Learning to assess the quality of stroke rehabilitation exercises

Conference Paper · March 2019
DOI: 10.1145/3301275.3302273

5 authors, including:

Sergi Bermúdez i Badia
Universidade da Madeira
152 PUBLICATIONS 1,608 CITATIONS

Some of the authors of this publication are also working on these related projects:

- MACBIOIDI: Promoting the cohesion of Macaronesian regions through a common ICT platform for biomedical R & D & I” (INTERREG program MAC/1.1.b/098) View project
- 3Dreams View project
Learning to Assess the Quality of Stroke Rehabilitation Exercises

Min Hun Lee
Daniel P. Siewiorek
Asim Smailagic
Carnegie Mellon University
mhlee@cmu.edu
dps.asimj@cs.cmu.edu

Alexandre Bernadino
Instituto Superior Técnico
alex@isr.tecnico.ulisboa.pt

Sergi Bermúdez i Badia
Madeira Interactive Technology Institute
sergi.bermudez@m-iti.org

ABSTRACT

Due to the limited number of therapists, task-oriented exercises are often prescribed for post-stroke survivors as in-home rehabilitation. During in-home rehabilitation, a patient may become unmotivated or confused to comply prescriptions without the feedback of a therapist. To address this challenge, this paper proposes an automated method that can achieve not only qualitative, but also quantitative assessment of stroke rehabilitation exercises. Specifically, we explored a threshold model that utilizes the outputs of binary classifiers to quantify the correctness of a movements into a performance score. We collected movements of 11 healthy subjects and 15 post-stroke survivors using a Kinect sensor and ground truth scores from primary and secondary therapists. The proposed method achieves the following agreement with the primary therapist: 0.8436, 0.8264, and 0.7976 F1-scores on three task-oriented exercises. Experimental results show that our approach performs equally well or better than multi-class classification, regression, or the evaluation of the secondary therapist. Furthermore, we found a strong correlation ($R^2 = 0.95$) between the sum of computed exercise scores and the Fugl-Meyer Assessment scores, clinically validated motor impairment index of post-stroke survivors. Our results demonstrate a feasibility of automatically assessing stroke rehabilitation exercises with the decent agreement levels and clinical relevance.

CCS CONCEPTS
• Computing methodologies → Intelligent agents; Activity recognition and understanding; • Applied computing → Health care information systems;

KEYWORDS
Intelligent Agent, Motion Analysis, Stroke Rehabilitation

1 INTRODUCTION

One of the effective stroke rehabilitation approaches is physical therapy intervention with task-oriented exercises [21, 39, 47]. During a therapy session, a therapist monitors patient’s performance and guides a patient with feedback. However, a post-stroke survivor may not receive timely and comprehensive rehabilitation due to the limited availability of therapists [36]. Alternatively, a therapist often prescribes in-home rehabilitation regimens, in which a post-stroke survivor independently participates in rehabilitation without any supervision of a therapist. A patient might feel uncertain whether he/she correctly performs an exercise and become unmotivated in rehabilitation regimens [2, 26]. A therapist encounters a challenge of tracking the progress of a patient and making an informative adjustment on rehabilitation regimens [15]. Thus, it becomes imperative to have a method to monitor and quantitatively assess in-home rehabilitation exercises.

Recent advances in sensors and machine learning algorithms offer a potential of computer-based in-home rehabilitation [20, 42]. Researchers have demonstrated systems that log measurements of movements (e.g. joint angles [4]), count repetitions of exercises, and detect incorrect movements [16, 19, 30] to support in-home rehabilitation. However, logging motion-related measurements and binary detection of incorrect movements are not intuitive and sufficient to follow the progress of a post-stroke survivor’s in-home rehabilitation. A therapist mentioned it is difficult to interpret sensor measurements (e.g. velocity of a joint) [16]. In addition, reporting the results of binary detection on an incorrect movement is limited for both therapists and patients to track diverse degrees of functional abilities. Instead of these approaches, prior studies on in-home rehabilitation exercise tools with therapists and patients [7, 16] highlighted the importance and need of having quantitative...
2 BACKGROUND AND RELATED WORK

