Generalized and efficient skill assessment from IMU data with applications in gymnastics and medical training

Article

Accepted Version

Generalized and Efficient Skill Assessment from IMU Data with Applications in Gymnastics and Medical Training

Aftab Khan, Toshiba Europe Limited, Bristol Research& Innovation Laboratory, UK
Sebastian Mellor, Newcastle University, UK
Rachel King, Department of Biomedical Engineering, University of Reading, UK
Balazs Janko and William Harwin, School of Systems Engineering, University of Reading, UK
R. Simon Sherratt, Department of Biomedical Engineering, University of Reading, UK
Ian Craddock, Department of Electrical and Electronic Engineering, University of Bristol, UK
Thomas Plotz, School of Interactive Computing, Georgia Institute of Technology, USA

Human activity recognition is progressing from automatically determining what a person is doing and when, to additionally analyzing the quality of these activities—typically referred to as skill assessment. In this chapter, we propose a new framework for skill assessment that generalizes across application domains and can be deployed for near-real-time applications. It is based on the notion of repeatability of activities defining skill. The analysis is based on two subsequent classification steps that analyze (1) movements or activities and (2) their qualities, that is, the actual skills of a human performing them. The first classifier is trained in either a supervised or unsupervised manner and provides confidence scores, which are then used for assessing skills. We evaluate the proposed method in two scenarios: gymnastics and surgical skill training of medical students. We demonstrate both the overall effectiveness and efficiency of the generalized assessment method, especially compared to previous work.

CCS Concepts: • Computing methodologies → Supervised learning by classification; Supervised learning by regression; • Information systems → Decision support systems; • Computing methodologies → Kernel methods; • Mathematics of computing → Time series analysis; • Applied computing → Health informatics;

Additional Key Words and Phrases: Activity recognition, skill assessment, machine learning for sports and health analytics

This work was partly performed under the SPHERE IRC funded by the UK Engineering and Physical Sciences Research Council (EPSRC) [Grant EP/K031910/1], and the RCUK Research Hub on Social Inclusion through the Digital Economy (SiDE) [Grant EP/G066019/1]. Authors' addresses: A. Khan (corresponding author), Toshiba Europe Limited, Bristol Research & Innovation Laboratory, Toshiba Europe Limited, 32 Queen Square, Bristol, UK; email: aftab.khan@toshiba-bril.com; S. Mellor, Newcastle University, Newcastle, UK; email: seb@jumpingrivers.com; R. King and R. S. Sherratt, Department of Biomedical Engineering, University of Reading, Reading, UK; emails: {rachel.king, r.s.sherratt}@reading.ac.uk; B. Janko and W. Harwin, School of Systems Engineering, University of Reading, Reading, UK; emails: {b.janko,w.s.harwin}@reading.ac.uk; I. Craddock, Department of Electrical and Electronic Engineering, University of Bristol, Bristol, UK; email: Ian.Craddock@bristol.ac.uk; T. Plötz, School of Interactive Computing, Georgia Institute of Technology, Atlanta; email: thomas.ploetz@gatech.edu.

1 INTRODUCTION

Human activity recognition (HAR) is a core component of many ubiquitous computing systems that aims to automatically infer the type of activities a person is engaging in and their spatio-temporal tags, that is, when and where an activity of interest happens. Inertial measurement units (IMUs) are typically the modality of choice for capturing activities, which is largely due to practical reasons, such as ubiquitous availability through smartphones or mainstream wearables, location independence as devices are typically with a user most of the time, and privacy preservation during recording (no cameras). Often machine learning methods are employed for sensor data analysis (cf., e.g., [8]). HAR has become a mature research field and the community is moving on from developing systems that help explore *what* and *when* something of interest is happening towards *how (well)* it is being performed. This emerging field is referred to as quality or skill assessment and the first systems have been developed for automated assessments in various sports, thereby providing insights into an athlete's skill (e.g., [26, 31]). In addition, skill assessment systems have been developed for training procedures in, for example, medical settings, where students learn and master surgical skills with automatically generated feedback (e.g., [38, 47]).

Automated quality or skill assessment is a challenging task and is therefore often constrained to specific domains. For example, in order to assess how well a patient is recovering from a leg injury, the quality of movements is analyzed. In such scenarios, specific sensing solutions need to be developed that are feasible for recording without burdening the patient unnecessarily to wear and maintain the device. Furthermore, specific movement parameters that are relevant for assessing the rehabilitation progress are monitored, which have to be defined by medical domain experts. As such, these skill assessment systems are often too specialized to be generalizable to other conditions or domains. System designers need to start over when moving to another domain, which comes with substantial effort and costs. A generalized notion of skill assessment is highly desirable.

The first steps towards generalizable skill assessment have already been taken. The underlying assumption is that skill (or quality) represents an inherent parameter that can be assessed when comparing activities that were recorded in similar circumstances. By fixing the activities performed, the automated assessment of sensor data streams can be focused on their quality, which leads to a technical formulation of the notion of skill. For example, Khan et al. [22] hypothesized that skill manifests at different levels of abstraction, which led to the development of a hierarchical assessment system that analyzes movement data at different levels of temporal context.

Such skill assessment frameworks are complex, which limits their real-time inference capabilities. This is a problem for systems that shall be used, for example, in sports coaching scenarios where an athlete requires immediate feedback in order to make a training session effective. More rapid availability of analysis results not only enables quicker feedback but also is a prerequisite for live coaching, for example; instead of providing after-the-fact analysis, a system could actively guide a user towards high-quality activities and thus facilitate skill acquisition.

Aiming for such application scenarios, in this article we propose an alternative definition of skill/quality of activities that enables less complex assessment systems. We argue that the quality of activities and thus the mastery (skill) of a person can be directly assessed by analyzing the way an activity is *repeated* individually or as part of a larger session. For example, if someone is highly skilled at performing a certain task, their repeat performance of the same task will be consistently similar, whereas for a novice one would expect high fluctuation. We present a new method and system for generalized, efficient skill assessment. The aim is to automatically determine skill through quality assessments of activity data, recorded through body-worn IMUs, in a manner that is domain independent and (near) real time. The method employs a two-tier classification mechanism with no domain-related knowledge embedded in the framework. Generalized skill classification is performed using confidence scores of a low-level statistical classifier that highlights the certainty in classifying low-level movements or activities. Feature extraction is performed using these confidence scores for classifying skill levels. Since the method is independent of both domain and underlying recognition methods—relying only on the activity data—it can easily be generalized. The proposed method is efficient and can be applied for assessing the quality of activities in near real time

Method development is motivated by a skill and quality assessment scenario in gymnastics, where athletes' performances are (manually) judged through certain predefined criteria. Aiming for automated coaching, a quality assessment system needs to analyze movement data both effectively and efficiently. Accurate, objective assessments need to be provided immediately after a gymnastics exercise has been performed, for which the presented system can be used. We demonstrate the effectiveness of the proposed method and system in a case study where five gymnasts, each wearing five tri-axial accelerometers on their waists and limbs, performed a range of routines that were then judged by a human expert. In order to demonstrate the generalizability of the proposed method, we also used it for analyzing an existing dataset of accelerometer-based surgical skill assessment [22]. For both case studies we not only evaluate the accuracy of the proposed method but also analyze the computational effort required for the automated assessments. The results demonstrate that the proposed system enables generalized automated skill assessment for both application domains with high accuracy, thereby being computationally efficient, which enables near-real-time analysis. Only minimal, technical adjustments are necessary to transfer between application domains, which is promising for skill assessment in general for the wider research community.

2 BACKGROUND

HAR is now a mature research field with a wealth of applications based on a range of sensing and modelling techniques. The wider ubiquitous computing literature is a rich resource for technical details as well as descriptions of specific application scenarios. Our work builds on this existing body of knowledge. Given that many of the HAR methods can now be considered common knowledge, and for the sake of brevity and focus of our presentation, we refrain from re-iterating the standard literature and refer the interested reader to more general surveys such as [8, 11, 28]. Instead, in what follows, we focus on existing work that is related specifically to automated skill and quality assessment from body-worn sensor data.

2.1 Skill Assessment vs. Activity Recognition

HAR typically focuses on the automatic determination of what actions or activities a person is engaging in and when. Activity recognition as such contributes to the broader area of behavior analysis. Behavior analysis in other disciplines, such as psychology or animal health and veterinary medicine, typically uses clear operational definitions of the phenomena that are of relevance—often formalized in, for example, behavior ethograms, taxonomies, or topographies. HAR, as it is, for example, conducted for human-computer interaction applications, often uses less strict, more common-sense-oriented definitions of the phenomena that are to be recognized. Only few frameworks exist that rigorously define the taxonomy of human activities. Bobick has defined three elementary levels that are of relevance for HAR: (1) movement, (2) action, and (3) activity [5]. Refined versions of this taxonomy exist [30], each of which is based on the hierarchy of contextual information that is required to describe the respective levels. From a technical standpoint, HAR translates into the automated analysis of time-series data that are recorded from a variety of sensors including cameras or body-worn IMUs. The task is to segment these time-series data into semantically cohesive portions, which are then classified; that is, labels are assigned automatically.

