Beyond Activity Recognition: Skill Assessment from Accelerometer Data

Aftab Khan†, Sebastian Mellor†, Eugen Berlin*, Robin Thompson†, Roisin McNaney†, Patrick Olivier†, Thomas Plötz†
†Open Lab, School of Computing Science, Newcastle University.
*Department of Computer Science, Technische Universität Darmstadt.
aftab.khan@newcastle.ac.uk

ABSTRACT
The next generation of human activity recognition applications in ubiquitous computing scenarios focuses on assessing the quality of activities, which goes beyond mere identification of activities of interest. Objective quality assessments are often difficult to achieve, hard to quantify, and typically require domain specific background information that bias the overall judgement and limit generalisation. In this paper we propose a framework for skill assessment in activity recognition that enables automatic quality analysis of human activities. Our approach is based on a hierarchical rule induction technique that effectively abstracts from noise-prone activity data and assesses activity data at different temporal contexts. Our approach requires minimal domain specific knowledge about the activities of interest, which makes it largely generalisable. By means of an extensive case study we demonstrate the effectiveness of the proposed framework in the context of dexterity training of 15 medical students engaging in 50 attempts of surgical activities.

Author Keywords
Activity Recognition; Skill Assessment; Rule induction; Classification; Accelerometer;

ACM Classification Keywords

INTRODUCTION
Human activity recognition (HAR) based on the analysis of inertial measurement data such as accelerometry is one of the core concerns of ubiquitous and wearable computing. It has traditionally focused on the automatic detection and classification of specific activities a person pursues in their environment [5]. A multitude of technical approaches has been proposed that enable the development of applications in domains as diverse as novel interaction techniques [21], supported support in smart environments [17], occupancy monitoring for energy efficient buildings [18], or automated health assessments [16, 30] or health care automation [6, 24, 31] to name but a few. Beyond the mere recognition of certain activities of interest, a number of application domains now require the analysis of the quality of a person’s activities. Such skill assessment is, for example, relevant for progress assessment in physical rehabilitation [26], or for coaching in certain sports [22].

Existing skill assessment approaches are typically driven by the specific application domain they have been designed for – in the sense that they aim for measuring certain pre-defined, very specific parameters that domain experts deem appropriate. For example, the ClimbaX system that assesses the skills of amateur rock climbers, measured four concrete parameters – power, control, stability, and speed – that are specifically relevant for judging the quality of rock climbing [22]. Without doubt such specialised assessment approaches fulfil their purpose of judging the quality of certain specific activities of interest through measuring key performance attributes. As such these methods are of value for those practical applications. However, chances are little that such bespoke assessment approaches generalise well beyond the application domain they were originally designed for. Consequently, moving on to quality assessment in a new, i.e., different application domain would require starting over again the development process through identifying those key performance attributes and, subsequently, conceiving measurement procedures that automatically infer said attributes from sensor data.

In this paper we take a different approach to skill assessment. We hypothesise that skill is an inherent feature of the quality of activities and how they are performed by individuals, which can be extracted solely using data-driven analysis techniques that operate on representative sample sets. We argue that skill assessment requires quality analysis at different levels of contextual, i.e., temporal abstraction and present a data-driven assessment method, which analyses accelerometer data at varying levels of abstraction using a hierarchical, stochastic rule induction framework. As such we substantially relax the requirements for automated skill assessment: whilst still relying on domain specific sample data (and meta-level ground truth classification), our analysis framework does not rely on explicit specifications of performance.
attributes but rather automatically incorporates those parts that are relevant for quality comparisons. With the proposed framework, the same learning approach could be readily applied in other domains.

The main contributions of this paper are as follows: i) First, we summarise the state-of-the-art in automated skill assessment in the field of ubiquitous and wearable computing. ii) Subsequently, we propose a data-driven, rule based skill assessment framework that can be used for quality analysis of human activities. This framework is based on a hierarchical, stochastic rule induction method that analyses activity data at different levels of abstraction thereby not being constrained by the domain itself. iii) We demonstrate the effectiveness of the developed framework in the context of a substantial case study in the field of dexterity training of medical students where we validate the effectiveness of the proposed framework. Using surgical instruments with embedded miniature accelerometers we recorded movement data from training sessions where medical students practiced basic surgical procedures such as stitching, knot tying, and instrument use. We show how our data-driven assessment framework can effectively replicate quality rankings based on comparative OSATS (“Objective structured assessment of technical skill[s] for surgical residents” [25]) criteria as domain experts would provide them. OSATS criteria were solely used for validating the proposed concept, and not for designing or training the assessment method. Finally, we outline how this framework can be used in real-life training scenarios where cohorts of medical students undergo standard dexterity training and an automated assessment framework would provide automated assessment feedback.

BACKGROUND

In the context of our work, we understand skill as relating to an individual’s ability to perform a task to a particular standard, repeatedly and consistently. Much work has been conducted in the field of activity recognition, examining methods to identify the execution of specific activities [5]. Whilst certain aspects of a task’s execution must first be known in order to assess an individual’s skill, the assessment of said skill goes beyond activity recognition. It requires to quantify what is traditionally dealt with in a subjective manner. Judges, teachers and coaches, to name but a few, employ skill assessment in the execution of their profession, relying on experience and intuition to accurately gauge the competence of the subject. Replicating that judgement is a complicated task, however attempts have been made to do so in a range of fields. The majority of reviewed applications rely upon a range of assumptions for the assessment of the quality with which a task is performed; from clear cut observations of power and composition for climbing; to template based performance assessments; to assessments through the use of rudimentary proxies for skill.

