stability and control, handling tools and materials with precision, guiding movements through hand–eye coordination, and using tactile feedback to sense texture, pressure, or resistance.

4.1 Multi-Modal Capturing of Manual Skills

We developed a multi-modal capturing solution that supports and can be extended with many data streams to capture all facets relevant for assessing a specific craftsmanship skill. Our capturing solution supports the Lab Streaming Layer (LSL) framework[6], which allows for easy temporal synchronization between many concurrent data streams. The current implementation supports optical motion tracking of rigid bodies and full-body pose estimation using OptiTrack or Qualisys, IMU-based body and tool tracking with Xsens, eye tracking with Tobii glasses, audio and RGB video recording, finger tracking with Manus gloves, and integration of off-the-shelf sensors (e.g., pressure or IMUs) via an ESP32 development board, for which we also implemented synchronization via LSL. These modalities span a spectrum that trades measurement accuracy against portability, and the most appropriate choice depends on the craft or task being captured. In this work, we prioritize accuracy and use optical infrared tracking, at the cost of requiring line-of-sight and a calibrated capture volume. More portable alternatives such as IMU-based body and tool tracking are not affected by occlusions and are better suited to less controlled environments. Advances in camera-based (RGB/RGB-D) markerless pose estimation are also expected to make these systems increasingly viable, potentially reducing the need for body-worn sensors. We further discuss these tradeoffs in Section 10.

Rather than committing to a fixed sensor setup, our multi-modal capturing system is designed to be reconfigurable and extensible, allowing for the tailoring of sensing modalities to the specific requirements of different tasks, skills, or training scenarios. For the handsaw covered in our Walkthrough (Section 2), this includes the optical tracking system OptiTrack to precisely track the body posture and the position and orientation of the saw and wooden board. An IMU sensor embedded in the handle of the saw will be used for sensor fusion, aiming to improve the robustness of the saw's orientation measurements against occlusions.

4.2 Skill-Specific Parameter Extraction

Pro-Skill's capturing system records a wide variety of data, but as not all recorded data is directly relevant for assessing the performance of a craftsmanship skill, this data needs to be converted into skill-specific parameters. Defining these skill-specific parameters is a one-time effort for every new skill in Pro-Skill and can be based on input from experts and literature on craftsmanship skills. We use the following skill-specific parameters for the handsaw: saw angle, saw amplitude, deviation from the traced line, distance between hip and elbow, stability of the elbow (elbow deviation), position of the non-dominant hand, and the speed of the sawing progress. Not all these parameters are directly available via the multi-modal capturing system and require additional calculations. For example, to determine the speed of the sawing progress, the sawing progress itself must first be determined, which requires the intersection point between the saw's cutting edge and the wooden board to be estimated and projected onto the traced line. Then, to express the speed of the sawing progress, the evolution of time over the sawing progress is considered (see Figure 5b). Also, the saw amplitude is determined from the range traversed by this intersection point along the saw's cutting edge.

For unknown skills and tools, these parameter calculations have to be defined once in Pro-Skill, but can then be reused for other tasks using the same tool. To avoid additional transformations of data, it is important that the

[6] https://labstreaminglayer.org

coordinate systems across tasks are defined consistently. The skill-specific parameters for the handsaw could, for example, be reused in tasks on making miter (Figure 6b) or bevel (Figure 6c) cuts, or even a downwards cut when the board is positioned on its narrow edge (Figure 6d).

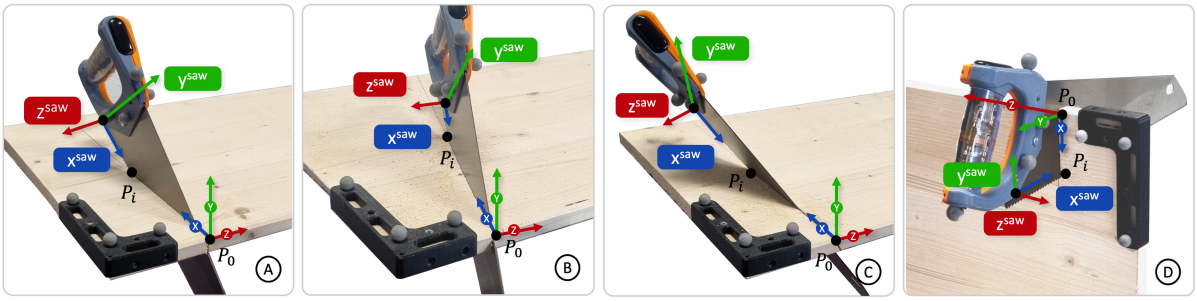


Fig. 6. Coordinate system definitions for the handsaw and the wooden board used in parameter extraction. The board axes x, y, z are defined at the reference point P_0, while the saw axes $x^{saw}, y^{saw}, z^{saw}$ are aligned with the tool. P_i marks the intersection of the saw with the board. (a) Straight rip cut, (b) miter cut, (c) bevel cut, and (d) downwards cut with the board positioned on its narrow edge. Consistent coordinate alignment allows the same parameters to be reused across related tasks.

5 Modeling of Skill Performance

To capture the main patterns and allowed variations of all parameters relevant for the considered craftsmanship skill, Pro-Skill constructs a probabilistic expert model using Trajectory parameterized Probabilistic Principal Component Analysis (TraPPCA) [1]. TraPPCA was selected for its ability to learn a representative probabilistic trajectory model from limited demonstrations, encoding demonstrated variability and incorporating an explicit progress relation. While this progress relation introduces certain challenges, it allows for direct mapping to specific execution phases, facilitating targeted feedback in the subsequent performance assessment (Section 6). The remainder of this section describes the concept of TraPPCA and its use within Pro-Skill to construct expert models of motor skills (Figure 7).

5.1 Background on TraPPCA

TraPPCA [1] is a learning-from-demonstration approach closely related to a simplified variational autoencoder, capable of learning a representative probabilistic trajectory model from a small number of demonstrations (e.g., 5

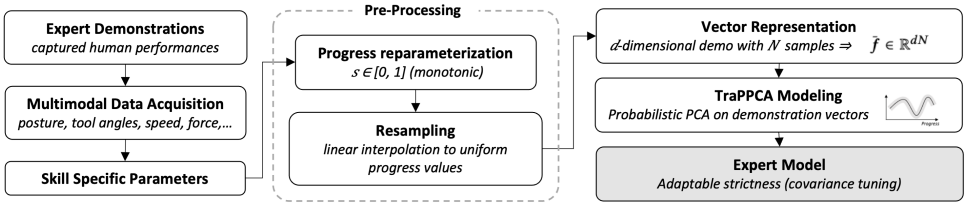


Fig. 7. Workflow for creating an expert model with TraPPCA. Expert demonstrations are first captured and pre-processed, using a progress variable to align them. The demonstrations are then modeled with TraPPCA, for which each demonstration is first rearranged in a single column vector. The result is a probabilistic expert model that captures both the mean trajectory and the allowed variability of the considered parameters. Model strictness can be tuned by adjusting the covariance.

demonstrations in [57] for a robotics application). Traditionally, TraPPCA has been used to learn robot motions and force and torque requirements of a robot based on human demonstrations [56, 57]. In addition, it has been used to model human motions for improving human-robot interaction [1, 7].

Unlike deep learning techniques, which require large datasets to generalize across diverse inputs, TraPPCA leverages assumptions to simplify the modeling process and reduce the required dataset. First of all, TraPPCA relies on an adaptation of Probabilistic Principal Component Analysis (PPCA) [4], inheriting its main assumptions. PPCA assumes that the modeled data is generated by a linear transformation from a lower-dimensional latent space, approximating the distributions by Gaussians and assuming isotropic Gaussian noise. However, in contrast to traditional PPCA, which considers independent observations of data points, TraPPCA considers independent observations of complete trajectories. To properly align these trajectories in the modeling process, TraPPCA requires that each demonstration is expressed with respect to a progress variable. This progress expresses the fraction of completion of the execution and must be monotonically increasing for each training demonstration. By aligning the training demonstrations based on this progress, TraPPCA is able to determine the correlations between the different progress samples and the different parameters over a larger horizon [1], resulting in a model with an explicit progress dependency.

Complementary to prior uses in robotics, this work uses TraPPCA for constructing expert models of human performances in craftsmanship activities. To satisfy the key requirement of TraPPCA, all recorded parameters relevant to skill execution must be expressible as a function of a monotonically increasing progress variable. The recorded parameters of each performance, therefore, first undergo pre-processing (Figure 7).

5.2 Pre-Processing of Demonstrations

First, each captured demonstration is reparameterized in function of a progress variable $s_i \in [0, 1]$, with $s_1 = 0$ corresponding to the first sample and $s_D = 1$ to the final sample of the recorded demonstration. As discussed in Section 5.1, the reparameterization using a progress variable ensures the alignment of demonstrations. Common choices for the progress variable are normalized time or spatial progress. Second, each demonstration is resampled using linear interpolation, such that the samples of all training demonstrations are located at the same, uniformly distributed progress values.

5.3 Creating Expert Craftsmanship Model

To create a TraPPCA model of a craftsmanship task, a d-dimensional demonstration, consisting of N samples, is rearranged in a single column vector $\bar{\boldsymbol{f}} \in \mathbb{R}^{dN}$, as described by Aertbeliën and De Schutter [1]. Here, d represents the number of extracted parameters of a demonstration (e.g., posture information, angles, speed). Applying PPCA on the single column vectors of all demonstrations results in the following linear latent variable model:

$$\bar{\boldsymbol{f}} = \bar{\boldsymbol{H}}\boldsymbol{\chi} + \bar{\boldsymbol{b}} + \bar{\boldsymbol{\epsilon}}. \tag{1}$$

Here, $\bar{\boldsymbol{H}} \in \mathbb{R}^{dN \times m}$ is a matrix containing m basis functions, which encode the variability in the demonstrations and are weighted by the latent variable vector $\boldsymbol{\chi} \in \mathbb{R}^m$, distributed as $\boldsymbol{\chi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_m)$. Furthermore, $\bar{\boldsymbol{b}} \in \mathbb{R}^{dN}$ represents the mean trajectory of the model and $\bar{\boldsymbol{\epsilon}} \in \mathbb{R}^{dN}$ is a Gaussian noise vector, with $\bar{\boldsymbol{\epsilon}} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_{dN})$.