Skill assessment, also referred to as quality assessment, goes beyond traditional activity recognition as it (additionally) aims at reasoning about the quality of what happened, thereby ranging from elementary movements to complex activities or behaviors [22]. The focus of skill assessment is on the process rather than on the result of the underlying activities. As such, the technical analysis focuses on other aspects of the sensor data and different parameters are quantified. These either are re-defined by domain experts, such as certain movement parameters, or cover more specific signal parameters such as smoothness or energy. With pre-defined parameters, typically shallow definitions of skill are used that measure relevant parameters on fixed-length temporal contexts (e.g., [26, 42]). For generalized skill assessment, generic signal representations are used that are not limited to specific domains. The state of the art here is to employ hierarchical notions of skill, that is, to analyze the time-series data in a range of temporal contexts [22]. Details are given below.

2.2 Areas of Skill Assessment

Two major areas of skill assessment have so far been described in the literature: (1) sports and (2) health and well-being. Without aiming to provide an exhaustive survey of the fields, which would be beyond the scope of this article, in what follows we summarize the key ideas for both areas.

2.21 Skill Assessment in Sports. Activity recognition in sports is typically relatively straightforward, aiming to, for example, recognize individual elements of the particular sport being analyzed. Examples of this include the classification of different serve types and other parameters of table tennis [4], classification of cricket shots [23], and classification of skateboard tricks [18]. The majority of these applications are based on supervised learning tasks. Due to the distinct character of the various activities of interest, high classification accuracies are achieved. More interesting and also more challenging than the mere characterization of specific elements of a sport is the automated characterization of their quality. Obvious applications for such skill assessments in sports could be within automated coaching. For any athlete it is important to receive feedback on their exercises—for health and safety reasons (e.g., to prevent injuries) and to improve their capabilities (e.g., to become more competitive). The majority of skill assessment applications in sports are based on the analysis of sets of movement parameters that are specific to the sport. For example, Ladha et al. [26] presented a system that measured the four key parameters of a climber's movements: (1) power, (2) control, (3) stability, and (4) speed. The actual assessment through such (or other) key parameters is then either based on regression approaches that predict concrete values—either of the key parameters or, derived from it, of quality scores as human judges would provide them—or on translating the problem to a conventional classification task where the classes are defined by, for example, levels of expertise.

2.22 Skill Assessment in Well-Being and Healthcare. The second major area for skill assessment relates to automated tracking of progress or decline in certain health and well-being challenges. For example, it is of great importance for doctors and carers to keep track of the cognitive decline of a dementia patient. Only if an accurate and continuously updated assessment of a patient's skills is maintained can the care and treatment program be tailored to the individual's needs. More and more technical systems are employed for health and well-being assessments outside clinical environments (e.g., [21]). Another application domain is

rehabilitation, where automated assessment methods aim at analyzing the overall quality of movements as it is of relevance for treatment programs after recovering from medical events such as stroke [13].

The majority of approaches for skill or quality assessment in health and well-being are tailored towards individual conditions. This is understandable as the main motivation for using such systems is to automatically monitor the change of specifically relevant parameters. Similar to the sports domain, these parameters are defined by domain experts (medics), and statistics about their occurrences during, for example, an exercise regime in rehabilitation are then reported to both patients and carers providing the base for optimization. Again, the technical process of skill (quality) assessment is translated into either a regression or a classification task and a wide range of techniques are employed.

2.3 Generalized Skill Assessment

The vast majority of automated skill assessment methods are domain specific; that is, they are designed specifically for individual application areas. This means that, for example, all skill- or quality-relevant parameters are pre-defined in collaboration with domain experts. For the aforementioned climbing assessment system [26], Olympic-level coaches were interviewed to specify what the key differences between an expert climber and a beginner are. This expert knowledge was then mapped to what was possible to sense, in this case using bodyworn accelerometers, and specific analysis methods (signal processing and machine learning) were developed. While such a procedure can be very successful for the specific application domain, it usually does not generalize towards other domains. In essence, system design has to start (almost) from "scratch" for every new application domain, which is the motivation for some related work—and this paper—aiming at generalized skill assessment.

Doughty et al. [17] introduced a vision-based method for skill ranking of videos from four different domains, using a supervised approach. Other work in computer vision has been focused on a specific domain, such as Olympic events [32]. An interesting and generalizable skill definition is provided by Velloso et al. [43] in which a benchmark-driven approach is introduced for assessing skill, that is, howwell an activity adheres to a benchmark specification. However, the technical approach is domain specific and evaluated using weight-lifting exercises.

Probably the most extensive approach to generalized skill assessment was presented by Khan et al. [22]. The idea there was that skill (equivalent to quality) can be considered an intrinsic parameter that can be learned through a weakly supervised analysis framework that exploits pairwise comparisons of activity sessions. It was argued that skill becomes manifest at various levels of temporal context that cannot be predefined per se without limiting the overall approach. This observation was exploited for a generalized assessment scheme that learns skill parameters using a hierarchical rule induction scheme [24]. Starting from a symbolic representation of raw sensor data [3], more and more abstract representations are generated through hierarchical aggregation. Based on such a multi-level representation of the sensor data, features are extracted using rule structures that include metrics directly associated with the complexity of the induced rules. For example, the number of levels determines the complexity of skill involved in this approach, which directly affects the size of feature vectors used for skill assessment. Features extracted using these rule structures are then used to train skill classification models, evaluated on a surgical skill assessment task as it is common for medical training. Other skill assessment methods also employ hierarchical analysis schemes (e.g., [37]), in which an activity assessment chain is proposed that analyzes both spatial and temporal semantics of movements.

3 GENERALIZABLE AND EFFICIENT SKILL ASSESSMENT

Compared to previous work in the field, in this article we propose an alternative notion of skill that fundamentally drives our framework for skill assessment. It mainly utilizes the repeatability and consistency aspects of skill and can be defined as follows:

Skill: The ability to do something, repeatedly, and with consistency, that is, low variability in execution

The emphasis here is on "repeatedly" such that, for example, a particular movement is more likely to be repeated with strong similarity (i.e., "consistency") by an individual who is skilled in that particular movement. This notion of repeatability is well accepted in several domains including both sports and healthcare. For example, in [1], the repeatability aspect in gymnastics is studied. It is shown that expert gymnasts showed better repeatability of the ankle trajectory. Similarly, in the coaching literature of cricket [45], 10,000 repetitions of batting shots are mentioned in order to play these shots at the highest skill level (i.e., play these shots instinctively). For stroke rehabilitation, several studies have shown that repetitive practice can improve strength after stroke [14, 16].

This general principle of repeatability is what—at a higher level of abstraction—we exploit in this article to develop a generalizable and efficient skill assessment method. Generalizable refers to the minimal, largely logistical effort that is necessary to transfer the skill assessment system from one application domain to another, and efficient refers to the requirement of many application domains where rapid feedback is needed. At the technical level we translate this notion of skill into utilizing confidence scores associated with classifying low-level actions or movements, which results in non-complex metrics that are then used for assessing skill. These metrics are generalizable as the notions of confidence scores, repeatability, and related

consistency are generic and thus not limited to specific application domains. For example, high variability in confidence scores can provide a good indication for low skill and vice versa.

The proposed definition is very different compared to the previously proposed generalized framework [22] where a hierarchical definition of skill was used to develop a generalizable skill assessment system (see Section 2.3 for a summary). According to that, skill could be represented with a mixture of rule structures computed over multiple partitions of the underlying sensor data where each rule structure was computed using a stochastic rule induction process that utilized sequential movement symbols.

Our new definition of skill results in a framework that differs from the previous one in two major aspects

(1) The model architecture is flat (without requiring an inefficient search through a large set of hierarchical rule trees for assessing skill [22]), resulting in models that are substantially less complex, which is beneficial for efficient evaluation.

(2) Assessment is based on confidence scores provided by low-level movement classifiers, which generalize well beyond application domain boundaries because most higher-level activities are based on such more general descriptors.

Figure 1 gives an overview of the proposed skill assessment system, and the following sections provide details of the two main system components.

3.1 Low-Level Movement Classification/Activity Recognition

To collect data for movement classification or activity recognition, IMUs can be used to record raw motion data, such as microelectromechanical (MEM) accelerometers and gyroscopes. Given their omnipresence through integration into many consumer devices such as smartphones or smartwatches, many HAR systems make use of such sensors to automatically identify various types of activities. In response to this, yet without limiting generalizability, our framework targets the analysis of three-axis accelerometer data; however, our framework can be used for any time-series data including most sensing solutions in the wider ubiquitous computing domain.

Two methods that can be used to treat the sensing data after pre-processing include:

(1) Segmenting the data into activity windows (sliding-window-based approach may also be used in this context) for further feature extraction and classification [8]. Based on such an activity recognition chain, movement data can be automatically transcribed into sequences of low-level activities—if annotated example data exists at this level of granularity.

(2) Converting the movement data into symbolic representations, for example, through unsupervised discretization methods (e.g., [3, 29]), which then serve as the basis for explicit segmentation procedures prior to feature extraction and classification as before.

The framework we propose contains both variants of low-level preprocessing (upper and lower parts of the central section in Figure 1, respectively). Depending on whether low-level ground truth annotation of sample data is available during model training, either the supervised (upper part) or the unsupervised (lower part) route is chosen.

For the supervised training path, standard activity recognizers can be employed and no particular preference is given here. State-of-the-art activity recognizers are distribution-based features, such as the Empirical Cumulative Distribution Function (ECDF) [19, 25, 35], that are computed per analysis frame. We employ ECDF features combined with other standard statistical features per segment (summarized in Table 1) that are then forwarded to the classifier.