Datta et al. [13] describe the use of 3D motion capture to evaluate the precision of a surgeon performing open surgery in a simulation, measuring the number of movements made, and the time in which the surgery was completed. Apart from this, skill assessment, largely using wearable sensing platforms and machine learning techniques for data analysis, plays a key role in the sports domain. For example, the SwimMaster system [7] employs sensor data to evaluate the efficiency of swimming strokes, providing information on technique that can be used in training. The GymSkill system [26] provides an implementation of a personal trainer for physical exercise using balance equipment. Able to provide an indication of the quality of the exercises, the system can also indicate errors in their execution. The ClimbAX system [22] measures key performance indicators of rock climbers aiming for objective coaching of amateur climbers. The point at which the degree and direction of analysis arrives at an approximation of skill assessment is indistinct, however within their domains each of these applications provide an assessment of some quality of the task under examination, be it speed, efficiency or orientation. Consequently it is the identification of this quality that is of prime interest.

Extraction of the pertinent features of performance for the assessment of skill is more straightforward in some domains than others. Through the identification of core skills required by rock climbers, Ladha et al. [22] defined a set of metrics by which to evaluate the quality of climbing. In these circumstances it is possible to observe already established metrics and directly relate them to the data obtained from sensors providing a clear description of the quality of the climb. Bouwsema et al. [10] make assumptions about metrics by which to measure skill whilst operating a prosthetic hand. They validate their approach through a correlation analysis between the measures they devised and a clinical assessment procedure for measuring capability with a prosthetic. Again this demonstrates how domain specific knowledge can facilitate the extraction of salient measures.

Ahmadi et al. [1] use accelerometry data to investigate the difference between sub-elites and developmental athletes. By identifying the areas in which these two categories differed most, they were able to rank skill-related features according to their relevance. This demonstrates that where clearly outlined metrics are not used, it is still possible to make an assessment by increasing the level of abstraction with which the performance is observed. MusicJacket [39], a wearable system to support teaching of good posture and technique to violin players, similarly uses the concept of an “ideal” performance with regards to bowing. By examining the build of the musician and the manner in which they naturally hold the violin, a trajectory for the bow is established that is considered to be most desirable. Adherence to said ideal is then considered to be an indicator of quality, without having to examine the sound that the instrument produces.

Other domains lend themselves less favourably to evaluation in this manner. Tasks that would not normally be considered for assessment, e.g., household chores and other everyday activities that able-bodied people perform without effort, but challenge, for example, a recovering stroke patient [15] do not normally come under this kind of scrutiny. As a result it is harder to elicit those characteristics that define the quality of execution. The approach used in [15] is to find a proxy for quality, in this case the speed at which patients walk, cycle,
climb stairs etc. Through the use of this proxy, an insight into the quality of the gait is provided allowing for monitoring of post-stroke patients’ progress towards recovery.

The main limitation on skill assessment as it stands, is the requirement of domain specific knowledge. In order to extract the parameters on which to assess the data there must be some level of expertise in the field to begin with, which imposes a restraint on the generalisation of such assessment methods. The framework presented in this paper looks to circumvent this restriction by introducing a technique through which the parameters of assessment can be established without prior expert knowledge.

**SKILL ASSESSMENT FROM ACCELEROMETER DATA**

A common definition of *skill* refers to “the ability to do something well; expertise” [29], which canonically focuses on procedural aspects. Whilst the assessment of the quality of, for example, a final product as it is produced by a craftsperson, could serve as a proxy for a judgment of its maker’s “abilities to do [the required workmanship] well”, it only allows for generic analysis as it does not necessarily unveil those parameters and features that discriminate an experienced worker from a novice. It is the procedural aspects that are of interest for applications of skill assessment. For example, situated, personalised coaching would effectively support the aforementioned craftsperson to become an expert in their field.

We approach the assessment of a person’s skills from a strictly activity based point of view, i.e., we translate the task of skill assessment into an analysis of the quality of activities. In well defined scenarios the latter then serves as an indirect measure for the human’s skills according to the aforementioned definition. In this framework, skill analysis that focuses on specific tasks necessarily has to be undertaken at varying levels of abstraction thereby taking into account different temporal contexts. For example, the ability to plan sequences of working steps considers a larger temporal context, possibly spanning across the whole of the task, while certain dexterity related parameters focus on much shorter, possibly repetitive activities such as specific gestures as they have to be performed when handling certain tools and utensils.