For a given number of latent variables m, the values for $\bar{\boldsymbol{H}}$, $\bar{\boldsymbol{b}}$ and σ are calculated from the provided training demonstrations employing maximum likelihood, using the eigenvalues and eigenvectors of the sample covariance matrix of $\bar{\boldsymbol{f}}$. This maximum likelihood approach and the choice of m are further explained in [1].

Afterwards, the strictness of the expert model can still be adjusted by increasing the covariance of the model with an additional covariance matrix, which could be used to adapt the performance assessment for different skill levels. Therefore, future work should investigate intuitive or automated approaches to determine this covariance matrix, which is outside the scope of this work.

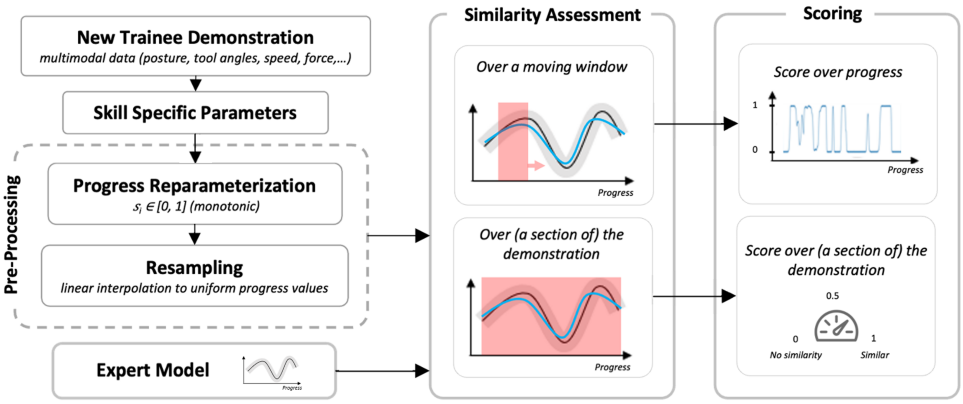


Fig. 8. Workflow for assessing new task executions. After pre-processing, the similarity to the expert model is evaluated either over a moving window or over (a section of) the demonstration. These similarity measures are mapped to scores between [0,1], expressed either as values over progress or as a single overall score.

6 Performance Assessment of Execution

Once an expert model for a craftsmanship task is constructed with Pro-Skill, new task executions are assessed based on their similarity to the expert model. To express this similarity, the deviation between the new execution and the expert model is determined using the squared Mahalanobis distance, a distance metric that takes into account the allowed variations in the expert TraPPCA model. Afterwards, these distances are mapped to interpretable scores that highlight deviations with respect to the expert model, enabling Pro-Skill to provide personalized feedback. This section discusses the different aspects of the performance assessment performed by Pro-Skill, shown in Figure 8.

6.1 Single-Parameter Performance Assessment

To evaluate the performance of a trainee's execution over progress values $s_i \in [s_{\min}, s_{\max}]$ with respect to a single parameter k, a score $c_k \in [0, 1]$ is determined. This score is based on the similarity between the parameter execution $\boldsymbol{f}_{\text{obs},k} \in \mathbb{R}^{N_k}$ after pre-processing (see section 5.2) and the corresponding part of the expert model $\bar{\boldsymbol{f}}_k \sim \mathcal{N}(\bar{\boldsymbol{b}}_k, \bar{\boldsymbol{\Sigma}}_k)$, with $\bar{\boldsymbol{b}}_k \in \mathbb{R}^{N_k}$ and $\bar{\boldsymbol{\Sigma}}_k \in \mathbb{R}^{N_k \times N_k}$. First, the squared Mahalanobis distance is calculated between $\boldsymbol{f}_{\text{obs},k}$ and $\bar{\boldsymbol{f}}_k$:

$$D_k = \text{SMD}[\boldsymbol{f}_{\text{obs},k}, \bar{\boldsymbol{f}}_k] = (\boldsymbol{f}_{\text{obs},k} - \bar{\boldsymbol{b}}_k)^T \bar{\boldsymbol{\Sigma}}_k^{-1} (\boldsymbol{f}_{\text{obs},k} - \bar{\boldsymbol{b}}_k), \tag{2}$$

which provides a similarity measure that accounts for the expert variability and the correlation between the progress samples encoded in the expert model. Then, the distance D_k is mapped to the score c_k using the following heuristic:

$$c_k = \exp(-0.5 \cdot D_k / N_k), \tag{3}$$

where 0 indicates poor correspondence and 1 indicates great correspondence to the expert. The heuristic is motivated by, but not a unique consequence of, the properties of the squared Mahalanobis distance. Specifically, the heuristic is inspired by the exponential term of the Gaussian probability density function associated with the squared Mahalanobis distance. Since D_k follows a χ^2-distribution with N_k degrees of freedom, its expected value equals N_k, the expected value of the χ^2-distribution. Therefore, to reduce the dependence of the score on the number of progress samples, D_k is divided by N_k. Nevertheless, the choice of N_k remains important, as it also determines the resolution of the comparison.

6.2 Combined Multi-Parameter Performance Assessment

To evaluate the performance of a trainee's execution over progress values $s_i \in [s_{\min}, s_{\max}]$ considering multiple parameters $k = 1 \dots d_c$, the squared Mahalanobis distance D_k is first calculated for each parameter individually (see section 6.1). Then, the mean squared Mahalanobis distance is calculated over the parameters:

$$D_{1\dots d_c} = \frac{\sum_{k=1}^{d_c} D_k}{d_c}, \tag{4}$$

compensating for the number of parameters. This computation assumes that the parameters are independent—a reasonable assumption, since including strongly correlated parameters would risk rewarding or penalizing the same aspect of the execution multiple times. Finally, the heuristic mapping of (3) is employed to obtain a score $c_{1\dots d_c} \in [0, 1]$, replacing D_k with $D_{1\dots d_c}$. Here, scores 0 and 1 again indicate poor correspondence and great correspondence to the expert, respectively.

6.3 Performance Assessment Over Progress

To evaluate the performance of a trainee's execution at each point throughout the execution, the performance assessments of sections 6.1 and 6.2 are performed over a (small) moving progress window, providing score signals in function of progress. By employing a moving window, the resulting score signals are less susceptible to minor, uninformative changes, making the score signals less noisy. Also, the moving window allows for accounting for local correlations between progress samples encoded in the expert model. The interpretation of these score signals is the same as before, with scores 0 and 1 indicating poor correspondence and great correspondence to the expert, respectively. Figure 5c shows examples of these progress-dependent score signals for a woodworking task involving a handsaw and illustrates how these signals can be used to provide personalized feedback. It contains score signals for individual parameters (sawing angle, sawing amplitude, saw deviation and progress speed), combined score signals over multiple parameters (posture, considering distance between hip and elbow, elbow deviation and 3D position of support hand), and a general score signal over all considered parameters.

7 Handling Non-Monotonic Task Progress

A core requirement of modeling craftsmanship skills with TraPPCA is that all parameters relevant to skill execution can be expressed as a function of a monotonically increasing progress variable. While this assumption holds naturally for common progress variables based on time and arc length — that is, the cumulative distance traveled along a path, it becomes more delicate when spatial progress is used as the progress variable. In tasks where spatial position (e.g., the advancement of a saw along a cutting line) defines progress, human behavior may not always follow a strictly forward trajectory. In such cases, non-monotonic progress may occur, where users temporarily move backward in the task space. This behavior is oftentimes natural or even intentional, for example, when confronted with friction, hesitation, or the need to recover from minor errors in craftsmanship tasks. In this section, we distinguish between two common forms of non-monotonic progress: retractions, where a user briefly moves back before continuing forward, and corrections, where users revise earlier parts of the task.

Retractions occur when the user goes backward from the current point in the execution and then continues forward without modifying the previous trajectory. For a handsaw (Section 2), this may happen when the saw becomes stuck. For a handheld drill, retractions may be caused by tool jamming or may be intentional, for instance, to release accumulated chips. When Pro-Skill records parameters as a function of a spatial progress variable, it is important to detect such retractions to prevent logging repeated samples for the same spatial positions. These repeated samples do not reflect a valid second execution of that part of the task, as material cannot be cut or drilled twice along the exact same path in a meaningful way. As a result, the additional samples are not only redundant, but may also distort the assessment by introducing misleading data. Pro-Skill detects retractions by

identifying significant decreases in spatial progress, followed by a return to approximately the same location where the retraction began. Samples recorded during these retractions are excluded from performance evaluation in the nominal assessment (Section 6). However, retractions, especially when frequently encountered, can be an indicator of skill level. Therefore, Pro-Skill logs each instance of retraction and presents this information as a separate metric in the performance assessment.

During *corrections*, users also go backward in the task, but unlike retractions, they do so to correct or recover from an error. For the handsaw (see Section 2), this may occur when the saw deviates too far from the traced line, and the user decides to move back and restart the cut from a more accurate location on the traced line. During such corrections, multiple recordings exist for the same spatial progress values, representing different attempts at the same part of the task. Pro-Skill must, therefore, detect corrections to ensure that all recorded samples at overlapping spatial progress are stored correctly. Each correction is assessed in the same way as a nominal execution (see Section 6), enabling both the original and revised attempts to be analyzed during offline performance evaluation.