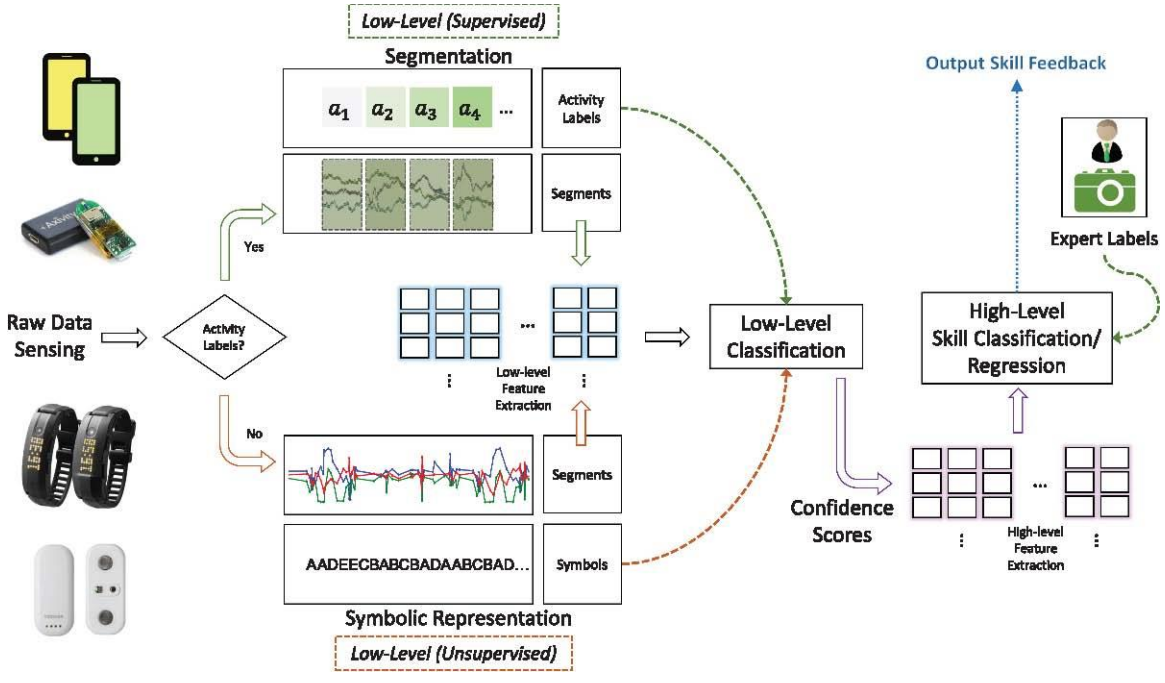


Fig. 1. Overview of the proposed skill assessment framework illustrating the two possible sub-approaches for the low-level classification stage (i.e., supervised and unsupervised activity/movement classification) and the high-level skill classification/ regression stage

Table 1. Features for Low-Level Symbol or Activity Classification.

Feature	Equation	Dimensions
Empirical Cumulative Distribution Function (ECDF)	$\{x, \exists j : P_c^j(x) = p_j\}$	20
Energy	$\frac{1}{N}\sum x^2$	1
Entropy	$\sum x \log(\lvert x \rvert)$	1
Mean (μ)	$\frac{1}{N}\sum x$	3
Standard Deviation	$\sqrt{\frac{1}{N}\sum (x_i - \mu)^2}$	3
Fast Fourier Transform (FFT)	$\sum x_j e^{-2\pi ijk/N}$	99
Correlation Coefficient	$\max(\frac{1}{N}\sum A_i \cdot B_{i-j})$	3
$\sum$		130

The statistical classification backend of our framework performs the automated recognition of low-level activities or symbol sequences as each is extracted in the segmentation stage. Note that in the context of an unsupervised approach, resulting symbols are used as representations ("labels") of the underlying sensor data and features are computed as before (Table 1). The resulting feature vectors, together with the symbols that serve as labels, are then fed into the classifier training stage. Depending on the complexity of a domain, the vocabulary size for these symbols may vary, which influences the output of the next processing stage (see below).

An important addition we contribute to the low-level classification of either activities or symbols is the incorporation and exploitation of the classifiers' confidence scores, which form the basis for the actual skill assessment. There are different ways of calculating confidence scores as, for example, reported for Support Vector Machines [33, 46]. The majority of the classifiers can be configured to provide these scores in the form of posterior probabilities for multi-class recognition scenarios:

$$\hat{Y} = \underset{y=1,..,M}{\operatorname{argmin}} \sum_{i=1}^{M} \hat{P}(i|f)C(y|i),$$

where Yˆ is the predicted classification and M represents the total number of classes, that is, either low-level activities or, in the unsupervised case, symbols. Pˆ(i|f) is the posterior probability of class i, using a classifier (e.g., SVM) for observation f∈Fv (where Fv is generated using statistical features as exemplified above) and C(y|i) is the cost of classifying the given observation as y when its true class is i. In this article, we use SVM for the low-level classification step. We then approximate the posterior classification probabilities using sigmoid regression applied to the SVM output [33].

3.2 High-Level Skill Assessment

We aim for skill assessment that is generalizable across application domains. Based on the confidence scores for automatically extracted segments as provided by the first stage of our analysis pipeline, the final stage of actual skill assessment is now extracting further information from the results of the first stage (right part of Figure 1). We generalize the concept of feature extraction as it is applied to sensor data streams (first stage of our pipeline) to the confidence score representations of the input data (at segment level). The first stage of our processing pipeline includes segmentation, which compresses the input data in time. The result is a more compact, semantically enriched representation of the sensor data that in itself represents a time series. The second stage of our pipeline analyzes this time-series data, in a similar way as the first stage does.

Given that segmented data is fed into the first stage of the framework, no further segmentation is needed and we proceed directly with feature extraction on the sequence of confidence scores. Standard statistical features are used here, resulting in a d=4 dimensional feature vector per symbol (unsupervised case) or activity (supervised case). This is generated using:

- posterior probabilities for individual symbols or activities (note that different lexicon sizes for the initial quantization step could be used in the unsupervised case, but an inventory of nine symbols represents a reasonable compromise between accuracy in representation and complexity [22] and thus efficiency of quantization) and
- four statistical features; mean, standard deviation, variance, and median.

In cases where multiple sensors are used, the final dimensionality of the feature representation multiplies accordingly. For example, if two sensors and nine symbols are used, for the unsupervised case, $d = 72$ (i.e., nine confidence scores per sample, and four features per symbol).

In the final stage, we feed the aforementioned feature vectors into the final classification backend using a standard statistical classifier or regressor; in Section 4, we evaluate the effectiveness of common variants that derive the final skill metrics. This classification stage is trained in a supervised fashion and thus requires input from expert human annotators at the skill level (see Section 4 for details in practical scenarios).

Overall, the proposed approach is substantially streamlined, compared to previous skill assessment schemes that required complex, deeply hierarchical modeling approaches. The benefits of our less complex approach are linked to faster training and, more importantly, to faster inference during deployment, which leads to (near) real-time assessments. Given that no domain knowledge is required for model design, training, and evaluation, the proposed methods are generalizable for sensor-data-based application scenarios.

3.3 Discussion

With a view on system effectiveness and efficiency, it is worth exploring in more detail the structural differences between the method we have proposed and the baseline approach introduced by Khan et al. [22].

3.31 Feature Dimensionality. With no dependency on multi-level representations, our framework is more efficient for skill assessment. Two subsequent processing stages produce rich and meaningful representations of the raw input data, in contrast to the pyramidal representations extracted previously. As a consequence, feature vectors have controlled, fixed sizes. Unlike the baseline approach, in the method presented here, input data are not required to be partitioned into several sections and features are extracted globally. If previously there were *s* sensors used, with p partitions, and if *d* dimensional features were extracted, then the dimensions of the feature vector are s×p×d. Note that for the majority of applications multiple sensors are used, which amplifies the problem. In our work, the size of the extracted feature vector is directly linked to the number of unique activities (or symbols), for which confidence scores are generated as a result of the low-level classification step, and independent of the number of sensors used and therefore not relying on pre-partitioning of the data. As such, our method provides a highly scalable solution where the dimensionality of the feature vectors remains controlled.

3.32 Skill Metrics. The problem of determining domain-specific metrics for skill assessment is avoided in our proposed approach as only the confidence scores from the low-level classification mechanism are used for high level skill classification. In the baseline approach, this was achieved by utilizing features from the rule induction process, which required learning rule structures and associated probabilistic relationships between various sets of movements in the symbol space. In contrast, we now perform

this using a classification procedure—a flat process, resulting in a faster availability of posterior probabilities, which are then utilized as skill metrics for assessing skill.

3.33 Reliance on Sequential Representations. One of the limitations discussed in the baseline approach is related to the requirement of sequential data for performing hierarchical rule induction. It is very difficult, and sometimes impossible, to process every dataset at such levels of representations—adding an extra level of complexity that can be avoided using our proposed approach. For example, in scenarios where activity data is separately collected (i.e., pre-segmented), rather than in a long session with multiple activities throughout a procedure/task, a meaningful rule topology cannot be generated (mainly due to very short temporal contexts). Using our approach, a trained activity recognition model can be used to test activities individually and produce confidence scores that can be used for skill assessment. For datasets where sequential representations are viable, and where low-level activities are practically unavailable (or are too abstract), our approach can be utilized to perform low-level movement classification. In such cases the repeatability aspect, associated with skill, is determined by these primitive movements and the confidence with which they are classified.