Aiming for generalisation, and thus for the provision of the basis for generic understanding of skills, we employ a hierarchical analysis approach that captures the essentials of activities, thereby explicitly taking differing levels of temporal context into account. Figure 1 gives an overview of our skill assessment framework. Without limiting the overall applicability of our approach to specific sensing modalities, we start from capturing raw activity data through direct movement sensing using tri-axial accelerometers (Figure 1a). Our analysis approach first abstracts from raw, typically noisy sensor readings through employing a symbolic representation framework that subsumes stationary periods in accelerometer data to discrete symbols [8] (Figure 1b). Based on this compact representation of sensor data, we then use a hierarchical rule induction scheme, in an unsupervised manner, that extracts the essentials of the studied activities at a range of automatically derived temporal context levels (Figure 1c).

Based on the automatically extracted, rule based representation of activity data we then extract features (Figure 1d) that characterise the extracted rule topology and thus, implicitly, describe the characteristics of the underlying activities that effectively capture characteristics that correlate to skills as they are relevant for the studied task (skills that are identified and judged by an expert (Figure 1f)). Through our novel, multi-
level representation of activity data we can now perform skill assessment thereby not structurally being dependent on task specifics. Instead, a skill assessment system for a specific task would merely require expert rankings – instead of absolute scores – of sample data. Such a relaxed condition effectively avoids expert bias as it is often observed in absolute scoring scenarios (independent of the concrete application domain). Effectively, we perform pair-wise comparisons of data samples (Figure 1g) and associated features (Figure 1d), which are then forwarded to a standard classification protocol (Figure 1c) identifying changes in skill and ranking with respect to a comparison cohort (Figure 1i). Using the skill change prediction scores and including the absolute skill assessment scores (Figure 1h) (if available), we can not only reliably rank new datasets (Figure 1i) (e.g., recorded from new participants of a study) according to their skill levels, but also predict an absolute value on the original absolute skill assessment metric (Figure 1j).

In contrast to existing approaches, our framework provides the basis for universal skill assessment, specialising towards particular domains solely through the analysis of domain specific data, thereby only requiring meta-annotation that indicates quality of sessions, and ultimately requiring only that these annotations are consistent in their relative assessments, thus reducing any concern over aforementioned experts bias, arising through variation in experience or intuition. In what follows we describe the technical details of our assessment framework.

A Symbolic Representation for Accelerometry Data
Activities of interest may be recorded with a variety of sensors, ranging from rich data sources such as audio and visual, to direct movement sensing using wearable accelerometers, or environmental sensors that indirectly record activity data such as PIR sensors, event logs, or even simple diary entries. For our developments we opportunistically employ accelerometer based activity sensing. However, the skill assessment framework itself is agnostic regarding the particular sensing modalities used.

Raw sensor data are inherently noisy and as such not well suited for direct analysis methods. For traditional activity recognition tasks, the first step of processing typically corresponds to feature extraction, often based on a sliding window procedure that processes analysis frames of consecutive sensor readings [12]. Whilst such approaches work well for some activity classification tasks, they don’t seem promising for more fine-grained quality or skill assessment as it is within our focus. Our approach requires de-noised but detailed enough input data in order to be able to learn the inherent skill features.

An alternative to aforementioned explicit feature extraction from raw sensor data is quantisation, which maps the underlying continuous time series data onto a discrete range of possible values, i.e., a finite lexicon. Widely used quantisation methods for generic time-series data include SAX, Persist, and ACA to name but a few [34]. More targeted towards the quantisation of accelerometer data, recently a very powerful alternative symbolic representation framework has been proposed, which is based on piecewise linear approximation of the raw sensor data [8]. We use this quantisation framework to achieve sensor data abstraction using the following two-step procedure: i) First, the original 3-dimensional acceleration time series is approximated by piecewise linear segments, thus reducing the amount of data points but preserving the actual shape of the signal. ii) In the second step these linear segments are then mapped to a symbolic representation, which allows to efficiently perform extraction of recurring patterns, subsequence matching, etc. Figure 2a illustrates the mapping of tri-axial accelerometer data to sequences of symbols from a discrete, finite lexicon. Technical details and performance analysis of the quantisation approach are explained in the original publication [8].

Multi-Level Contextual Analysis of Activity Data
Through applying the symbolic representation framework as described above we translate raw sensing data into symbolic sequences of basic “stroke”-like movement primitives. Based on these symbolic sequences, we now extract features that capture the characteristics of the underlying activity data at varying levels of contextual abstraction. This general approach responds to our hypothesis that skill – or synonymously: the quality of activities performed – becomes manifest in activity data at a range of different levels of abstraction with varying temporal horizons. We employ a hierarchical stochastic rule induction framework that provides access to sequential symbolic data in different temporal contexts, thereby effectively abstracting away from noise and insignificant procedural variations [23, 27, 28]. Such models have previously been used for, e.g., unsupervised analysis of sports games and their regularities with a view on anomaly detection that can be linked back to progress (or decline) of technical skills [20].

By means of automatically derived, hierarchical rule-based representation of activity data we are then able to classify changes in skill levels. In what follows we describe the technical details of our approach to multi-level contextual analysis of activity data, and how we use this framework for skill assessment.