The above retraction and correction handling addresses occasional non-monotonic behavior within tasks that have an inherently linear structure. We discuss the implications of the monotonicity requirement for the scope of tasks Pro-Skill can model in Section 10.

8 Evaluation

To evaluate both the effectiveness of TraPPCA expert models in capturing craftsmanship and the accuracy of Pro-Skill's automated assessments, we conducted a study comparing Pro-Skill's evaluations to manual assessments by human experts. Specifically, we compared expert judgments of trainees' task performances with Pro-Skill's assessments of those same performances, using expert models trained on demonstrations from those same experts. The study was approved by the university's Social and Societal Ethics Committee (SMEC REC/SMEC/2024-2025/25 CS).

8.1 Craftsmanship Task and Apparatus

For this study, we use the basic woodworking task covered in the Walkthrough (Section 2): making a straight cross-cut, along a traced line, on a solid pine wooden board (40cm/15.7inches wide and 1.8cm/0.71inches thick) using a traditional handsaw (Bahco ProfCut saw). The capturing setup is an optical tracking system (OptiTrack Flex 13 cameras capturing an area of $5 \times 3 \times 2.5$ m (approx. $200 \times 120 \times 100$ in)) that precisely tracks the position and orientation of the saw and the wooden board. Participants wear a motion capture suit to precisely track their full body posture during the task execution.

For the purpose of this study, we added three additional RGB cameras to the setup: one body-worn RGB camera (GoPro HERO 13) and two external RGB cameras to capture the performances from both first-person and third-person perspectives. These additional camera streams are also synchronized using the Lab Streaming Layer (LSL) framework as discussed in Section 4. They are, however, only used during the manual assessment by human experts and not by Pro-Skill for the automated assessment.

8.2 Expert Data Collection Procedure

To create a representative expert model for the handsaw task, we recruited four subjects (all male, ages 29–40; height $M = 179.5$ cm, $SD = 8.4$ cm, range 167–185 cm) who are highly skilled in using a handsaw and instructed them to perform the handsaw task. Before the start of this study, a short questionnaire was given to report on their woodworking experience. Participants were given minimal instructions about the task to avoid influencing their expert intuition on how to perform the handsaw task. Participants were simply instructed to *"cut along the traced line and follow it as accurately as possible."*. All four experts repeated the task 8 times while being recorded.

8.3 Trainee Data Collection Procedure

We recruited 14 participants (13 male, 1 female, ages 21–55; height $M = 177.4$ cm, $SD = 8.0$ cm, range 159–185 cm). All participants had limited prior woodworking experience. However, their observed performance ranged from absolute beginners to more competent, intermediate-level executions. Their performances on the handsaw task will be evaluated by Pro-Skill and by our four human experts. Participants were given the same basic instructions as the experts (Section 8.2): to cut along the traced line as accurately as possible. Before performing the actual task, each participant completed a brief test run to become familiar with the saw and basic handling. The main task was then carried out once by each participant.

During pilot sessions, some trainees' saw blades bent significantly after getting stuck in the material. Since the saw is tracked as a rigid body, this bending introduced inaccuracies in the saw-board intersection point used to calculate deviation from the traced line, making this parameter unreliable for trainee performances. We therefore added a scanning solution: a ceiling-mounted camera and a green screen behind the board, isolating the cut edge's contour, from which we sample the contour and calculate the offset from the traced line.

8.4 Manual Assessments by Human Experts Procedure

To verify Pro-Skill's automated assessment of trainee performances, we asked the same four experts described in Section 8.2 to manually assess the trainee recordings. Before introducing our evaluation parameters, we first asked the experts what they considered important criteria during a skillful handsaw performance. Because different parameters can be more relevant at different moments of the task, we structured the questionnaire into three phases: the start of the cut (first 10% of the cut), the main cutting phase (middle 70%), and the completion of the cut (last 20%). Experts were asked to list the criteria they considered most relevant in each.

To further validate our selection of parameters, we also asked all experts how they felt about the importance of our six (categories of) skill-specific parameters (saw angle, utilized blade length, deviation from target line, progress speed, body posture, and overall performance). A five-point Likert scale was used, ranging from *not important* to *very important*.

After this, the experts evaluated the performances of the trainees. It is important to note that the experts assessed the trainees' performances only based on RGB videos, consisting of three synchronized recordings from different perspectives, as well as by observing a top-down photo of the resulting cut. Motion tracking data was not available to our experts to ensure their evaluation relied only on visual analysis, similar to how they would typically evaluate in a training environment. The 14 trainee performances were presented to our experts in a different random order. Additionally, the experts were provided with the RGB video recordings from three different perspectives of their own performances, as well as their own top-down photos of the resulting cuts. The human experts were instructed to use their own recordings as a reference for their assessment of trainees.

The experts were asked to assess each of the 14 trainee performances by watching the synchronized video recordings and comparing them to their own performances of the task. Each performance was graded using 5-point Likert-scale questions ranging between *very low similarity* and *very high similarity* to their own performance. The assessment included ratings for the (categories of) skill-specific parameters (saw angle, saw amplitude, deviation from the traced line, progress speed, and body posture) as well as an overall performance rating. The experts were given as much time as needed to perform their evaluations and could pause or replay videos as desired. In addition to these 14 evaluations, one additional evaluation at the start of the study served as a dry run to become familiar with the evaluation procedure.

8.5 Study Results

8.5.1 Expert Validation of Performance Parameters. The experts' responses regarding important criteria for skillful handsaw use align well with the parameter set used in Pro-Skill. Additionally, some alternative criteria were mentioned, such as *'fluent movement throughout the cut'* and *'correct amount of force'*. While these aspects are indirectly related to some of the existing parameters, they could be explicitly included in the future to further improve Pro-Skill. Overall, these findings show that our selected parameters for training Pro-Skill's model are parameters that are also considered by experts during manual performance assessments.

The experts' responses when rating the importance of our selected parameters during each phase of the cut showed that the experts strongly agree with our parameter set. Across all phases, the experts considered our parameters to be important ($M = 3.80$, $SD = 1.17$). In the middle phase, all parameters were rated highly important ($M = 4.61$, $SD = 0.57$), indicating a strong consensus that each contributes to assessing performance. In the start phase, deviation from the traced line ($M = 4.50$, $SD = 1.00$) and body posture ($M = 4.50$, $SD = 0.58$) were emphasized, while progression speed through the cut was rated lowest ($M = 1.75$, $SD = 0.50$). In the end phase, overall importance ratings declined slightly, but body posture remained consistently valued ($M = 4.50$, $SD = 0.58$). These findings further support the relevance of the selected parameters and validate their inclusion in the expert model used by Pro-Skill. Although the system does not explicitly assign different weights to parameters based on the execution phase, such differences are implicitly captured through the modeling of expert variability. Parameters that are less important tend to show greater variability across expert demonstrations, as experts pay less attention to their consistency. As a result, Pro-Skill naturally learns a broader tolerance range for those parameters, meaning trainees are not penalized as heavily for variations in execution style where consistency is less critical.

8.5.2 Validation of Expert Proficiency. To validate the proficiency of the experts, we compared their task execution time and deviation from the traced cutting line with the trainees. The experts completed the task faster and with less variability in duration than the trainees. Each trainee performed the task once ($M = 124.4$ s, $SD = 97.0$ s, range = 36–398 s, $N = 14$). Each expert performed the task eight times; we first averaged within each expert and then summarized the four experts' means. On this basis, the experts completed the handsaw task in $M = 35.8$ s (range = 27–57 s, $N = 4$). Additionally, experts showed lower deviation from the traced cutting line ($M = 0.46$ mm, $SD = 0.55$ mm) compared to trainees ($M = 3.96$ mm, $SD = 3.71$ mm). These differences demonstrate that the experts not only deviated less from the traced cutting line, but also achieved substantially greater efficiency (speed) across repeated executions. Together, these results confirm the expert status of our four participants and ensure a reliable foundation for training the expert models used in Pro-Skill.

8.5.3 Pro-Skill's Assessment of Performances. For this analysis, Pro-Skill's automated assessment was applied to the main cutting phase of the handsaw task. A separate expert model was learned from each expert's demonstrations, using the skill-specific parameters and spatial progress variable described in the Walkthrough (Section 2). To focus on the central portion of the task execution, the segment between 10% and 80% progress was extracted for analysis. For each of the 14 trainees, the same central portion of the task execution was considered. The single-parameter performance assessment (Section 6.1) was used to grade the individual parameters of all trainee performances, including sawing angle, sawing amplitude, saw deviation, and progress speed. The multi-parameter performance assessment (Section 6.2) was used to determine a score for posture, which consists of parameters including hip-elbow distance, elbow deviation, and 3D position of the support hand, and to determine an aggregated score, considering all parameters. These performance assessments result in continuous scores ranging from 0 to 1, which were mapped to the five-point Likert scale used in the manual expert evaluations to allow for direct comparison. To perform this mapping, the range of continuous scores between 0 and 1 was divided into 5 equal parts, each corresponding to an answer on the Likert scale (e.g., score range 0-0.2 corresponds to

Table 1. Inter-rater agreement between the expert rater and the expert TraPPCA model. *Reported* gives the baseline κ with its 95% confidence interval (sampling uncertainty over trainees) and Landis–Koch category. *Bin sensitivity* gives the 95% interval of κ across 1000 bin sets whose edges were jittered by SD = 0.05, and the range of Landis–Koch categories it spans.