4 CASE STUDIES AND EXPERIMENTAL RESULTS

In order to evaluate the effectiveness, efficiency, and generalization capabilities of the proposed framework, we conducted two case studies. These serve as realistic examples of typical skill assessment scenarios. The framework has been deployed "as is" for both domains with only minimal, logistical modifications implemented related to specifics of sensor hardware and so forth. No further adaptations have been introduced when switching between domains. We explore computational efficiency with a view on the envisioned application scenarios where rapid feedback generation is required as outlined before. Where possible, we also draw comparisons to previously proposed skill assessment frameworks illustrating the general progress made in the field of automated skill assessment.

For the experimental evaluation of the two case studies, we focus on two levels: (1) overall skill assessment capabilities and efficiency of operation and (2) detailed analysis of individual components of the assessment pipeline as explained in Section 3. In what follows we will present the results separately for both case studies and will draw comparative conclusions.

Table 2. Overview of the Collected Participant Data for the Gymnastics Use Case

	Gender	Age	Height (mm)	Weight (Kg)	Total Sessions	Unique Activities	Total Activities	Skill Level (Average Deductions)
P1	Female	11	1,371	28.9	4	7	75	0.2286 ± 0.0705
P2	Female	12	1,546	35.2	3	9	75	0.1500 ± 0.0934
P3	Female	16	1,572	48.7	1	7	21	0.1667 ± 0.0658
P4	Male	19	1,792	73.6	1	7	21	0.1286 ± 0.0845
P5	Male	20	1,815	73.6	1	7	21	0.1381 ± 0.0973

4.1 Assessment of Gymnastics Skills

Our first case study focuses on the assessment of gymnastics skills. Gymnastics is a very popular but demanding sport requiring dynamic strength, flexibility, and balance to perform complex and challenging tasks. Gymnasts are also vulnerable to injury, so a gymnast benefits greatly by ensuring that skills are practiced and performed correctly to reduce the chances of sustaining an injury [6]. In gymnastics, deterministic models have been used to relate performance outcomes (e.g., points awarded during a competition) and the biomechanical features that produce the outcome [12]. These models are specific to the activity and built based on expert knowledge and are possibly subjective. Accelerometry has been investigated for estimation of general athletic skill by examining body sway [27]. More specifically towards gymnastics and injury prevention, accelerometer data were investigated to determine how well they are able to estimate load during impacts [2, 9, 40].

For our gymnastics case study we recruited a total of five participants with various levels of experience and expertise. Each participant was equipped with five tri-axial accelerometers that were attached to all four limbs and the torsi of the athletes, and performed multiple repetitions (in sessions) of at least seven unique exercises as they are common in apparatus gymnastics. Ethical approval was granted by the Research Ethics Committee of the university that oversaw the case study. All participants provided informed written consent, and for participants under the age of 16, informed written consent was also provided by a parent or guardian. Table 2 summarizes the dataset that was collected during the case study.

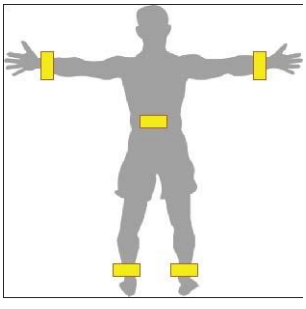
The nine different activities that were performed by the athletes can briefly be described as follows:

(1) Straddle to handstand (HS): On gymnastic blocks, starting from the straddle sit position, the legs are raised till they are above the gymnast's head in the handstand position. This position is held and then the gymnast returns to

the straddle sit position.

(2) Back flic (BF): From a stationary standing position, the gymnast jumps backwards, inverting the body such that the hands are next to contact the floor. The gymnast then pushes off from the hands to return to a standing position.
(3) Round-off back flic (ROBF): From a running start, the gymnast performs a half cartwheel, landing with both feet making contact with the floor at the same time and rotating such that the gymnast is facing back along the direction they just traveled (this is the round-off). Maintaining the momentum, the gymnast then performs a back flic.
(4) Backwards walkover (BWO): From a stationary standing position the gymnast leans backwards in a controlled movement leading with one leg till the hands meet the floor. The leading leg continues over the gymnast with the trailing leg following till the foot meets the floor and the gymnast returns to a standing position.
(5) Round-off tuck-back (ROTB): From a running start, the gymnast performs a round-off, then jumps backwards, tucking the knees into the chest and performing one backwards rotation before landing.
(6) Tuck-front somersault (TFS): From a running start, the gymnast jumps up, tucking the knees into the chest and performing one forward rotation of the body and landing with both feet together.
(7) Round-off back flic tuck-back (ROFTB): From a running start, the gymnast performs a round-off back flic, then jumps backwards, tucking the knees into the chest and performing one backwards rotation before landing.
(8) Round-off straight-back (ROSB): From a running start, the gymnast performs a round-off, then jumps backwards, keeping the body straight and performing one backwards rotation before landing.
(9) Multiple backward flics (MF): From a running start, the gymnast performs a round-off, then three back flics consecutively.

Fig. 2. Illustrations of the on-body locations of the five sensors.



We collected sensing data over the course of six sessions. On the day of data collection, participating gymnasts had to present in full good health and without injuries. The five accelerometers were attached to the waists, right and left wrists, and right and left ankles of participating gymnasts. Anthropometric data was also collected from each participant describing their age, weight, height, length of the arms and legs, and location of the sensors with respect to the body (Figure 2). However, the anthropometric data was not used in the context of this work. During each session, participants were asked to perform a selection of the described activities based on their experience and skill level. Each activity was performed three consecutive times. In addition to the sensor data, videos of the activities were recorded for post-analysis and scoring (see below). Three example activities are illustrated in Figure 3. For the safety of the gymnasts, the front-tuck somersault, round-off tuck-back, round-off straight-back, and multiple flics were performed on an inflatable gymnastics air floor.

All data was collected at 100Hz, with an accelerometer data range of +/-16g and a resolution of 16 bits. Sensor data were synchronized [34] and annotated using the collected video using ELAN software [44]. After the data were collected, each activity was scored using the deductions method by a gymnastics coach by viewing the videos. For each activity, {0, 0.1, 0.2, 0.3} of a point is deducted from a highest possible score of 1 based on five criteria: technical skill, momentum, flexibility, control, and style. The lower the deduction, the better the activity is performed, as shown in Tables 2 and 3. This scoring is representative of the official scoring system of British gymnastics[1] where a gymnast will start with a score of 10.0 and have points deducted for various faults in the execution. This scoring is performed by human experts (judges). In this way the scoring used in our experiments is similar to that which the gymnasts will be familiar with while training. It can be seen that participant 4 has the lowest amount of average deduction, which represents the highest-skill participant, compared with participant 1, who has the highest average deduction, representing the lowest skill. Skill scores are collected for each individual activity and therefore the proposed framework is used to automatically classify/estimate these scores per activity performed as detailed below.

[1]https://www.british-gymnastics.org/scoring-guide.

Low-level activity annotation is provided for the gymnastics case study and as such the assessment system follows the supervised path in the first stage of the pipeline (Figure 1). For overall effectiveness it is of importance that the first stage—supervised classification—works properly. This will be evaluated first before we discuss the results of the second stage, that is, the actual skill classification and runtime efficiency.

Fig. 3. Illustrations of three activities performed by the participants.

(a) Backwards walkover.

(b) Round-off back flic.

Table 3. Overview of the Gymnastic Routines for All Attempts and Associated Skill Levels

Activity	Performed by (# Participants)	Total Examples (# Activities)	Average Skill Level (μ, σ Deductions/Activity)
Straddle to handstand	2	21	0.2000 ± 0.0949
Back flic	5	30	0.2400 ± 0.2358
Round-off back flic	5	30	0.1800 ± 0.0847
Backwards walkover	5	30	0.2167 ± 0.1783
Round-off tuck-back	5	30	0.1867 ± 0.0900
Tuck-front somersault	5	30	0.1867 ± 0.1106
Round-off back flic tuck-back	4	18	0.1944 ± 0.1056
Round-off straight-back	4	18	0.1167 ± 0.0985
Multiple backward flics	2	6	0.1833 ± 0.0753
$\sum$		213	

4.1.1 Low-Level Classification-Activity Recognition Results. Data related to gymnastics routines was separately collected and as such no other segmentation method was required. For these segments we then computed statistical features as described in Section 3.1.

In line with the general routine for validating HAR systems [20], we employed three types of cross-validation schemes to evaluate the activity recognition (and later skill assessment) performance:

(i) Validation type 1: k-fold cross-validation in which the training set in each fold contains randomized data from all participants and all movement types. This type of validation is the most common in the field of HAR using body-worn sensors where typically only limited sets of annotated sample data are available and thus economic use of it is imperative. We report these results as comparison baseline.

(ii) Validation type 2: Leave-one-user-out cross-validation in which the test set in each run contains data only from one participant with training performed using the data from all the other users. This validation protocol is considered the hardest yet most realistic [20] because it tests how well a model generalizes towards users whose data were not part of the training step.

(iii) Validation type 3: Leave-one-attempt-out cross-validation in which the training set contains data from the same participant and different activities. Multiple tests are performed using individual samples of all the activities. This could be considered as a traditional leave-one-out cross-validation scheme but repeated separately for all users.