Stochastic Rule Induction
Stochastic rule induction refers to the extraction of rule structures in a probabilistic fashion from a set of observations. To this end, we employ a generic hybrid model that comprises two main analysis methodologies, namely a Multi-Level Chinese Takeaway Process – MLCTP – and the Cartesian product Label-based Hierarchical Bottom-up Clustering – CLHBC [20] originally proposed in the context of automated sports video annotation [3, 4, 14, 19, 40]. We employ this technique as it was found to be beneficial in a number of analysis domains where stochastic rule induction played a key role for the analysis of, for example, sports games [20]. MLCTP represents a hierarchical extension of the classical Chinese restaurant process [2] and the Stick-Breaking construction paradigm [36]. Chinese restaurant processes (and the stick-breaking construction) belong to a group of non-parametric stochastic processes that are naturally capable of representing grouped data and have been used in a number
of different applications ranging from topic modelling [9] to
genetics [33] and hierarchical clustering of images [32].

MLCTP is typically used for constructing multi-layer Markov
models with implicit memory represented in the hierarchy,
which has shown to be very effective for probabilistic model-
ing of complex activity data. CLHBC provides a com-
plementary modelling approach, which is based on the label
structures of sequential data. Note that these labels do not
necessarily correspond to ground truth annotations but rather
to – even artificial – symbolic representations. In our case
the hierarchical rule structures are established by considering
symbolic labels associated with the activity data as extracted
in the first step of our workflow.

MLCTP-CLHBC corresponds to a three stage process: i) first,
rule topologies are generated by MLCTP; ii) (generic) labels are associated with various rule structures, using the
CLHBC model; and iii) transition probabilities are finally
computed which are associated with the induced rule space
(see Figure 2b for an illustration). This procedure is per-
formed in a completely data driven fashion and is independent
of the type of sequential data provided to the method.

The number of levels, and the topology structure itself, are
determined by MLCTP. Bottom-level states are associated with
the input number of symbols (as extracted from the raw sen-
sor data). A bijective mapping of leaf states (bottom level in
Figure 2b) to the observations is introduced, which is used to
compute the transition likelihoods between observations $E_{O-1}$ to $E_O$: $O \in \{1, 2, 3, ..., G_H\}$ where $G_H$ refers to the
total number of unique observations, i.e., in our case the total
number of unique symbols generated by the quantisation step.
The established hierarchical structure is then populated with
transition probabilities and represented with a single transi-
tion matrix using the normalised products of all parent state
transitions (linked via connected nodes) such that:

$$L(E_O|E_{O-1}) = \frac{1}{D} \prod_{v=H} P(z_{y}^{O}|z_{y}^{O-1})$$

where $L(E_O|E_{O-1})$ is the augmented likelihood for MLCTP
generated state transitions between events, $E_{O-1}$ and $E_O$. $D$
is the normalisation constant and $H$ is the total number of
levels. $z_{y}^{O}$ represents state $\zeta$, indexed by $y$, with its parent
state $x$ and is at level $v$, for an input observation index $O$.
More details of this method can be found in [20].

Hierarchical representations of rule structures enable linkage
between observed activities and their progressively abstract
meanings as we go higher up in the rule hierarchy. In the
case of sensing systems, this means high level abstraction
from low-level raw/symbolic observations to the correspond-
ing contextual representations that might be more indicative
of skill or quality. Through considering structural properties
of the induced rule hierarchy we can now analyse activity
data from a more abstract and at the same time very thor-
ough standpoint. Figure 2b illustrates an example topology
for activity data as extracted using MLCTP, with transition
probabilities of the induced rule hierarchy being calculated
using Eq. 1. The bottom level corresponds to the number of
discrete symbols as extracted from the raw sensor data, and
higher levels represent said abstract representations of con-
textual correlations of activity data represented by their transi-
tion matrices.

Classification of Skill

The multi-level analysis of symbolic activity data obtained
through the aforementioned stochastic rule induction process
allows us to calculate contextual representations at different
levels of abstraction in the form of aforementioned transition
matrices. We argue that these matrices capture information
that is relevant for a quality analysis of the underlying activity
sequences that goes beyond the recognition of their con-
stituent actions.

In the next step of our analysis procedure we use the gener-
ated transition matrices and the rule topology to extract mean-
meaningful features that will eventually provide access to those
characteristics inherent to the analysed activities that are rele-
vant for skill and quality assessment. Since we aim for gener-
alization of our assessment framework we do not restrict ourselves unnecessarily to particular application domains. The features are generic in the sense that they characterise structural properties of the extracted rule topologies.

**Rule Induction Metrics**

Rule-based representations of symbolic activity data contain the primary source of abstract information for feature extraction. We can cluster these representations into two groups: a) Complexity metrics of the induced rule hierarchy, b) Metrics associated with the induced transitions (Table 1).

Complexity metrics include the number of nodes, number of levels, total number of branching points and the average number of branching points per level indicating the complexity of the induced rule hierarchy. The more complex a set of sequential symbolic representations is, the more complex the hierarchical structure becomes. We also compute the log-likelihood of the complete transition matrix representing the amount of state changes between activities. A log-likelihood of the diagonal and the trace of the hierarchical transition matrices are also computed highlighting the amount of time spent in the same state. Rule topologies are extracted using two hyper-parameters (controlling the depth (i.e., the number of levels) and the distribution of the state transition probabilities; both linked with the two stages of the MLCTP method [20]). We use these two hyper-parameters as well in the feature space.