	Reported				Bin sensitivity	
Parameter	Cohen's κ	95% CI		Agreement	95% CI	Agreement range
		Lower	Upper		(Lower, Upper)	
Sawing Amplitude	0.429	0.216	0.642	Moderate	(0.38, 0.48)	Fair–Moderate
Saw Deviation	0.789	0.668	0.909	Substantial	(0.76, 0.82)	Substantial–Almost perfect
Posture	0.492	0.278	0.706	Moderate	(0.42, 0.57)	Moderate
Progress Speed	0.854	0.775	0.933	Almost perfect	(0.84, 0.87)	Almost perfect
Sawing Angle	0.643	0.494	0.791	Substantial	(0.59, 0.70)	Moderate–Substantial
Overall	0.849	0.775	0.923	Almost perfect	(0.80, 0.90)	Substantial–Almost perfect

the lowest discrete score 1 and answer *'very low similarity'*). These boundaries were defined in advance and applied consistently across all assessments, rather than being selected based on the collected scores. Because experts may differ in how they use the rating scale, fine-tuning the boundaries for each expert would likely yield higher agreement. We deliberately refrain from this, keeping the mapping consistent, as there is no principled automatic approach to determine such per-expert corrections. Additionally, we assess the sensitivity of the reported agreement to the placement of these boundaries in Section 8.5.4 (Table 1).

8.5.4 Agreement Analysis. To determine the level of agreement between the manual expert evaluations and Pro-Skill's assessment of the trainees, we conducted a set of inter-rater agreement analysis [10] for each of the skill-specific parameters. Because a separate model was trained for each of the four experts and evaluated against that expert's own ratings of all 14 trainees, each parameter's agreement is computed over 56 paired model–expert assessments. These 56 paired assessments are not all independent, as they comprise repeated assessments of the same 14 trainees across the four experts. The resulting agreements are reported both pooled (Table 1) and per expert (Appendix A).

Comparing Pro-Skill's scores to the expert ratings reveals *almost-perfect* agreement for overall performance ($\kappa = 0.85$, 95% CI $[0.78, 0.92]$) and for progress speed ($\kappa = 0.85$, 95% CI $[0.78, 0.93]$). Deviation from the target line ($\kappa = 0.79$, $[0.67, 0.91]$) and sawing angle ($\kappa = 0.64$, 95% CI $[0.49, 0.79]$) both show *substantial* agreement. In contrast, posture ($\kappa = 0.49$, 95% CI $[0.28, 0.71]$) and sawing amplitude ($\kappa = 0.43$, 95% CI $[0.22, 0.64]$) yield only *moderate* agreement.

Per-expert analyses (see Appendix A) show that the overall performance agreement is high for all raters (E1 $\kappa = 0.90$, E2 $\kappa = 0.88$, E3 $\kappa = 0.86$, E4 $\kappa = 0.74$). However, agreement on single parameters varies more substantially across raters. This variability is especially pronounced for parameters that are more difficult to isolate visually. For instance, in the case of sawing amplitude, E2's agreement with Pro-Skill was near zero ($\kappa = 0.14$, with a confidence interval that included 0), while E3 showed almost perfect agreement ($\kappa = 0.85$). In contrast, parameters that can be independently verified, such as deviation from the traced line, which can be assessed by inspecting a top-down photo of the cut (E1 $\kappa = 0.83$, E2 $\kappa = 0.93$, E3 $\kappa = 0.72$, E4 $\kappa = 0.68$), and progress speed, which can be derived from the duration of the recording (E1 $\kappa = 0.83$, E2 $\kappa = 0.79$, E3 $\kappa = 0.88$, E4 $\kappa = 0.89$), show less variation across raters. These results underscore the challenge of analytically assessing individual performance parameters solely from video recordings, even among expert woodworkers.

To assess how sensitive these results are to the boundaries used to map Pro-Skill's continuous scores onto the five-point Likert scale, we recomputed the weighted Cohen's κ across 1000 binning schemes with the inner edges jittered by Gaussian noise ($\sigma = 0.05$, a quarter of a bin width; Table 1). Progress speed, saw deviation,

Table 2. Inter-rater agreement among the four experts for each performance parameter. IRA is the mean pairwise weighted Cohen's κ (quadratic weights), with the 95% confidence interval (CI) representing the overall envelope (minimum lower bound to maximum upper bound) across all six expert pairs. Overall consistency across all raters is evaluated using ICC(3,1) with its 95% CI.

Parameter	IRA				ICC		
	Cohen's κ	95% CI		Agreement	Point estimate	95% CI	
		Lower	Upper			Lower	Upper
Sawing Amplitude	0.722	0.313	1.000	Substantial	0.772	0.577	0.908
Saw Deviation	0.889	0.593	1.000	Almost perfect	0.905	0.804	0.964
Posture	0.631	0.227	0.999	Substantial	0.735	0.521	0.891
Progress Speed	0.871	0.490	1.000	Almost perfect	0.879	0.755	0.954
Sawing Angle	0.689	0.155	1.000	Substantial	0.701	0.474	0.875
Overall	0.854	0.682	0.979	Almost perfect	0.865	0.730	0.948

and overall performance were most stable, while posture, sawing angle, and sawing amplitude varied more. For every parameter, however, the variation induced by shifting the bin boundaries was smaller than the sampling uncertainty already reflected in the reported confidence intervals, indicating that the sensitivity of the reported agreement to the precise placement of the bin edges is low.

8.5.5 Robustness to Body Size. To investigate whether body size influences the assessment, we tested whether trainees whose height differed more from their reference expert were assessed less accurately. Although several skill-specific parameters are expressed as relative measurements and angles, some parameters, such as the sawing angle a person naturally adopts, may still be affected by body build. For all 56 trainee–expert pairs, we correlated the absolute height difference between the trainee and the expert with the absolute discrepancy between the expert's rating and Pro-Skill's score, using Spearman's rank correlation. Across all six parameters, we found no significant association (Spearman r between -0.21 and 0.22, all $p > 0.10$). We note that height is a coarse proxy for body build and that the spread of expert heights is limited, so this is a bounded check rather than a guarantee of body-proportion invariance.

8.5.6 Inter-Expert Agreement. To contextualize the agreement levels reported above, we also assessed how consistently the four experts rated the same trainee performances. Table 2 reports, for each parameter, ICC(3,1) across all four experts and the mean quadratically weighted Cohen's κ across the six pairwise expert comparisons (E1–E2, E1–E3, E1–E4, E2–E3, E2–E4, and E3–E4). Across parameters, ICC values ranged from 0.70 to 0.90, while mean pairwise weighted κ ranged from 0.63 to 0.89, indicating substantial to almost-perfect inter-expert agreement. Agreement was highest for saw deviation, progress speed, and overall performance, and lowest for sawing angle, sawing amplitude, and posture.

8.5.7 Comparison with DTW Baseline. To verify that Pro-Skill improves assessment quality beyond simpler template-based approaches, its assessments were compared to a baseline using Dynamic Time Warping (DTW) [44]. First, each execution was resampled using linear interpolation to match the number of samples used in Pro-Skill's assessments, with samples now uniformly distributed over the execution time. Then, z-score normalization was applied to each parameter using mean and standard deviation derived from the expert executions, compensating for differences in parameter scales and units. Subsequently, pairwise distances between expert executions and between trainee and expert executions were computed using standard DTW with Euclidean distance, both for individual parameters (sawing amplitude, saw deviation, progress speed, sawing angle) and for multiple parameters (posture, overall). To account for variations in warping path length, each DTW distance was divided

Table 3. Inter-rater agreement for performance parameters between expert rater and DTW baseline.

Parameter	Cohen's Kappa κ	95% Confidence Interval		Agreement
		Lower	Upper	
Sawing Amplitude	0.345	0.171	0.518	Fair
Saw Deviation	0.709	0.575	0.844	Substantial
Posture	0.385	0.213	0.557	Fair
Progress Speed	0.678	0.558	0.798	Substantial
Sawing Angle	0.317	0.145	0.489	Fair
Overall	0.336	0.207	0.465	Fair

by the corresponding path length. Finally, scores $c_b \in [0, 1]$ were obtained using an exponential decay:

$$c_b = \exp\left(-\lambda \frac{d_{\text{train}}}{d_{\text{expert}}}\right), \tag{5}$$

where d_{train} is the mean distance between the trainee execution and the expert executions, and d_{expert} is the mean pairwise distance among the expert executions. The scaling factor λ was set to 0.1, such that a trainee whose mean distance to the expert executions equals the mean distance between expert executions receives a score of approximately 0.9. This matches the intended interpretation of the score, where expert-level performance corresponds to a high, but not necessarily perfect, score, as is also the case for Pro-Skill.

The results are reported both pooled (Table 3) and per expert (Appendix B). Comparison with Table 1 shows that Pro-Skill achieves higher agreement with the manual expert ratings than the DTW baseline for all parameters. Improvements in Cohen's κ range from 0.08 to 0.51 (mean 0.21), with the largest improvement observed for overall performance.

8.5.8 Decoupling Training and Rating Experts. Earlier, each expert model was validated against the same expert who provided its training demonstrations, a pairing that could in principle inflate agreement. To test this, we compared each expert model against the ratings of the other experts. This decoupling of the training expert from the rating expert changed agreement only marginally: for overall performance, the weighted Cohen's κ decreased from 0.85 to 0.80, while the largest change for any parameter was 0.07 (Appendix C). This shows that the original pairing did not substantially inflate agreement for the handsaw task.

8.6 Discussion of Study Results

Our evaluation highlights promising alignment between Pro-Skill's automated assessments and human expert judgments. While the overall agreement was strong, particularly for general performance and more easily observable parameters, it is important to reflect on what these results tell us about the nature of performance evaluation in manual craftsmanship. In particular, the varying degrees of agreement across parameters raise questions about how experts interpret and apply evaluation criteria compared to the models used in Pro-Skill.