For activity classification we employed a standard support vector machine classifier with RBF kernel. Results for all three cross-validation schemes are shown in Figure 4. The difference in the number of activities for the leave-one-user-out cross-validation scheme is due to the fact that not all participants performed all activities. As such, certain cases were excluded in which a test gymnast performed activities that other participants did not perform. Validation type 3 (i.e., the leave-one-activity-out cross-validation scheme) produced the best performance with a class-weighted F1 score of 0.9765, *k*-fold produced a similar performance of 0.9622 (with $k = 10$), while the leave-one-user-out validation scheme had an F1 score of 0.9131. However, in all three cases, AR performance is at a sufficiently high level for skill assessment.

4.1.2 Skill Assessment Results. The actual skill assessment is performed in the second stage of the processing pipeline (high-level classification). Posterior probabilities for the nine classes (activities) in the first classification stage are translated into time series and then fed into the second stage of the pipeline. This is achieved by curating a sequence of posterior probabilities where each sample represents a set of confidence scores for all possible activities for that segment. Feature extraction is then performed using this sequence as outlined in Section 3.2. It is worth noting that in the context of a sliding-window-based approach, integration of a null class would be necessary, for example, to cover periods of still standing between the activities of interest. The null-class-related posterior probabilities may be ignored when performing skill assessment in such a scenario. In this work, the first CV case has nine activities and therefore at each prediction stage, there are nine posteriors that are directly used as features for skill assessment. In the second CV case, as the number of activities is reduced to five, the corresponding number of features is also reduced.

Since the skill ground-truth scores are in the form of deductions, that is, $l \in \{0, 0.1, \ldots, 1\}$, we can perform both classifications using the observed skill deductions or regression. Both of these approaches are extensively evaluated below using various classification and regression strategies.

Classification results. Classification results for skill assessment are summarized in Table 4. It can be clearly seen that SVM significantly outperforms other approaches (*k*-NN and Decision Tree) and produces class-weighted F1 scores of over 0.82

in both cross-validation schemes. Confusion matrices for skill classification using SVM in all three CV schemes are also shown in Figure 5.

Regression results. Regression results are summarized in Table 5. We use four different kinds of regression methods, including support vector regression (SVR) with radial basis function kernel [41], Gaussian Processes (GPs) with a squared exponential kernel [36], Regression trees [7], and Linear regression [10]. It can be seen that GPs produced the best regression results with minimum errors both in terms of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) metrics.

Fig. 4. Activity classification results for (a) k-fold, (b) leave-one-user-out cross, and (c) leave-one-activity-out validation schemes.

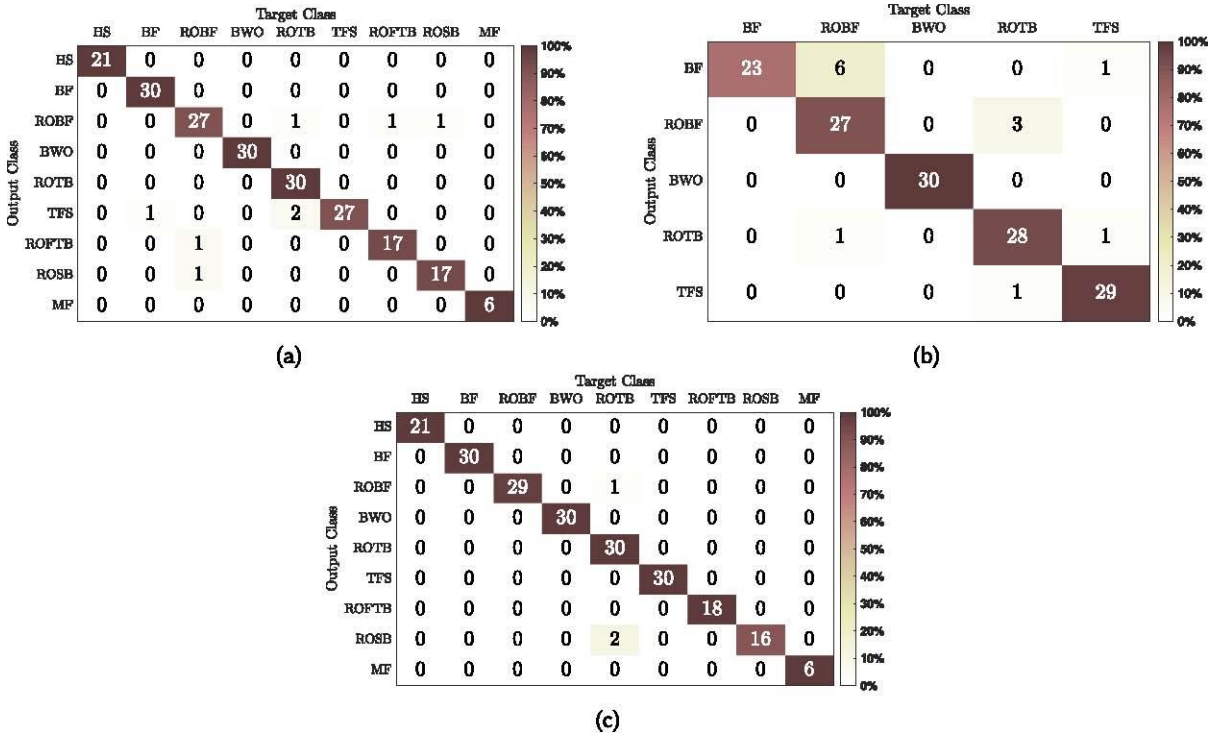


Table 4. Skill Classification Results Using Various Classification Strategies and Three Cross-Validation Methods: Type 1: k-fold, Type 2: Leave-One-User-Out, and Type 3: Leave-One-Activity-Out

	Validation Type 1			Validation Type 2			Validation Type 3		
	Prec.	Rec.	F1 Score	Prec.	Rec.	F1 Score	Prec.	Rec.	F1 Score
SVM	**0.8142**	**0.8125**	**0.8126**	**0.8991**	**0.8828**	**0.8828**	**0.8231**	**0.8173**	**0.8153**
Decision Tree	0.4075	0.4087	0.4078	0.5360	0.5360	0.5379	0.5696	0.5721	0.5705
k-NN	0.7145	0.7163	0.7151	0.5483	0.5554	0.5448	0.6620	0.6618	0.6615

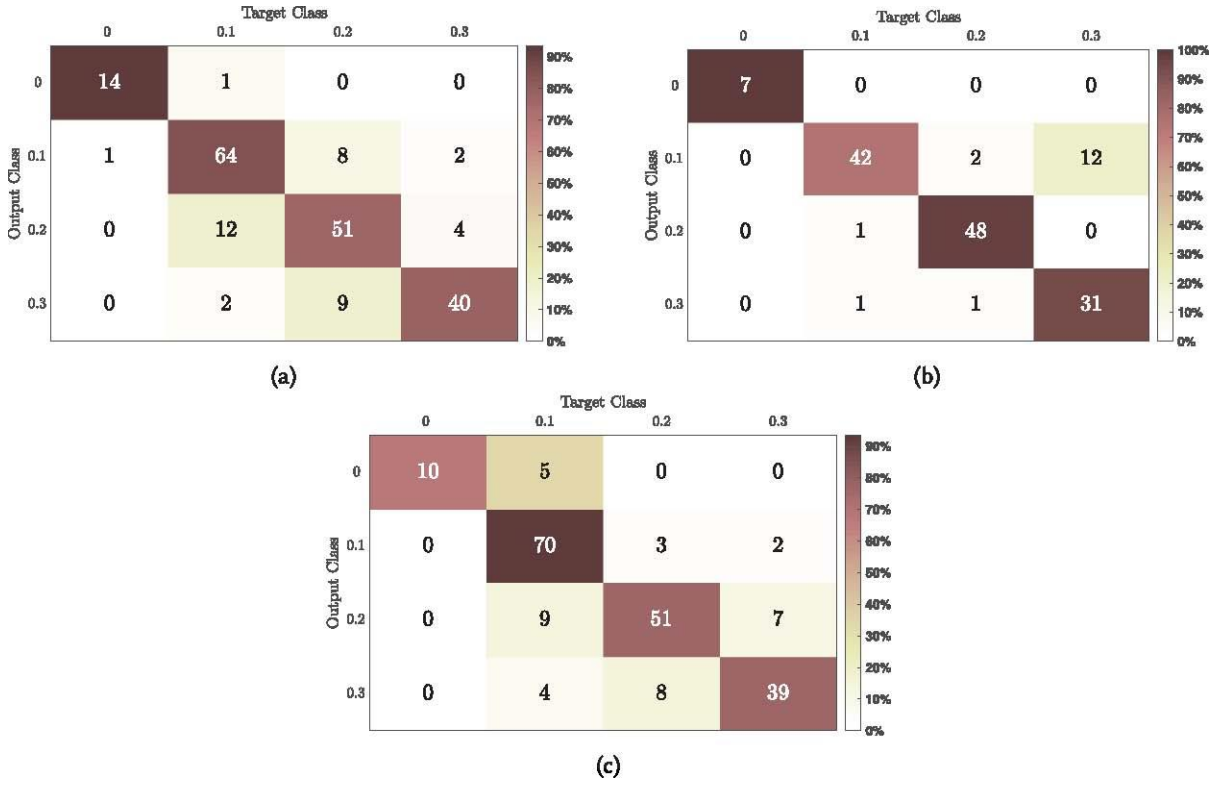


Fig. 5. Skill classification results for the three cross-validation methods used during the low-level activity classification; (a) k-fold, (b) leave-one-user-out, and (c) leave-one-activity-out cross-validation methods.