It is with these features that we can reliably capture the complexity of the sequential activity data at various levels of abstractions that could be linked with the skill of a person. Table 1 gives an overview of these metrics which are calculated for each of the p sensors over the whole session. In addition to this and to highlight various aspects of a session over a time, these rule metrics are also calculated over q partitions of a session. This allows for complete abstract representations of an entire session and eventually enable skill assessment.

**Skill Ranking: Ground Truth for Bootstrapping**

A logical approach for actual skill assessment would now be to train a statistical classification backend such that it would replicate ground truth assessments regarding skill. With our multi-level representation of activity data this would be straightforward through presenting pairs of activity data and their associated absolute skill labels to a standard training procedure, thereby effectively mapping the classification problem from identifying activities to predicting skill classes instead. This has been done before (e.g., [37, 38]), however, we do not believe that this is appropriate given the complexity of the problem.

Any automatic classification system requires some sort of ground truth annotation of sample data, be it for the actual training procedure or for the subsequent validation step (typically both). When it comes to skill assessment it is a widely known fact that expert assessment are typically not as reliable as they should be, which is referred to as judge bias. It is simply a very challenging task to reliably, i.e., in a reproducible way, produce absolute scorings of something as complex as the quality of activities, and thus the skills an actor actually possesses. In response to this, our framework has been designed to not require absolute skill or quality assessments as input for its training process. Instead, we limit ourselves to the absolute minimum annotation, namely relative skill rankings. Arguably, it is much easier for a human annotator to compare two instances of the same activities and to then judge which of the two was of better or worse quality. With such a weak annotation we can then automatically bootstrap our assessment framework for a specific domain.

**Skill Ranking: Automated Assessment**

Class annotations for bootstrapping the assessment system are derived from the relative differences between instances of activities of interest – subsequently referred to as attempts. As mentioned before, by using the differences we avoid any annotator bias which may arise from their own experience or intuition and effectively define a protocol for both bootstrapping assessment system for particular domains as well as for using and maintaining it (cf. case study).

Skill assessment is thus translated into a classification task where activity data from new attempts, represented by means of aforementioned features, is first classified regarding three coarser levels of i) improvement; ii) consistency; and iii) deterioration of skill. This classification is enough to sort the individual attempts based solely on the relative assessments. For example, an attempt classified as an improvement over

---

### Table 1: Structural features extracted from rule topologies for skill assessment (per segment; see text for explanation).

<table>
<thead>
<tr>
<th>No.</th>
<th>Feature Type</th>
<th>Dim.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity metrics of the induced rule hierarchy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>i)</td>
<td>Total number of nodes</td>
<td>1</td>
</tr>
<tr>
<td>ii)</td>
<td>Total number of levels</td>
<td>1</td>
</tr>
<tr>
<td>iii)</td>
<td>Total number of branching points</td>
<td>1</td>
</tr>
<tr>
<td>iv)</td>
<td>Average number of branching points per level</td>
<td>1</td>
</tr>
<tr>
<td>v)</td>
<td>Number of symbols used to create rule tree</td>
<td>1</td>
</tr>
<tr>
<td>Metrics associated with the induced transitions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vi)</td>
<td>Log-likelihood (hierarchical transition matrix): ( \sum \log(L(O</td>
<td>E_{O-1})) )</td>
</tr>
<tr>
<td>vii)</td>
<td>Log-likelihood (hierarchical transition matrix (diagonal)): ( \sum \log(\text{diag}(L(O</td>
<td>E_{O-1}))) )</td>
</tr>
<tr>
<td>viii)</td>
<td>Trace of the hierarchical transition matrix: ( \text{tr}(L(O</td>
<td>E_{O-1})) )</td>
</tr>
<tr>
<td>ix,x</td>
<td>Hyper-parameters of the induced rule structures and transition matrices</td>
<td>2</td>
</tr>
</tbody>
</table>

Total = 10 × p sensors × (1 overall + q partitions)
SESSION: SKILLS DEVELOPMENT

Figure 3: Example of how the relative predictions are used to bootstrap a classification on an arbitrary scale. Red line indicates new predicted position. Symbols $>$, $=$, and $<$ indicate relative predictions. Figure 3(a) shows $n = 23$ sorted reference attempts, possible insertion indices for a new attempt are given below $(-23, . . . , 23)$; (b) shows a new attempt classified as better than 3 attempts $(1, 2, 3)$ shown in green, similar to 5 $(4, 5, 6, 7, 8)$ in white, and worse than 15 in purple of the reference attempts, the resulting position is $-12$ (the same as attempt 6) resulting in an absolute prediction of 2.

every other attempt would clearly fall at the head of the ranking (see Figure 3c), possibly with an absolute skill assessment greater than those already witnessed within the reference (training) data set. An attempt mostly classified as similar to the others would fall towards the median of the reference data. By using a simple sorting algorithm described below we can suggest a position within the original group of attempts. We can further use the absolute skill assessment values for the original attempts, if provided, to map the ordered attempts to this given scale. Figure 3 also shows how these attempts may fall on a scale and how a new attempt may be classified with any of the given values, between any two, or at the limits of the group.