The recruited experts were clearly very skilled in using a handsaw, but the manual evaluation task, judging isolated aspects of performances, posed challenges. While there was an almost perfect agreement between the human experts and Pro-Skill for overall performance evaluation and parameters that were directly observable (such as cut deviation or progress speed), the experts reported that it was more difficult to judge parameters like posture or amplitude in isolation. As one rater (E1) noted (paraphrased): "When a performance looks wrong overall, it is hard to really judge just the sawing amplitude. My judgment is shaped by the entire action." This highlights the natural tendency of expert practitioners to evaluate craftsmanship holistically, rather than through discrete parameter scoring.

Pro-Skill, by contrast, can evaluate each parameter individually, based on consistent and objective sensor data. The system focuses specifically on how each parameter evolves during task execution, and is unaffected by the overall impression or by other performance parameters. While our system's assessments do not always align with human ratings on every aspect, its consistency and specificity can enhance training by surfacing details that might otherwise go unnoticed. In this way, Pro-Skill can also offer an additional layer of analysis on top of human judgment that can contribute to clearer, more targeted feedback in the development of craftsmanship skills.

The advantage of Pro-Skill over the DTW baseline may be attributed to several methodological features. First, deviations from the expert model are evaluated relative to the expert variability, allowing different levels of variability throughout the execution, while also accounting for correlations between samples. Second, the employed distance measure is dimensionless and independent of differences in units or scale between parameters, such that additional parameter normalization is not required for the performance assessment when combining multimodal parameters. Third, Pro-Skill's reparameterization and resampling with respect to a representative progress variable ensures that each part of the execution contributes equally to the assessment, while execution speed is explicitly represented by an additional parameter rather than implicitly affecting the assessment. Finally, the scoring heuristic is motivated by the properties of the employed distance measure, providing a principled basis for converting deviations into performance scores. In contrast, the DTW baseline offers no analogous basis for the scoring heuristic and the resulting scores depend on a manually selected scaling factor.

Finally, decoupling the expert who provided the demonstrations from the expert whose ratings were used for validation changed agreement only marginally. This suggests that the reported agreement was not primarily caused by matching each model to the same expert who trained it. However, it does not necessarily imply that the demonstrations of the different experts are sufficiently similar to be combined into a single expert model: where valid execution styles differ significantly, the variability of the combined expert model would also include intermediate executions that match none of these styles, allowing them to be scored as acceptable.

9 Proof-of-Concept Demonstrations on Additional Tasks

To illustrate Pro-Skill's reconfigurability beyond the evaluated handsaw task, we applied the system to two additional tasks. These are proof-of-concept demonstrations of the capturing, modeling, and assessment pipeline: unlike the handsaw study in Section 8, they were not assessed against human expert judgments. They show that Pro-Skill can be *configured* for tasks with different parameters, progress variables, and sensor setups, whereas validating the accuracy of the resulting assessments would require studies of the same rigor as our primary evaluation. The two tasks are: (i) a downwards cut into a wooden board positioned on its narrow edge, which demonstrates Pro-Skill's ability to switch between tasks using the same set of skill-specific parameters; and (ii) drilling a hole into a vertically-mounted aluminum plate with a cordless drill, which demonstrates that the system can capture, model, and assess a skill with very different characteristics using a distinct set of sensors.

9.1 Vertical Cuts with Limited Depth

Pro-Skill supports easy switching between related tasks by reusing the same skill-specific parameters, provided that coordinate systems are consistently defined (see section 4.2). To demonstrate this, we applied the system to a new handsaw task: making a partial, vertical cut into a board positioned on its narrow edge (Figure 9a). Aside from defining the coordinate transformations and capturing a small set of new expert demonstrations, no further setup changes were needed.

The resulting expert model captured key features of the new task, such as horizontal saw blade alignment (saw angle close to zero), consistent posture similar to the original task, and adjusted non-dominant hand placement for board stability (Figure 9b). A sample trainee execution showed clear divergence in sawing amplitude and deviation from the traced line, both of which were accurately reflected in Pro-Skill's parameter-specific and

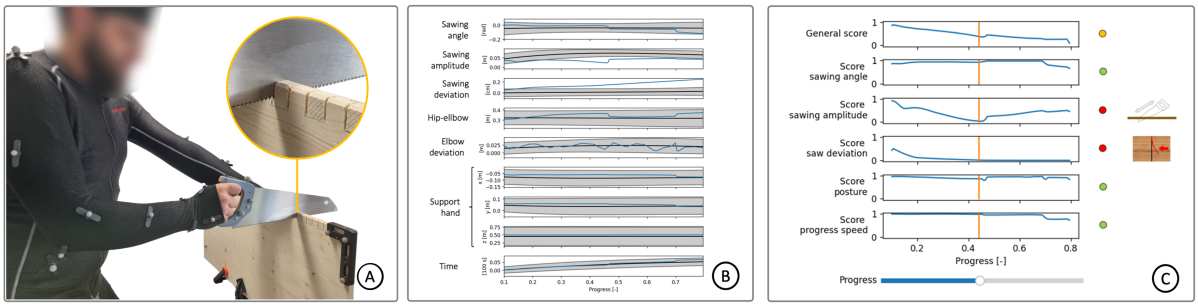


Fig. 9. Vertical sawing task and assessment. (a) Setup with a rotated board. (b) Expert TraPPCA model for key parameters and sample trainee execution. (c) Parameter-specific and combined scores over progress indicating where the trainee performance diverges from the expert model, supporting detailed feedback.

general performance scores (Figure 9c). These observations illustrate how Pro-Skill can be reconfigured for related tasks with minimal additional setup.

9.2 Drilling Holes in Aluminum Plates

To illustrate that Pro-Skill can be reconfigured for a fundamentally different skill, we applied the system to a new task: drilling a hole at shoulder height into a vertically-mounted aluminum plate using a handheld cordless drill. The multi-modal capturing system was reconfigured with a distinct sensor setup (Figure 10a): (1) a load cell behind the plate to measure exerted force, (2) IMU-based Manus gloves for tracking hand and finger motion (e.g., pressing the drill trigger), and (3) the XSens Awinda system for full-body motion and tool tracking via body- and tool-mounted IMUs. This setup differs entirely from the one used in the handsaw task (Section 2), illustrating the reconfigurability and extensibility of Pro-Skill's capturing architecture.

Skill-specific parameters included: exerted force, variation in tool orientation, and the relative positions of the dominant and non-dominant hands with respect to the sternum. Since the aluminum plate was positioned with respect to the user's sternum to promote consistent task execution, defining the hand-position parameters relative to the sternum was a natural choice. An expert TraPPCA model was then constructed (Figure 10b), using normalized time as the progress variable—appropriate given the limited spatial displacement when drilling through thin material. The model shows a gradual build-up of force peaking around 100 N as the drill penetrates the plate, which aligns with typical thrust forces required to drill into aluminum with a handheld drill. The model also shows minimal variation in tool orientation throughout the execution. The hands remain close together, consistent with the dominant hand operating and the non-dominant hand stabilizing the drill.

In contrast, a trainee's performance exhibited an unexpected dip in force around 35% progress and greater tool orientation variation at the beginning and end of the task, though hand positioning remained consistent with the expert model. These observations were clearly reflected in the performance assessment (Figure 10c): a drop in the force-related score after 35%, lower scores for variation of tool orientation at the beginning and end, and consistently high scores for hand position. The general performance score also followed the expected pattern, confirming that the assessment aligns with observed deviations. Together, these observations demonstrate that Pro-Skill can capture and model skills with fundamentally different characteristics and sensor configurations.

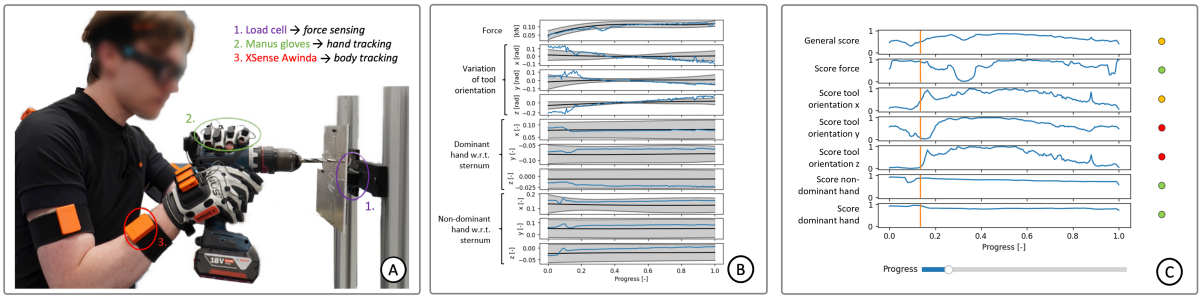


Fig. 10. Drilling task and assessment. (a) Sensor setup (load cell, IMU gloves, IMU-based tracking). (b) Expert TraPPCA model (force ramp, orientation stability, hand positions) and sample trainee execution. (c) Parameter-specific and combined scores over progress, with score dips highlighting suboptimal force profile and orientation stability.

10 Limitations and Future Work

While Pro-Skill offers a solid foundation for automated skill assessment, it also highlights areas for further refinement and exploration.