4.1.3 CPU Times. In this section, we show that the proposed approach can be deployed in (near) real-time for assessing skill. As soon as a session ends with multiple activities, results from the two main parts (low-level activity recognition and high-level skill classification) can be produced. In Figure 6, we show the computation time (in seconds) for these two stages. In all cases, the complete skill assessment procedure, from reading the raw data to producing skill assessment results, can be performed in less than half a second on average (adding computation times for both parts). This means that the proposed approach can be deployed for assessing skill (almost) immediately by providing skill feedback at the end of each session (or activity in this scenario). All experiments were performed using an Intel Core i7-6700K CPU at 4.00GHz×8 with 32GB of RAM.

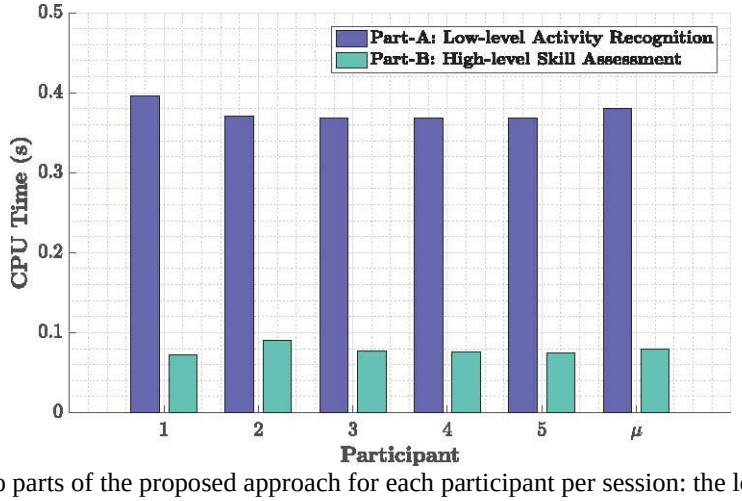


Fig. 6. CPU times for the two parts of the proposed approach for each participant per session: the low-level activity recognition part and the high-level skill assessment part.

Table 5. Skill Assessment Results Using Various Regression Strategies and Three Cross-Validation Methods as Used for Classification.

	Validation Type 1		Validation Type 2		Validation Type 3	
	RMSE	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
Support Vector Regression	0.0595	0.0422	0.0797	0.0635	0.0648	0.0465
Gaussian Processes	**0.0535**	**0.0395**	**0.0699**	**0.0541**	**0.0527**	**0.0402**
Regression Trees	0.0976	0.0720	0.0829	0.0617	0.0771	0.0505
Linear Regression	0.0901	0.0718	0.0882	0.0748	0.0583	0.0410

4.2 Assessment of Surgical Skills

Aiming to explore the generalization capabilities and for direct comparison to previous work, we conducted a second case study. Much previous work on skill assessment was dedicated to the analysis of surgical performances as they are routinely performed by medical students as part of their training. Various approaches, based either on the analysis of video data or on the assessment of tri-axial accelerometry, have been proposed that all focused on the automated generation of quantitative, objective judgments on the quality of certain surgical procedures as they would be provided by more senior and hence experienced surgeons (e.g., [22, 39]).

In our second case study we employed our novel skill assessment framework on the accelerometry dataset that was recorded for the aforementioned studies. Tri-axial accelerometers were attached to surgical equipment (forceps and needle holder), and movement data during surgical standard tasks was captured for a total of 15 participants and a total of 50 attempts of particular surgical procedures such as stitching or knot tying. Participants had varying levels of expertise and thus skill while attempting the same task of suturing a wound using a replica pad.

Table 6. Overview of Annotated Activities and Skills for n = 50 Attempts in the Surgical Skill Assessment Study

Activities		Skills					
Activity	Total Annotations	OSATS Measure	1	2	3	4	5
Using Forceps with Needle	1,706	Respect for Tissue	0	3	14	23	10
Using Needle-Holder	644	Time and Motion	7	10	17	13	3
Making a Stitch	957	Instrument Handling	5	13	16	10	6
Attempting a Knot	376	Suture Handling	9	9	18	12	2
Overall Procedure	56	Flow of Operation	1	8	18	15	8
		Knowledge of Procedure	2	15	8	16	9
		Overall Performance	3	12	15	17	3
$\sum$	3,739	$\sum$	27	70	106	106	41

More than 10 hours of sensor data were collected for this dataset. Ground-truth annotation was provided at two levels: low-level activity annotation and higher-level skill assessment. For the former, all 50 sessions were annotated by two independent, trained human coders who labeled the sessions with regards to the elementary activities relevant for surgical procedures (Table 6, left). For the second level of assessment, seven different skill measures based on the objective structured assessment of technical skill (OSATS) criteria [15] were annotated by an expert observing the sessions. These measures are labeled as (1) *Respect for Tissue*, (2) *Time and Motion*, (3) *Instrument Handling*, (4) *Suture Handling*, (5) *Flow of Operation*, (6) *Knowledge of Procedure*, and (7) *Overall Performance* (Table 6, right). This assessment is similar to the gymnastics scoring in that a human expert derives scores relating to aspects of the physical performance, although in this case the aspects remain separate and infer levels of additional knowledge. This scoring is suitable for critical assessment of the individual as in a competition or training.

Since generalized skill assessment has been performed previously using this dataset [22], the main aim here is to evaluate the two aspects of the proposed framework: accuracy and efficiency.

As detailed above, raw data was captured using accelerometers attached to the instruments being used for the surgical training. Although activity labels are available for this dataset, they are at a very abstract level and sporadic. Therefore, they do not provide a sufficient granularity level for skill assessment. It is for this reason that symbolic representations [3] were used as a first step (see Figure 1). The resulting symbols provide labels for segmented raw data. Features were then extracted for symbol classification (as explained in Section 3). Lowlevel classification is performed with regard to nine quantization symbols (for both

instruments) using SVMs (according to the results of our detailed analysis as discussed in the previous section). Posterior probabilities were computed for each modality.

For the high-level skill assessment, expert skill labels are used to train classification models. Skill labels are used in the same fashion as in the baseline approach; as such, the change in skill quality rather than the absolute skill is classified (i.e., there are three classes representing *deterioration*, *consistency*, and *improvement* in skill). We perform evaluations using the same cross-validation strategies as previously reported [22] (leave-one-attempt out cross-validation schemes) and make comparisons between the two approaches.

In Figure 7, we show skill assessment performance using the same classifier and a leave-one-attempt-out cross-validation scheme. It can be seen that in some cases, the proposed approach provides significantly better F1 scores (for skills such as instrument handling, suture handling, and knowledge of procedure). For other skill metrics, there is no significant difference in performance.

The skill assessment accuracy is achieved at substantially reduced computational costs and thus significantly reduced processing times. The baseline approach can be split into two main parts: (A) the hierarchical rule induction part and (B) the high-level skill classification part. The approach proposed in this article is also split into two main parts: (A) low-level activity/movement classification and (B) a similar high-level skill classification.

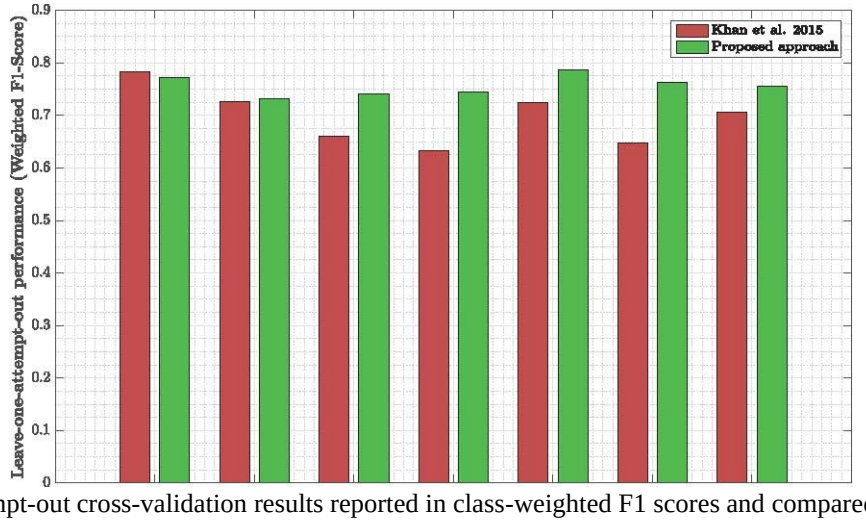
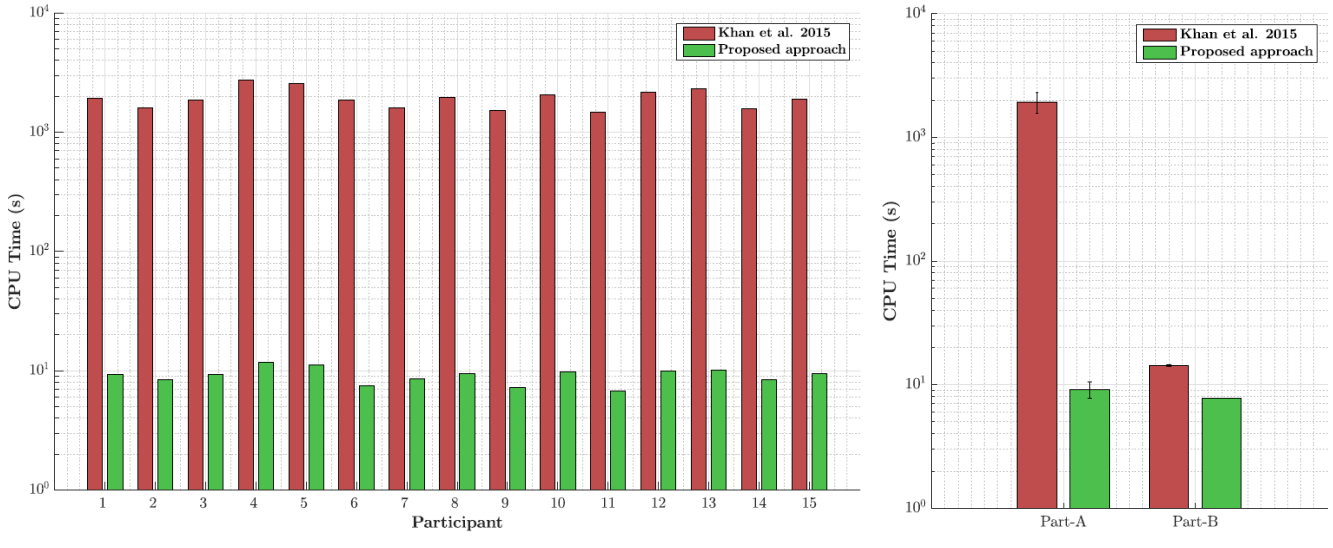