By taking a reference cohort of $n$ task attempts for which we know at least the relative assessments we can produce a ordered list. We then classify new attempts against all original attempts on a scale $l_i \in \{-1, 0, 1\}$ equating to deterioration, consistency, and improvement. The sum $Pos = \sum l_i$ of these assessments will provide the position within the ordered list for our new task attempt. We can see that the minimum for this sum is $-n$ and the maximum is $n$, where $Pos = -n$ would indicate ranking the new attempt before all other attempts and $Pos = n$ after all other.

Figure 3 demonstrates an example with 23 attempts in the training data (reference cohort), and two new attempts for which the pair-wise relative assessment values have been predicted. Relative predictions are indicated by the symbols $>$, $=$, and $<$. The resulting new attempt predicted position is shown by the red line. Additionally, the original 23 attempts also have absolute assessments on a scale of 1 to 5. By overlaying these values onto the sorted list and positioning the new attempts at their respective indices we can achieve both a relative assessment, and an absolute prediction value for a given skill assessment measure.

CASE STUDY: SURGICAL SKILLS

To validate our framework we consider a scenario of surgical skills training. As part of their training, medical students have to undergo dexterity practice in order to master basic instrument handling for elementary suturing skills (e.g., incisions, suturing, knot tying etc.). Objective quality feedback is an essential part of this training, which helps students to learn from mistakes and eventually to improve. However, manual evaluation of these exercises requires a significant amount of time and expert supervision, which places a strain on senior personnel and thus typically represents a significant bottle-neck in medical schools. Moreover, even experienced surgeons often do not agree on the (absolute) quality of what they observe. As such an element of judge bias further complicates an objective assessment, which renders the domain an ideal testbed for our automated approach.

We envision independent dexterity training for medical students using tissue simulators and sensor equipped instruments (Figure 4). Students work on a task that includes elementary suturing activities and subsequently our automated assessment approach analyses the recorded movement data w.r.t. their quality. Aiming for objective assessments every recorded attempt is compared to the – over time growing – comparison cohort for overall assessment. Furthermore, the system provides domain related quality feedback in terms of readily available skill assessment measures based on the Objective Structured Assessment of Technical Skills (OSATS) criteria, which is the gold standard for skill annotation in the domain [25].

Dataset

We attached miniature tri-axial accelerometers to a set of standard surgical instruments (Figure 4b), which allowed us to directly record instrument movements (sampling rate: 50 Hz). A total of 15 participants engaged in a total of 50 attempts of surgical activities. The same surgical procedures were replicated allowing us to focus the analysis on the skill associated with these sessions. The procedure is that of suturing a wound on a replica pad as can be seen in Figure 4a and usually entails an initial knot, followed by several stitches, and a final knot. There is significant variety between participants and attempts with regard to total duration, number of knots, success of knots, and use of each instrument.

The resulting dataset comprises of approx. $5\frac{1}{8}$ hours of sensor data per surgical tool (needle holder and forceps i.e.,
As expert skill assessment annotation in medical scenarios proves very costly in time, we obtained the time of one expert to watch all 50 sessions and provide the absolute skill assessment scores for 7 OSATS measures. The 7 OSATS criteria we use are: a) Respect for Tissue, b) Time and Motion, c) Instrument Handling, d) Suture Handling, e) Flow of Operation, f) Knowledge of Procedure, and g) Overall Performance. A summary of absolute OSATS assessments for the dataset is shown in Table 3. The ranking approach as described in the methodology section would not need this absolute expert assessment, requiring only relative differences between sessions, and thus less skilled annotators with weaker domain specific knowledge can be used, making annotation of the sessions more accessible.

Note that a standard low-level activity recognition task on this dataset is not very challenging. A total of 4 basic relevant activities would have to be recognised: ‘knot tying’, ‘active use of forceps’, ‘active use of needle holder’, and ‘performing a stitch’. For comparisons we bootstrapped a standard HAR pipeline (as described in Bulling et al. [12]) including basic preprocessing such as normalisation, frame extraction using a sliding window procedure (with frames of 1s each and 50% overlap between subsequent frames), extraction of statistical features resulting in 46-dimensional (23 statistical features [12] × p = 2 sensors) features. We then used a standard random forest classifier [11] with 100 trees using a stratified 10-fold cross-validation approach. Without further optimisation we achieved activity recognition accuracies of > 80%.

Classification of skill change

In this paper, we focus on abstract representations of activity data as sets of features for classifying change in quality, which we treat as indicator for skill. This allows us not to limit the general applicability of our approach. Our first set of validation experiments within the chosen domain of surgical skill assessment employs the proposed framework (with a lexicon size of 9 symbols for quantization) in combination with a standard Support Vector Machines (SVM; with RBF kernel) backend for classifying relative skill change. Optimised slack and the kernel parameter are selected using a standard grid-search procedure [35] separately for each of the OSATS criteria.

In the first experiment, we use a leave-ten-attempts-out cross validation protocol to classify differences in the OSATS scores into three categories i.e., three levels of change representing improvement (+4, +3, +2), consistency (+1, 0, −1), and deterioration of skill (−2, −3, −4) – equidistantly defined. Table 4a shows the mean weighted F1 scores (and the corresponding precision and recall values) across 5 distinct sets of 10 attempts each for all OSATS scores. In the second experiment, a leave-one-attempt-out cross validation protocol is employed to classify changes in OSATS scores into the same aforementioned three categories of skill change. Results for this experiment are also included in Table 4a.