First, craftsmanship performance is rarely uniform, there are often multiple valid ways to execute the same task, as well as occasional deviations driven by environmental or material conditions. Each expert model in this paper was trained on demonstrations from a single expert and therefore captured one individual's execution style rather than a general notion of expert performance. While TraPPCA can incorporate the variation across a single expert's demonstrations, combining fundamentally different styles into one model would blur the boundary between acceptable and suboptimal performances. In practice, this implies that a trainee is assessed against the model of the instructor whose technique they are taught, or that multiple models, one per accepted style, are maintained, with the closest-matching model used for assessment. A related challenge is that anomalies, such as compensatory actions in response to irregularities (e.g., knots in the wood), may fall outside the nominal variability of a single model and risk being incorrectly penalized. Future work could address these challenges by automatically discovering execution styles and selecting the appropriate model, and by distinguishing intentional expert adaptations from actual performance errors, for example, clustering expert demonstrations into distinct styles in the model's parameter space, building a separate model per style, and assessing a new execution against the closest-matching model. Such flexibility would further enhance the robustness and fairness of Pro-Skill.

Second, the requirement of a monotonically increasing progress variable delimits the class of tasks that Pro-Skill can currently model. Tasks with an inherently linear structure, such as cutting, drilling, and welding or brazing, map naturally onto a spatial or temporal progress variable. In contrast, craftsmanship activities dominated by iterative, non-linear workflows, such as wood carving or surface finishing, involve revisiting areas and refining earlier work in ways that cannot readily be expressed as a function of a single monotonic progress variable. Our handling of retractions and corrections addresses occasional non-monotonic behavior within otherwise linear tasks, but is not designed for workflows where revisiting constitutes the dominant mode of work. Pro-Skill is therefore best positioned to assess linear tasks. Supporting inherently non-linear skills remains an open challenge for future work.

Third, Pro-Skill requires identifying the skill-specific parameters for evaluating a craftsmanship skill, which is not always straightforward. At present, this process relies on a one-time effort of an expert for each new skill, which can be time-consuming and risks overlooking subtle yet important aspects of the skill. Related work has already proposed data-driven methods to automatically extract the most relevant aspects of different tasks from human demonstrations. For example, ExerAIde [41] systematically discerns nuanced kinematic differences

between healthy and pathological exercise movements, identifying the most critical joints for this distinction. However, ExerAIde is specifically designed for 4D skeleton keypoint data, while Pro-Skill considers extensible, multi-modal measurement data. Also, ExerAIde only identifies which measurements are most relevant, while Pro-Skill's parameters often result from aggregated information across multiple measurements, enabling more interpretable feedback. As illustrated by the above considerations, determining the skill-specific parameters of Pro-Skill presents a more challenging problem compared to existing approaches. To address these challenges, future work could explore semi-automated methods for parameter discovery, for instance, by using expert-in-the-loop machine learning approaches, clustering techniques to identify recurring patterns, or interactive tools that help experts label and validate candidate parameters based on sensor traces.

Fourth, Pro-Skill's current feedback provides insights on individual aspects of the performance, but lacks reasoning about relations between parameters (e.g., limited sawing amplitude may contribute to slow progress) and does not yet provide concrete guidance to effectively reach expert performance. Also, Pro-Skill currently assesses individual task executions and is not yet embedded in longitudinal training programs that guide learners across multiple sessions. In particular, while our evaluation demonstrates that Pro-Skill's assessments agree with expert judgments, it does not yet demonstrate that trainees learn more effectively when receiving this feedback. A natural next step is a controlled study in which trainees repeat the task after reviewing Pro-Skill's feedback, comparing their improvement against trainees receiving no feedback or conventional instructor feedback. Such a study can build directly on the assessment validity established here. Future research could also explore how Pro-Skill's automated assessments can be integrated into adaptive curricula, in which feedback and guidance evolve dynamically in response to learner progress. Going further, future versions of Pro-Skill could embed more conceptual understanding of the skill to better contextualize assessments, for example, explaining not only that the saw amplitude was insufficient, but also why larger strokes support a more efficient cut. Such explanations could be supported by contextual animations that illustrate the implications of technique adjustments, analogous to CODA's animated explanations of constraints [58]. Building on this, Pro-Skill could drive adaptive tutorials that personalize instructional content based on inferred expertise level. When low proficiency or recurring KP issues are detected, the system could recommend targeted learning resources (e.g., short videos, step-by-step instructions, or focused practice drills), and progressively adapt guidance as competence improves. Conversely, for advanced users, it could suggest expert tips, alternative workflows, or tool choices to support continued refinement without interrupting practice. Importantly, such content adaptation should be designed to promote reflection and autonomy rather than dependency, leveraging Pro-Skill's post-hoc KP-oriented analysis to support deliberate practice.

Fifth, a fundamental challenge in evaluating craftsmanship skills is the lack of an objective ground truth. Unlike domains where performance can be directly measured through standardized outcomes, manual skills are often assessed through expert observation, an approach that is inherently subjective. In our study, while experts showed high agreement on overall performance, their evaluations of individual parameters (such as posture or stroke amplitude) varied. This reflects the difficulty of consistently judging discrete aspects of embodied actions, especially when those aspects are interdependent or difficult to observe directly. This challenge underscores the need for new approaches to establish more reliable baselines for evaluating skill performance. One promising direction is to incorporate expert-in-the-loop systems, where domain experts use sensor-derived feedback to refine their evaluations. By making subtle deviations more visible, such tools could help align subjective assessments across raters. Assessing the outcome of a task can also serve as a partial validation of skill performance. While in our handsaw task, the end result (a straight cut) lacks precise, objective grading criteria, other domains, such as welding or machining, have well-established procedures for evaluating outcomes through non-destructive or destructive testing methods. Future research could explore how such outcome-based evaluations might provide more objective reference points for validating automated skill assessments.

Finally, Pro-Skill's low demonstration requirement does not necessarily imply a low deployment cost. Our handsaw study deliberately used optical infrared tracking to prioritize sensing accuracy, isolating the question of assessment validity before considering the effects of more practical sensing setups. However, Pro-Skill's pipeline is not tied to this sensing modality. Our drilling demonstration (Section 9) already operates entirely on IMU-based body and tool tracking together with off-the-shelf force sensing, and the platform additionally supports markerless motion capture. The pipeline therefore already supports more portable and deployable sensing. What remains to be established is whether assessment accuracy is preserved under these more practical sensing configurations. Determining the sensing fidelity required by each parameter, and validating reduced sensor setups accordingly, is an important direction for future work.

11 Conclusion

In this paper, we introduced Pro-Skill, a platform for multi-modal capturing and automated assessment of manual craftsmanship skills. Pro-Skill contributes an extensible, LSL-based multi-modal capturing system together with techniques for modeling and assessing skill performance. These techniques capture the natural variability of skilled execution from only a handful of expert demonstrations, making the system well-suited to ad-hoc training contexts.

Through a study on a handsaw task, we validated Pro-Skill's assessments against expert judgments, finding almost-perfect agreement on overall performance and strong alignment on parameters such as saw angle, deviation, and progress speed, while also showing complementary value where the system consistently analyzes parameters that are difficult to judge by eye. We further illustrated Pro-Skill's reconfigurability through proof-of-concept demonstrations of two additional tasks, vertical sawing and handheld drilling, showing that it can be configured for different skills, tools, and sensing setups.

Pro-Skill targets manual craftsmanship skills characterized by nominally monotonic progress, providing interpretable, parameter-level feedback on Knowledge of Performance without requiring large training datasets or continuous expert supervision. In doing so, it lays a foundation for automated skill assessment in workshops, vocational training, and other structured learning environments.

12 Acknowledgments

This work was supported by Flanders Make (Flanders Make is the Flemish Strategic Research Center for the manufacturing industry), through the Flanders Make SBO project AUTOCRAFT.

References

[1] Erwin Aertbeliën and Joris De Schutter. 2014. Learning a Predictive Model of Human Gait for the Control of a Lower-limb Exoskeleton. In *5th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics*. 520 – 525.

[2] Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. 2019. EduSense: Practical Classroom Sensing at Scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (Sept. 2019), 1–26. doi:10.1145/3351229

[3] Fraser Anderson, Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2013. YouMove: enhancing movement training with an augmented reality mirror. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, Scotland, United Kingdom) *(UIST '13)*. Association for Computing Machinery, New York, NY, USA, 311–320. doi:10.1145/2501988.2502045

[4] Christopher M. Bishop. 2006. *Pattern recognition and machine learning*. Springer, Chapter 12.

[5] Sanne M. B. I. Botden, Sonja N. Buzink, Marlies P. Schijven, and Jack J. Jakimowicz. 2008. ProMIS Augmented Reality Training of Laparoscopic Procedures Face Validity. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare* 3, 2 (2008), 97–102. doi:10.1097/SIH.0b013e3181659e91

[6] Sylvain Calinon, Florent Guenter, and Aude Billard. 2007. On Learning, Representing, and Generalizing a Task in a Humanoid Robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 37, 2 (2007), 286 – 298. doi:10.1109/TSMCB.2006.886952

[7] Thomas Callens, Arthur van der Have, Sam Van Rossom, Joris De Schutter, and Erwin Aertbeliën. 2020. A Framework for Recognition and Prediction of Human Motions in Human-Robot Collaboration Using Probabilistic Motion Models. *IEEE Robotics and Automation Letters (RAL) paper presented at the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* 5, 4 (2020), 5151 – 5158.

[8] Marianna Capecci, Maria Gabriella Ceravolo, Francesco Ferracuti, Sabrina Iarlori, Ville Kyrki, Andrea Monteriù, Luca Romeo, and Federica Verdini. 2018. A Hidden Semi-Markov Model based approach for rehabilitation exercise assessment. *Journal of Biomedical Informatics* 78 (2018), 1–11. doi:10.1016/j.jbi.2017.12.012

[9] Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 4724–4733.

[10] J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37.