Fig. 7. Leave-one-attempt-out cross-validation results reported in class-weighted F1 scores and compared against the baseline p values using a two-sample t -test are given below



(a) CPU times for hierarchical rule induction-based approach (baseline) and low-level movement recognition.

(b) Mean CPU times for both low and high level compared against the baseline

Fig. 8. CPU time comparison.

CPU time comparisons are made in Figure 8(a) for part A and Figure 8(b) for part B. It can be seen that there is a substantial difference in CPU times (shown in seconds) for all of the participants. Note, these times correspond to the deployment/test, that is, inference stage. The baseline approach on average takes more than 1,000s, whereas the newly proposed approach for the same participants takes less than 10s on average (see Figure 8(a)). For part B, the difference is lower but still significant; 14.2785*s* +/- 0.1885 for the baseline approach and 7.7133*s* +/- 0.0154 for the proposed approach. However, part B in the baseline approach relies on the hierarchical rule induction of part A, which means that in domains where skill is even more complex and where more modalities are used, the total number of features would also increase, resulting in more computation time for part B.

5 CONCLUSIONS, LIMITATONS AND FUTURE WORK

Skill (or quality) assessment represents an extension of HAR. The focus is to go beyond assessing what is happening when, by focusing on how well certain activities are performed. Prominent applications for skill assessment are automated coaching in skill acquisition and training for certain professions, and in sports. Few first systems for skill assessment based on body-worn sensors have been introduced to the wider ubiquitous computing community. The majority of previous work focuses on methods that are tailored towards specific domains such as coaching in climbing, tennis, or cricket. Beyond such specialized systems, there is little work related to generalized skill assessment that is transferrable between application domains without substantial (re-)modeling efforts. The work presented in this article addresses this general direction of research.

We developed a new approach for skill assessment that is automated, generalized, and at the same time accelerated compared to previously proposed techniques. The new approach does not rely on a set of parameters that explicitly define skill. Since the definition of skill could be very specific to individual domains, a generic definition of skill is essential for generalized approaches, yet those definitions used in previous work are very complex, thus limiting applicability, especially when (near) real-time assessments are desired.

Our definition of skill is reduced to the notion of repeatability and consistency; that is, skill manifests itself in higher similarity of repeated activities. We employ the repeatability aspect of skill encoded in the confidence scores with which activities are classified. We showed that with such a simpler definition of skill, the resulting framework becomes less complex and thus better scalable: feature dimensionality remains controlled when adding sensors, and session length is not a limiting factor.

We demonstrated the effectiveness of the proposed approach in two case studies: (1) automated quality scoring in gymnastics and (2) surgical skill assessment. For the former we recorded data from gymnasts who wore a number of IMUs while performing a range of exercises. They were scored by professional judges and the task was to reproduce these manual scorings through our automated method. The second case study was based on an existing dataset of medical students practicing certain surgical procedures, thereby using sensor-equipped utensils. The latter dataset has been used in previous work and mainly served for comparison to the state of the art. The achieved results show great promise in terms of absolute assessment accuracy, generalizability across domains with minimal (mainly logistical) effort, and substantially improved efficiency. With our skill assessment system it is possible to automatically provide near-real-time feedback of high quality.

There may be certain applications where this notion of repeatability and consistency may not apply, such as in applications where activities are not repeatable in nature. In such a context, the framework of [22] may be used that, although not optimal for real-time deployment, provides an extensive approach to generalized skill assessment that is not limited by the repeatability aspect of activities.

In the proposed and similar other frameworks, skill assessment is restricted to domains where activity data is available; for example, in applications where skill is assessed based on other factors such as mental aptitude, other modalities may be used in order to assess skill effectively. One example of this could be in affective computing, where skill may be assessed via stress analysis. The definition of skill is relatively generic in the proposed and similar frameworks, but they are usually applied in a domain-specific manner. Transfer learning may be a candidate approach that can be used in the future in order to reduce training times when assessing skills in physically similar domains (tennis/badminton or cricket/baseball).

Further evaluation of the proposed framework may be performed by considering skill assessment in the context of a long-term deployment. Models would need to be accordingly trained in order to further test the generalization performance over such a deployment. Such a study would also present its own unique challenges in handling variations in sensor placement, clothing, and equipment setup.

REFERENCES

[1] Ludovic Baudry, Ludovic Seifert, and David Leroy. 2008. Spatial consistency of circle on the pedagogic pommel horse: Influence of expertise. *Journal of Strength and Conditioning Research* 22, 2 (2008), 608–613.

[2] Karen T. Beatty, Andrew S. Mcintosh, and Bertrand O. Frechede. 2006. Methodfor analysing the risk of overuse injury in gymnastics. In *Proc. Int. Symp. on Biomechanics in Sports*. 1–4.

[3] Eugen Berlin and Kristof Van Laerhoven. 2012. Detecting leisure activities with dense motif discovery. In *Proc. ACM Int. Joint Conf. Ubiquitous and Pervasive Comp. (UbiComp'12)*. ACM, New York, NY, 250–259. DOI: https://doi.org/10.1145/2370216.2370257

[4] Peter Blank, Benjamin H. Groh, and Bjoern M. Eskofier. 2017. Ball speed and spin estimation in table tennis using a racket-mounted inertial sensor. In *Proc. Int. Symp. Wearable Computing (ISWC'17)*. ACM, New York, NY. DOI: http://doi.acm.org/10.1145/3123021.3123040

[5] Aaron F. Bobick. 1997. Movement, activity and action: The role of knowledge in the perception of motion. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences* 352, 1358 (Aug.1997), 1257–1265. DOI: https://doi.org/10.1098/rstb.1997.0108

[6] Elizabeth J. Bradshaw and Patria A. Hume. 2012. Biomechanical approaches to identify and quantify injury mechanisms and risk factors in women's artistic gymnastics. *Sports Biomechanics* 11, 3 (2012), 324–341. DOI: https://doi.org/10.1080/14763141.2011.650186

[7] Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. 1984. *Classification and Regression Trees*. Chapman & Hall, New York, NY. 358 pages. http://www.crcpress.com/catalog/C4841.htm

[8] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *Computing Surveys* 46, 3 (2014), 1–33. DOI: https://doi.org/10.1145/2499621

[9] Lauren Burt, Geraldine Naughton, and Raul Landeo. 2007.Quantifyingimpactsduringbeam and floor training in pre-adolescent girls from two streams of artistic gymnastics. In *Proc. Int. Symp. on Biomechanics in Sports*. 354–357.

[10] Samprit Chatterjee and Ali S. Hadi. 2006. *Regression Analysis by Example*. John Wiley & Sons, Inc. DOI: https://doi.org/10.1002/0470055464

[11] Liming Chen, J. Hoey, C.D. Nugent, D. J. Cook, and Zhiwen Yu. 2012.Sensor-based activity recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 6 (Nov. 2012), 790–808. DOI: https://doi.org/10.1109/TSMCC.2012.2198883

[12] John W. Chow and Duane V. Knudson. 2011. Use of deterministic models in sports and exercise biomechanics research. *Sports Biomechanics* 10, 3 (Sept. 2011), 219–233. DOI: https://doi.org/10.1080/14763141.2011.592212

[13] Ruth H. Da-Silva, Frederike van Wijck, Lisa Shaw, Helen Rodgers, Madeline Balaam, Lianne Brkic, Thomas Ploetz, Dan Jackson, Karim Ladha, and Christopher I. Price. 2018. Prompting arm activity after stroke: A clinical proof of concept study of wrist-worn accelerometers with a vibrating alert function. *Journal of Rehabilitation and Assistive Technologies Engineering* 5 (May 2018), 1–8. DOI: https://doi.org/10.1177/2055668318761524

[14] Ruth H. Da-Silva, Frederike van Wijck, Lisa Shaw, Helen Rodgers, Madeline Balaam, Lianne Brkic, Thomas Ploetz, Dan Jackson, Karim Ladha, and Christopher I. Price. 2018. Prompting arm activity after stroke: A clinical proof of concept study of wrist-worn accelerometers with a vibrating alert function. *Journal of Rehabilitation and Assistive Technologies Engineering 5* (2018), 2055668318761524. DOI: https://doi.org/10.1177/2055668318761524

[15] Vivek Datta, Simon Bann, Mirren Mandalia, and Ara Darzi. 2006. The surgical efficiency score: A feasible, reliable, and valid method of skills assessment. *American Journal of Surgery* 192, 3 (Sept. 2006), 372–378. DOI: https://doi.org/10.1016/j.amjsurg.2006.06.001

[16] Davide G. de Sousa, Lisa A. Harvey, Simone Dorsch, and Joanne V. Glinsky. 2018. Interventions involving repetitive practice improve strength after stroke: A systematic review. *Journal of Physiotherapy* 64, 4 (2018), 210–221. DOI: https://doi.org/10.1016/j.jphys.2018.08.004

[17] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. 2017. Who's better? Who's best? Pairwise deep ranking for skill determination. In *Proc. Conf. Computer Vision and Pattern Recognition (CVPR'17)*.