Across all OSATS, we achieve a mean of the weighted F1 scores 0.64 and 0.71 for both of the aforementioned experiments respectively with a standard deviation of 0.05 each.

Table 2: Overview of annotated activities for n = 50 attempts.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Total annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using Forceps with Needle</td>
<td>1,706</td>
</tr>
<tr>
<td>Using Needle-Holder</td>
<td>644</td>
</tr>
<tr>
<td>Making a Stitch</td>
<td>957</td>
</tr>
<tr>
<td>Attempting a Knot</td>
<td>376</td>
</tr>
<tr>
<td>Overall Procedure</td>
<td>56</td>
</tr>
</tbody>
</table>

Table 3: Overview of expert annotated OSATS measures for all attempts.

<table>
<thead>
<tr>
<th>OSATS measure</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respect for Tissue</td>
<td>0</td>
<td>3</td>
<td>14</td>
<td>23</td>
<td>10</td>
</tr>
<tr>
<td>Time and Motion</td>
<td>7</td>
<td>10</td>
<td>17</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>Instrument Handling</td>
<td>5</td>
<td>13</td>
<td>16</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Suture Handling</td>
<td>9</td>
<td>9</td>
<td>18</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>Flow of Operation</td>
<td>1</td>
<td>8</td>
<td>18</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>Knowledge of Procedure</td>
<td>2</td>
<td>15</td>
<td>8</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>Overall Performance</td>
<td>3</td>
<td>12</td>
<td>15</td>
<td>17</td>
<td>3</td>
</tr>
</tbody>
</table>
Table 4: (a) Results of relative classification $x \in \{-1, 0, 1\}$ for all 50 attempts and 7 OSATS. (b) Skill ranking results showing weighted accuracy for correctly predicting OSATS values 1 to 5 for each measure, over 50 leave-one-out cross validation tests. Predictions within ±1 are considered correct.

<table>
<thead>
<tr>
<th>OSATS criteria</th>
<th>(a) Leave-ten-attempts-out</th>
<th>Leave-one-attempt-out</th>
<th>(b) Predicted OSATS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Weighted [P,R,F1 scores])</td>
<td>(Weighted [P,R,F1 scores])</td>
<td>(from ranked L-1-O attempts)</td>
</tr>
<tr>
<td>Respect for Tissue</td>
<td>[0.74, 0.73, 0.73]</td>
<td>[0.79, 0.78, 0.79]</td>
<td>0.87</td>
</tr>
<tr>
<td>Time and Motion</td>
<td>[0.67, 0.66, 0.66]</td>
<td>[0.73, 0.73, 0.73]</td>
<td>0.78</td>
</tr>
<tr>
<td>Instrument Handling</td>
<td>[0.64, 0.62, 0.62]</td>
<td>[0.67, 0.66, 0.66]</td>
<td>0.73</td>
</tr>
<tr>
<td>Suture Handling</td>
<td>[0.60, 0.57, 0.58]</td>
<td>[0.66, 0.67, 0.67]</td>
<td>0.64</td>
</tr>
<tr>
<td>Flow of Operation</td>
<td>[0.68, 0.66, 0.66]</td>
<td>[0.73, 0.73, 0.73]</td>
<td>0.69</td>
</tr>
<tr>
<td>Knowledge of Procedure</td>
<td>[0.61, 0.59, 0.59]</td>
<td>[0.66, 0.65, 0.66]</td>
<td>0.74</td>
</tr>
<tr>
<td>Overall Performance</td>
<td>[0.69, 0.67, 0.67]</td>
<td>[0.71, 0.71, 0.71]</td>
<td>0.76</td>
</tr>
<tr>
<td>Mean</td>
<td>[0.66, 0.64, 0.64]</td>
<td>[0.71, 0.71, 0.71]</td>
<td>0.74</td>
</tr>
<tr>
<td>Std.</td>
<td>[0.05, 0.05, 0.05]</td>
<td>[0.05, 0.05, 0.05]</td>
<td>0.07</td>
</tr>
</tbody>
</table>

This indicates a high performance of our framework for skill change classification across various indicators of skill quality. As can also be seen in Table 4a, our rule induction based framework has been successful in classifying skill changes at various levels of abstract representations for example, the mean weighted F1 scores for Respect for Tissue and Overall Performance (0.76 and 0.69 on average, for the two experiments respectively), indicate high performance for skills that can be respectively considered as: (i) low-level – a skill directly related to movements – and (ii) high-level – an abstract evaluation of the whole attempt as an abstract performance indicator.

**Skill Ranking and Absolute Assessment Scores**

In the second set of validation experiments we target the envisioned use case of an automatic skill assessment system for supporting independent dexterity training of medical students, assuming an existing system – as it has, for example, been bootstrapped as described before – new training sessions would be compared to the existing comparison cohort and quality feedback generated accordingly. From an experimentation point of view, we simulate the introduction of a new attempt into a reference cohort of – in this case 49 – previous attempts by using leave-one-attempt-out cross validation results to predict relative assessments. We now use the ranking procedure described in the previous section to position each of these left-out attempts within the reference cohort. This subsection describes stages (h), (i), and (j) from Figure 1.