[11] Nathan Devrio and Chris Harrison. 2022. DiscoBand: Multiview Depth-Sensing Smartwatch Strap for Hand, Body and Environment Tracking. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. ACM, Bend OR USA, 1–13. doi:10.1145/3526113.3545634

[12] Nathan DeVrio, Vimal Mollyn, and Chris Harrison. 2023. SmartPoser: Arm Pose Estimation with a Smartphone and Smartwatch Using UWB and IMU Data. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 79, 11 pages. doi:10.1145/3586183.3606821

[13] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. 2023. Project Aria: A New Tool for Egocentric Multi-Modal AI Research. arXiv:2308.13561 [cs]. doi:10.48550/arXiv.2308.13561

[14] P.M. Fitts and M.I. Posner. 1967. *Human Performance*. Brooks/Cole Publishing Company.

[15] Jun Gong, Fraser Anderson, George Fitzmaurice, and Tovi Grossman. 2019. Instrumenting and Analyzing Fabrication Activities, Users, and Expertise. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3290605.3300554

[16] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina González, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbeláez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shout, and Michael Wray. 2024. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 19383–19400. doi:10.1109/CVPR52733.2024.01834

[17] Shlomi Haar, Guhan Sundar, and A. Aldo Faisal. 2021. Embodied virtual reality for the study of real-world motor learning. *PLOS ONE* 16, 1 (Jan. 2021), e0245717. doi:10.1371/journal.pone.0245717

[18] Tomasz Hachaj and Marek R. Ogiela. 2014. Rule-based approach to recognizing human body poses and gestures in real time. *Multimedia Systems* 20, 1 (Feb. 2014), 81–99. doi:10.1007/s00530-013-0332-2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 770–778.

[20] Min-Chun Hu, Chi-Wen Chen, Wen-Huang Cheng, Che-Han Chang, Jui-Hsin Lai, and Ja-Ling Wu. 2015. Real-Time Human Movement Retrieval and Assessment With Kinect Sensor. *IEEE Transactions on Cybernetics* 45, 4 (2015), 742–753. doi:10.1109/TCYB.2014.2335540

[21] Jun-Da Huang. 2011. Kinerehab: a kinect-based system for physical rehabilitation: a pilot study for young adults with motor disabilities. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*. ACM, Dundee Scotland, UK, 319–320. doi:10.1145/2049536.2049627

[22] Kevin Huang. 2018. Exploring In-Home Monitoring of Rehabilitation and Creating an Authoring Tool for Physical Therapists. (7 2018). doi:10.1184/R1/6716594.v1

[23] Auke Jan Ijspeert, Jun Nakanishi, Heiko Hoffmann, Peter Pastor, and Stefan Schaal. 2013. Dynamical Movement Primitives: Learning Attractor Models for Motor Behaviors. *Neural Computation* 25, 2 (2013), 328–373. doi:10.1162/NECO_a_00393

[24] A. Iosifidis, A. Tefas, and I. Pitas. 2012. View-Invariant Action Recognition Based on Artificial Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* 23, 3 (March 2012), 412–424. doi:10.1109/TNNLS.2011.2181865

[25] Timothy Judkins, Dmitry Oleynikov, and Nicholas Stergiou. 2006. Real-time augmented feedback benefits robotic laparoscopic training. *Studies in health technology and informatics* 119 (Feb. 2006), 243–8.

[26] Xiao Ke, Huangbiao Xu, Xiaofeng Lin, and Wenzhong Guo. 2024. Two-Path Target-Aware Contrastive Regression for Action Quality Assessment. *Information Sciences* 664 (2024), 120347.

[27] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-Play Acoustic Activity Recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. ACM, Berlin Germany, 213–224. doi:10.1145/3242587.3242609

[28] Gierad Laput, Chouchang Yang, Robert Xiao, Alanson Sample, and Chris Harrison. 2015. EM-Sense: Touch Recognition of Uninstrumented, Electrical and Electromechanical Objects. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. ACM, Charlotte NC USA, 157–166. doi:10.1145/2807442.2807481

[29] Gierad Laput, Yang Zhang, and Chris Harrison. 2017. Synthetic Sensors: Towards General-Purpose Sensing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 3986–3999. doi:10.1145/3025453.3025773

[30] Markus Laube, Michael Haslgrübler, Behrooz Azadi, Bernhard Anzengruber-Tánase, and Alois Ferscha. 2022. Skill Level Detection in Arc Welding towards an Assistance System for Workers. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, Corfu Greece, 73–82. doi:10.1145/3529190.3529206

[31] Yalin Liao, Aleksandar Vakanski, and Min Xian. 2020. A Deep Learning Framework for Assessing Physical Rehabilitation Exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28, 2 (Feb. 2020), 468–477. doi:10.1109/TNSRE.2020.2966249

[32] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video Swin Transformer. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 3192–3201.

[33] M. Merleau-Ponty. 2002. *Phenomenology of Perception*. Routledge. https://books.google.be/books?id=q3HwhfjRmswC

[34] Helen C. Miles, Serban R. Pop, Simon J. Watt, Gavin P. Lawrence, and Nigel W. John. 2012. A review of virtual environments for training in ball sports. *Computers & Graphics* 36, 6 (Oct. 2012), 714–726. doi:10.1016/j.cag.2012.04.007

[35] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. IMUPoser: Full-Body Pose Estimation using IMUs in Phones, Watches, and Earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 529, 12 pages. doi:10.1145/3544548.3581392

[36] David Monaghan, Freddie Honohan, Amin Ahmadi, Troy McDaniel, Ramin Tadayon, Ajay Karpur, Kieran Morran, Noel E. O'Connor, and Sethuraman Panchanathan. 2016. A Multimodal Gamified Platform for Real-Time User Feedback in Sports Performance. In *Proceedings of the 24th ACM International Conference on Multimedia* (Amsterdam, The Netherlands) *(MM '16)*. Association for Computing Machinery, New York, NY, USA, 708–710. doi:10.1145/2964284.2973815

[37] Alexandros Paraschos, Christian Daniel, Jan R. Peters, and Gerhard Neumann. 2013. Probabilistic Movement Primitives. In *Advances in Neural Information Processing Systems*, Vol. 26.

[38] Hyung Kun Park and Woohun Lee. 2016. Motion Echo Snowboard: Enhancing Body Movement Perception in Sport via Visually Augmented Feedback. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. ACM, Brisbane QLD Australia, 192–203. doi:10.1145/2901790.2901797

[39] Michael Polanyi. 1966. *The tacit dimension*. Anchor, Garden City, NY.

[40] D. Pye. 1995. *The Nature and Art of Workmanship*. Herbert Press. https://books.google.be/books?id=z-o9AQAAIAAJ

[41] Ahmed Qazi and Asim Iqbal. 2024. ExerAIde: AI-assisted Multimodal Diagnosis for Enhanced Sports Performance and Personalised Rehabilitation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 3430–3438.

[42] Raf Ramakers, Danny Leen, Jeeeun Kim, Kris Luyten, Steven Houben, and Tom Veuskens. 2023. Measurement patterns: User-oriented strategies for dealing with measurements and dimensions in making processes. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[43] Daniel Roetenberg, Henk Luinge, and Per Slycke. 2009. Xsens MVN: Full 6DOF Human Motion Tracking Using Miniature Inertial Sensors. (2009).

[44] H. Sakoe and S. Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing* 26, 1 (1978), 43 – 49.

[45] Alan W Salmoni, Richard A Schmidt, and Charles B Walter. 1984. Knowledge of results and motor learning: a review and critical reappraisal. *Psychological bulletin* 95, 3 (1984), 355.

[46] R.A. Schmidt, T.D. Lee, C. Winstein, G. Wulf, and H.N. Zelaznik. 2018. *Motor Control and Learning: A Behavioral Emphasis*. Human Kinetics. https://books.google.be/books?id=EvJ6DwAAQBAJ