[18] Benjamin H. Groh, Martin Fleckenstein, Thomas Kautz, and Björn Eskofier. 2017. Classification and visualization of skateboard tricks using wearable sensors. *Pervasive and Mobile Computing (PMC)* 40, C (Sept.2017), 42–55. DOI: https://doi.org/10.1016/j.pmcj.2017.05.007

[19] Nils Hammerla, Reuben Kirkham, Peter Andras, and Thomas Plötz. 2013. On preserving statistical characteristics of accelerometry data using their empirical cumulative distribution. In *Proc. Int. Symp. Wearable Computing (ISWC).* ACM, New York, NY, USA, 65–68. DOI: https://doi.org/10.1145/2493988.2494353

[20] Nils Y. Hammerla and Thomas Plötz. 2015. Let's (not) stick together: Pairwise similarity biases cross-validation in activity recognition. In *Proc. Ubicomp*.

[21] Jesse Hoey, Thomas Plötz, Dan Jackson, Andrew Monk, Cuong Pham, and Patrick Olivier. 2011. Rapid specification and automated generation of prompting systems to assist people with dementia. *Pervasive and Mobile Computing (PMC)* 7, 3 (June 2011), 299–318. DOI: https://doi.org/10.1016/j.pmcj.2010.11.007

[22] Aftab Khan, Sebastian Mellor, Eugen Berlin, Robin Thompson, Roisin McNaney, PatrickOlivier, and Thomas Plötz. 2015. Beyond activity recognition: Skill assessment from accelerometer data. In *Proc. Int. Joint Conf. Ubiquitous and Pervasive Comp. (UbiComp'15)*. ACM, 1155–1166. DOI: https://doi.org/10.1145/2750858.2807534

[23] Aftab Khan, James Nicholson, and Thomas Plötz. 2017. Activity recognition for quality assessment of batting shots in cricket using a hierarchical representation. In *Proc. Interactive, Mobile, Wearable and Ubiquitous Computing (IMWUT)* 1, 3 (Sept. 2017), 62:1–62:31. DOI: https://doi.org/10.1145/3130927

[24] Aftab Khan, David Windridge, and Josef Kittler. 2014. Multi level Chinese takeaway process and label-based processes for rule induction in the context of automated sports video annotation. *IEEE Transactions on Cybernetics* 44, 10 (Oct. 2014), 1910–1923. DOI: https://doi.org/10.1109/tcyb.2014.2299955

[25] Hyeokhyen Kwon, Gregory D. Abowd, and Thomas Plötz. 2018. Adding structural characteristics to distribution-based accelerometer representations for activity recognition using wearables. In *Proc. Int. Symp. Wearable Computing (ISWC'18)*. 72–75.

[26] Cassim Ladha, Nils Y. Hammerla, Patrick Olivier, and Thomas Plötz. 2013. ClimbAX: Skill assessment for climbing enthusiasts. In *Proc. ACM Int. Joint Conf. Ubiquitous and Pervasive Comp. (UbiComp'13)*. ACM, 235–244. DOI: https://doi.org/10.1145/2493432.2493492

[27] Claudine J. C. Lamoth, Rob C. van Lummel, and Peter J. Beek. 2009. Athletic skill level is reflected in body sway: A test case for accelometry in combination with stochastic dynamics. *Gait & Posture* 29, 4 (June 2009), 546–551. DOI: https://doi.org/10.1016/j.gaitpost. 2008.12.006

[28] Oscar D. Lara and Miguel A. Labrador. 2013. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys & Tutorials* 15, 3 (2013), 1192–1209. DOI: https://doi.org/10.1109/surv.2012.110112.00192

[29] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. 2007. Experiencing SAX: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery* 15, 2 (April 2007), 107–144. DOI: https://doi.org/10.1007/s10618-007-0064-z

[30] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. 2006. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding* 104, 2–3 (Nov. 2006), 90–126. DOI: https://doi.org/10.1016/j.cviu.2006.08.002

[31] Andreas Möller, Luis Roalter, Stefan Diewald, Matthias Kranz, Nils Hammerla, Patrick Olivier, and Thomas Plötz. 2012. GymSkill: A personal trainer for physical exercises. In *Proc. IEEE Conf. Pervasive Comp. and Communication (PerCom'12)*. 213–220. DOI: https://doi.org/10.1109/PerCom.2012.6199869

[32] Paritosh Parmar and Brendan Tran Morris. 2017. Learning to score olympic events. In *Proc. Conf. Computer Vision and Pattern Recognition (CVPR'17)*.

[33] John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances In Large Margin Classifiers*. MIT Press, 61–74.

[34] Thomas Plötz, Chen Chenn, Nils Y. Hammerla, and Gregory D.Abowd. 2012. Automatic synchronization of wearable sensors and video cameras for ground truth annotation –A practical approach. In *Proc. Int. Symp. Wearable Computing (ISWC'12)*.100–103. DOI: https://doi.org/10.1109/ISWC.2012.15

[35] Thomas Plötz, Nils Y. Hammerla, and Patrick Olivier. 2011. Feature learning for activity recognition in ubiquitous computing. In *Proc. Int. Joint Conf. Artificial Intelligence (IJCAI'11)*. AAAI Press, 1729–1734. DOI: https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-290

[36] Carl Edward Rasmussen and Christopher K. I. Williams. 2006. In *Gaussian Processes for Machine Learning*. MIT Press.

[37] Martin Seiffert, Flavio Holstein, Rainer Schlosser, and Jochen Schiller. 2017. Next generation cooperative wearables: Generalized activity assessment computed fully distributed within a wireless body area network. *IEEE Access* 5 (Sept. 2017), 16793–16807. DOI: https://doi.org/10.1109/access.2017.2749005

[38] Yachna Sharma, Vinay Bettadapura, Thomas Plötz, Nils Hammerla, Sebastian Mellor, Roisin McNaney, Patrick Olivier, Sandeep Deshmukh, Andrew Mccaskie, and Irfan Essa. 2014. Video based assessment of OSATS using sequential motion textures. In *Proc. 5th Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI'14)*.

[39] Yachna Sharma, Thomas Plötz, Nils Hammerla, Sebastian Mellor, Roisin McNaney, Patrick Olivier, Sandeep Deshmukh, Andrew Mccaskie, and Irfan Essa. 2014. Automated surgical OSATS prediction from videos. In *Proc. IEEE Int. Symposium on Biomedical Imaging (ISBI'14)*. DOI: https://doi.org/10.1109/isbi.2014.6867908

[40] Chantal Simons and Elizabeth J. Bradshaw. 2016. Reliability of accelerometry to assess impact loads of jumping and landing tasks. *Sports Biomechanics* 15, 1 (Jan. 2016), 1–10. DOI: https://doi.org/10.1080/14763141.2015.1091032

[41] Alex J. Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Compu*ting 14, 3 (Aug. 2004), 199–222. DOI: https://doi.org/10.1023/B:STCO.0000035301.49549.88

[42] Robin Thompson, Ilias Kyriazakis, Amey Holden, Patrick Olivier, and Thomas Plötz. 2015. Dancing with horses: Automated quality feedback for dressage riders. In *Proc. ACM Int. Joint Conf. Ubiquitous and Pervasive Comp. (UbiComp'15)*. ACM, 325–336. DOI: https://doi.org/10.1145/2750858.2807536

[43] Eduardo Velloso, Andreas Bulling, Hans Gellersen, Wallace Ugulino, and Hugo Fuks. 2013. Qualitative activity recognition of weight lifting exercises. In *Proc. Int. Conf. Augmented Human*. ACM, New York, NY, 116–123. DOI: https://doi.org/10.1145/2459236.2459256

[44] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proc. Int. Conf. Language Resources and Evaluation Conference (LREC'06)*.

[45] Bob Woolmer, Timothy Noakes, Helen Moffett, and Fay Lewis. 2008. *Bob Woolmer's Art and Science of Cricket*. New Holland London.

[46] Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. 2004. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research* 5 (Dec. 2004), 975–1005. http://dl.acm.org/citation.cfm?id=1005332.1016791

[47] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L. Sarin, Thomas Plötz, Mark A. Clements, and Irfan Essa. 2016. Automated video based assessment of surgical skills for training and evaluation in medical schools. *International Journal of Computer Assisted Radiology and Surgery* 11, 9 (Aug. 2016), 1623–1636. DOI: https://doi.org/10.1007/s11548-016-1468-2