With 49 relative predictions of one attempt versus the rest we can deduce an index position with which to insert this new attempt, thus providing a relative measure to assess the overall performance of the participant. Using the 7 described OSATS criteria, which range from physical skill to assumed knowledge, allows us to evaluate the use of this framework on a variety of different skill measures. For each of the 50 attempts (in a leave-one-attempt-out manner) we can examine the ranking given for each OSATS measure. To evaluate the success of this ranking we take the resultant ranking further by overlaying the original OSATS scores for the other 49 attempts. Using this scale and the new position of the single attempt, we can determine whether the ranked position provides a good indicator of the original OSATS measure, and thus a good ranking position in general. The classification results we have used to build the ranking maps absolute differences of ±1 to the class consistency, with greater differences up to ±4 to improvement or deterioration respectively. We therefore consider a predicted absolute value of ±1 as correct. These absolute values can be equal to, less than, greater than, or between any original value. This is an extension of the original OSATS range. See Figure 3c for an example of when a 4.5 may be predicted.

Table 4b shows the results of evaluating the absolute OSATS measures in this fashion. The success value for each OSATS measure is given by a weighted average of correct classification of each original score. Overall the success across all tested OSATS is 74%. This suggests a good ability for the framework suggested to be used for assessing skill expressed in a number of different manners. We also show an ability to reproduce an arbitrary absolute skill assessment measure.

Figure 5 in conjunction with Table 4 shows the distribution of these predictions for each OSATS measure and the weighted accuracy of these predictions. From the original expert annotation of the training data on a scale of 1 to 5, the relative assessment and ranking of each test session has been used to predict a new score within the ordering from $< 1$ (poorer than all others), within or between any existing group, up to $> 5$ (better than all existing). The diagonal region within black lines is considered to be a good prediction. In Figure 5h we show the overall prediction accuracies of all OSATS measures as well as the confusion matrices individually (a-g). We can see the majority of predictions falling within reasonable bounds, with only (d) showing lesser performance (64%) due to heavily weighted expert annotations.

**Summary**

The results of our validation studies demonstrate the practical value of our automated skill assessment framework. Responding to the inevitable judge bias problem we do not attempt to predict absolute skill scores. Instead we mimic a relative assessment procedure as it would realistically be pursued. In our concrete case study we demonstrated that our automated approach can correctly rank the quality of suturing attempts in more than three fourth of the cases (with varying accuracies for the two studied protocols). Whilst this still might seem “off” the “ground truth” the result should be contextualised with the overall practical value of automatically generated quality feedback for, in this case, medical students who – without our approach – often do not receive any or at least very biased feedback on their dexterity performance.
We achieve the presented results with a framework that does not incorporate any kind of domain knowledge at system design, which makes it applicable in other tasks and domains with minimal effort.

**Limitations and Future Work**

In this study, further challenges were also identified for future work. For example, the rule induction process only relies on sequential symbolic data while the length in time is not explicitly utilized. For some domains, this could be a very important skill defining parameter that could be either embedded in the model or used as a feature. Also, partitioning in the modelling – even though it achieves non-reliance on the domain – may also require tuning as an extra parameter for a new domain. These partitions and the induced rule hierarchies can be explored for respectively identifying parts of a session and activities of interest that play an important role for skill representation. Related to this, a top-down approach to identifying skill defining activities in time could also be of great importance for some applications.

**CONCLUSION**

In this paper, we presented an automated skill assessment framework for the context of human activity recognition systems. Skill assessment is hard to quantify and typically requires large amounts of domain specific prior knowledge that negatively affects generalisation. We proposed a completely data-driven skill quantification framework using a hierarchical and stochastic rule induction method that provided abstraction from the observed raw accelerometer data without any domain-specific knowledge. We demonstrated the ability of this system in a surgical activity data by assessing surgical skills of 15 participants over a total of 50 attempts.

The presented approach goes beyond previously proposed techniques. For example, the ClimbAX [22] system for analysing the performance of rock climbers uses a set of parameters specifically identified based on domain knowledge. Using the rule induction process defined in this paper, features of the data are automatically extracted that describe skill without the requirement of specific domain knowledge.

Our framework enables us to observe change in skill quality that opens up substantial opportunities for practical applications. For example, in healthcare scenarios a decision maker—a clinician or otherwise—would utilise the automatically extracted information about an individual’s skill improvement or degradation to adapt care or treatment programs. Specifically, tracking skill maintenance in Dementia patients is typically performed with respect to a certain set of activities. Our skill change assessment would enable effective comparisons against normal activity sessions to highlight the scale of the patient’s problems. The same assessment could also be utilised to understand a patient’s skill degradation — as an indicator of degrading condition – or improvement — as an indicator of response to medicines, for example. Alternatively, the ranking process can be utilised to identify the amount of skill degradation to prioritise medical care via performing comparative analysis of different patients’ skill.

**Acknowledgment**

This research was funded by the RCUK Research Hub on Social Inclusion through the Digital Economy (SiDE) (EP/G066019/1).
REFERENCES