[47] Richard A Schmidt, Douglas E Young, Stephan Swinnen, and Diane C Shapiro. 1989. Summary knowledge of results for skill acquisition: support for the guidance hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15, 2 (1989), 352.
[48] Chuan-Jun Su, Chang-Yu Chiang, and Jing-Yan Huang. 2014. Kinect-enabled home-based rehabilitation system using Dynamic Time Warping and fuzzy logic. *Appl. Soft Comput.* 22 (Sept. 2014), 652–666. doi:10.1016/j.asoc.2014.04.020
[49] Shogo Tanaka. 2011. The notion of embodied knowledge. 149–157.
[50] Richard Tang, Xing-Dong Yang, Scott Bateman, Joaquim Jorge, and Anthony Tang. 2015. Physio@Home: Exploring Visual Guidance and Feedback Techniques for Physiotherapy Exercises. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, Seoul Republic of Korea, 4123–4132. doi:10.1145/2702123.2702401
[51] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. 2020. Uncertainty-Aware Score Distribution Learning for Action Quality Assessment. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition)*. IEEE, 9836–9845.
[52] Balasaravanan Thoravi Kumaravel, Fraser Anderson, George Fitzmaurice, Bjoern Hartmann, and Tovi Grossman. 2019. Loki: Facilitating Remote Instruction of Physical Tasks Using Bi-Directional Mixed-Reality Telepresence. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. ACM, New Orleans LA USA, 161–174. doi:10.1145/3332165.3347872
[53] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *IEEE International Conference on Computer Vision*. IEEE, 4489–4497.
[54] Dishita G Turakhia, Yini Qi, Lotta-Gili Blumberg, Andrew Wong, and Stefanie Mueller. 2021. Can Physical Tools that Adapt their Shape based on a Learner's Performance Help in Motor Skill Training?. In *Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, Salzburg Austria, 1–12. doi:10.1145/3430524.3440636
[55] Eduardo Velloso, Andreas Bulling, and Hans Gellersen. 2013. MotionMA: motion modelling and analysis by demonstration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Paris France, 1309–1318. doi:10.1145/2470654.2466171
[56] Cristian Vergara, Joris De Schutter, and Erwin Aertbeliën. 2019. Combining Imitation Learning With Constraint-Based Task Specification and Control. *IEEE robotics and automation letters* 4, 2 (2019), 1892–1899.
[57] Cristian Vergara, Joris De Schutter, and Erwin Aertbeliën. 2020. Learning robust manipulation tasks involving contact using trajectory parameterized probabilistic principal component analysis. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 8336–8343.
[58] Tom Veuskens, Florian Heller, and Raf Ramakers. 2021. CODA: A Design Assistant to Facilitate Specifying Constraints and Parametric Behavior in CAD Models.. In *Graphics Interface*. 87–96.
[59] Angchi Xu, Ling-An Zeng, and Wei-Shi Zheng. 2022. Likert Scoring with Grade Decoupling for Long-term Action Assessment. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition)*. IEEE, 3222–3231.
[60] Li Ye, Simin Yang, Xueyan Zhou, and Yuxi Lin. 2023. Supporting traditional handicrafts teaching through eye movement technology. *International Journal of Technology and Design Education* 33, 3 (July 2023), 981–1005. doi:10.1007/s10798-022-09748-z
[61] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. 2021. Group-aware Contrastive Regression for Action Quality Assessment. In *IEEE International Conference on Computer Vision*. IEEE, 7899–7908.
[62] Boyu Zhang, Jiayuan Chen, Yinfei Xu, Hui Zhang, Xu Yang, and Xin Geng. 2024. Auto-Encoding Score Distribution Regression for Action Quality Assessment. *Neural Computing and Applications* 36, 2 (2024), 929–942.
[63] Kanglei Zhou, Ruizhi Cai, Liyuan Wang, Hubert P. H. Shum, and Xiaohui Liang. 2024. A Comprehensive Survey of Action Quality Assessment: Method and Benchmark. arXiv:2412.11149 [cs.CV] https://arxiv.org/abs/2412.11149
[64] Kanglei Zhou, Junlin Li, Ruizhi Cai, Liyuan Wang, Xinxgxing Zhang, and Xiaohui Liang. 2024. CoFInAl: Enhancing Action Quality Assessment with Coarse-to-Fine Instruction Alignment. In *International Joint Conference on Artificial Intelligence (IJCAI)*.
[65] Kanglei Zhou, Yue Ma, Hubert P. H. Shum, and Xiaohui Liang. 2023. Hierarchical Graph Convolutional Networks for Action Quality Assessment. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 12 (2023), 7749–7763.

Appendices

A Individual Inter-Rater Agreement Between Experts and the Models Trained on Their Performances

Table 4. Inter-rater agreement for performance parameters between expert 1 and their expert TraPPCA model.

Parameter	Cohen's Kappa κ	95% Confidence Interval		Agreement
		Lower	Upper	
Sawing Amplitude	0.383	-0.052	0.818	Fair
Saw Deviation	0.826	0.675	0.977	Almost perfect
Posture	0.609	0.253	0.964	Substantial
Progress Speed	0.826	0.724	0.981	Almost perfect
Sawing Angle	0.693	0.474	0.913	Substantial
Overall	0.900	0.800	1.00	Almost perfect

Table 5. Inter-rater agreement for performance parameters between expert 2 and their expert TraPPCA model.

Parameter	Cohen's Kappa κ	95% Confidence Interval		Agreement
		Lower	Upper	
Sawing Amplitude	0.136	-0.134	0.400	Slight
Saw Deviation	0.934	0.862	1.000	Almost perfect
Posture	0.716	0.453	0.980	Substantial
Progress Speed	0.785	0.536	1.000	Substantial
Sawing Angle	0.536	0.215	0.857	Moderate
Overall	0.876	0.751	1.000	Almost perfect

Table 6. Inter-rater agreement for performance parameters between expert 3 and their expert TraPPCA model.

Parameter	Cohen's Kappa κ	95% Confidence Interval		Agreement
		Lower	Upper	
Sawing Amplitude	0.850	0.753	0.947	Almost perfect
Saw Deviation	0.721	0.448	0.994	Substantial
Posture	0.385	0.070	0.699	Fair
Progress Speed	0.876	0.789	0.964	Almost perfect
Sawing Angle	0.667	0.352	0.981	Substantial
Overall	0.863	0.664	1.000	Almost perfect

Table 7. Inter-rater agreement for performance parameters between expert 4 and their expert TraPPCA model.

Parameter	Cohen's Kappa κ	95% Confidence Interval		Agreement
		Lower	Upper	
Sawing Amplitude	0.564	0.321	0.808	Moderate
Saw Deviation	0.679	0.398	0.960	Substantial
Posture	0.117	-0.435	0.669	Slight
Progress Speed	0.894	0.800	0.987	Almost perfect
Sawing Angle	0.703	0.476	0.929	Substantial
Overall	0.741	0.609	0.872	Substantial

B Individual Inter-Rater Agreement Between Experts and DTW Baseline

Table 8. Inter-rater agreement for performance parameters between expert 1 and their DTW baseline.

Parameter	Cohen's Kappa κ	95% Confidence Interval		Agreement
		Lower	Upper	
Sawing Amplitude	0.663	0.355	0.971	Substantial
Saw Deviation	0.450	0.203	0.697	Moderate
Posture	0.400	0.121	0.679	Fair
Progress Speed	0.583	0.393	0.773	Moderate
Sawing Angle	0.596	0.264	0.928	Moderate
Overall	0.552	0.285	0.820	Moderate

Table 9. Inter-rater agreement for performance parameters between expert 2 and their DTW baseline.

Parameter	Cohen's Kappa κ	95% Confidence Interval		Agreement
		Lower	Upper	
Sawing Amplitude	0.161	0.011	0.311	Slight
Saw Deviation	0.706	0.462	0.951	Substantial
Posture	0.477	0.198	0.755	Moderate
Progress Speed	0.796	0.631	0.962	Substantial
Sawing Angle	0.243	-0.037	0.523	Fair
Overall	0.368	0.165	0.572	Fair

Table 10. Inter-rater agreement for performance parameters between expert 3 and their DTW baseline.

Parameter	Cohen's Kappa κ	95% Confidence Interval		Agreement
		Lower	Upper	
Sawing Amplitude	0.486	0.230	0.741	Moderate
Saw Deviation	0.820	0.648	0.992	Almost perfect
Posture	0.667	0.344	0.990	Substantial
Progress Speed	0.759	0.556	0.962	Substantial
Sawing Angle	0.216	-0.041	0.472	Fair
Overall	0.144	-0.031	0.320	Slight

Table 11. Inter-rater agreement for performance parameters between expert 4 and their DTW baseline.

Parameter	Cohen's Kappa κ	95% Confidence Interval		Agreement
		Lower	Upper	
Sawing Amplitude	0.373	0.002	0.743	Fair
Saw Deviation	0.751	0.514	0.988	Substantial
Posture	0.102	-0.035	0.238	Slight
Progress Speed	0.566	0.350	0.781	Moderate
Sawing Angle	0.320	-0.102	0.742	Fair
Overall	0.243	0.108	0.378	Fair

C Matched Versus Decoupled Inter-Rater Agreement

Table 12. Matched versus decoupled agreement between the Pro-Skill models and the expert raters. Matched agreement averages the diagonal of the model×rater matrix (model and rater correspond to the same expert); decoupled agreement averages the off-diagonal cells (model and rater correspond to different experts). Weighted Cohen's κ (quadratic).

Parameter	Matched (diagonal)	Decoupled (off-diagonal)	Difference
Sawing Amplitude	0.48	0.41	+0.07
Saw Deviation	0.79	0.75	+0.04
Posture	0.46	0.47	-0.01
Progress Speed	0.85	0.86	-0.01
Sawing Angle	0.65	0.68	-0.03
Overall	0.84	0.80	+0.05

Table 13. Full model×rater agreement matrix per parameter (weighted Cohen's κ, quadratic). Rows are the Pro-Skill models trained on each expert's demonstrations; columns are each expert's manual ratings. Bold diagonal cells are the matched pairs (model and rater correspond to the same expert); off-diagonal cells are the decoupled pairs. The means of these cells are summarised in Table 12.

		Rating expert			
Parameter	Model	E1	E2	E3	E4
Overall	E1	**0.90**	0.85	0.89	0.75
	E2	0.82	**0.88**	0.82	0.80
	E3	0.76	0.69	**0.86**	0.76
	E4	0.85	0.70	0.90	**0.74**
Progress Speed	E1	**0.85**	0.83	0.89	0.86
	E2	0.83	**0.78**	0.91	0.82
	E3	0.84	0.82	**0.88**	0.87
	E4	0.87	0.82	0.92	**0.89**
Saw Deviation	E1	**0.83**	0.87	0.85	0.73
	E2	0.83	**0.93**	0.91	0.85
	E3	0.62	0.76	**0.72**	0.80
	E4	0.53	0.65	0.63	**0.68**
Sawing Angle	E1	**0.69**	0.75	0.75	0.70
	E2	0.66	**0.54**	0.63	0.81
	E3	0.59	0.64	**0.67**	0.74
	E4	0.60	0.61	0.63	**0.70**
Posture	E1	**0.61**	0.64	0.54	0.54
	E2	0.43	**0.72**	0.85	0.49
	E3	0.57	0.39	**0.38**	0.19
	E4	0.35	0.30	0.34	**0.12**
Sawing Amplitude	E1	**0.38**	0.42	0.42	0.64
	E2	0.02	**0.13**	0.06	0.38
	E3	0.60	0.76	**0.85**	0.64
	E4	0.30	0.37	0.36	**0.56**