2.1 Challenges of Stroke Rehabilitation Practices

Post-stroke survivors should receive an early and extensive rehabilitation program to prevent disability and stroke recurrence. Promoting task-oriented exercise is a popular strategy to improve functional ability and lower a chance of having recurrent stroke [39]. However, it is expensive and difficult for post-stroke survivors to receive the administration of an individualized rehabilitation session [36]. Instead, a post-stroke survivor engages in in-home rehabilitation without any supervision of a therapist.

Both stroke survivors and therapists encounter following challenges to pursue in-home rehabilitation. First, stroke survivors may have low participation in rehabilitation due to several reasons [2, 9]. Most of them expressed anxieties about a lack of information and support from professionals [2, 26]. They described the need of a trainer, who provides motivation and coaching on their performance [9]. Specifically, they mentioned that a way to see physical improvement after exercising would be a good source of exercise motivation [9]. According to the studies of [25, 38], they found that viewing their scores on a screen made them motivated to beat their previous scores. In addition, as patient’s self-report is a primary source of a therapist to follow the adherence and progress of a patient, a therapist has limited quantitative performance data to understand patient’s progress [15]. A therapist has difficulty with adjusting in-home rehabilitation regimens and deriving the general predictability of post-stroke motor recovery [15]. As a first step to address these challenges, this paper mainly explores the feasibility of developing a computer-assisted method to assess stroke rehabilitation exercises.

2.2 Computer-Assisted Rehabilitation Monitoring

One preliminary approach to monitor in-home rehabilitation is logging the measurements of a sensor. Huang utilized inertial sensors to measure joint movements (e.g. repetitions, rotation velocity, frequency, range of motion) for balance rehabilitation [16]. However, according to the user study of deploying a monitoring system [16], physical therapists mentioned that basic measurements (e.g. degrees/sec) were not intuitive to assess a patient’s compliance to prescription. Physical therapists preferred to have easily comprehensible data, so that they could spend more time with patients [16].

Another approach is to apply gesture recognition [10, 24] that classifies incorrect movements. Chang et al. utilized the rules of six joint angles to identify the accuracy of movements for upper limb rehabilitation [6]. Pogorele et al. utilized k-nearest neighbors and neural network algorithms to recognize four gait related problems with body-worn tags and wall-mounted sensors [35]. Das et al. applied a Support Vector Machine (SVM) classifier to distinguish mild and severe symptoms of Parkinson’s Disease using full body motion capture data from four Parkinson’s patients [10]. Su et al. computed joint positions and the speed of completing an exercise and applied neural networks and fuzzy logics to classify the quality

performance results. Performance results can reinforce patient’s motivation [9, 25, 38] and patient’s adherence to rehabilitation regimens [7]. In addition, they are valuable for therapists to understand patient’s performance and adjust regimens.

In this paper, we describe an approach of automatically assessing stroke rehabilitation exercises using a RGB depth sensor and machine learning algorithms. Our approach leverages a threshold model with binary classifications [14], providing the ability to assess performance qualitatively and quantitatively (Figure 1). Our insight is that assessing the quality of stroke rehabilitation exercises has the property of both classification and regression with categorical scales (e.g. ‘0: cannot perform’, ‘1: partially perform’, ‘2: fully perform’ [37]). The state-of-the-art approaches to classify categorical scales assume that the categorical response is a measured latent continuous variable, which can be modeled with intervals on the real line [14, 22]. This assumption allows to leaning a set of thresholds to divide data into categorical responses and quantifying the performance of a movement using a threshold model with binary classifications.

To demonstrate our proposed method, we recruited two therapists to specify the experimental designs (i.e. three performance components of task-oriented rehabilitation exercises, three upper-limb exercises). We then collected a dataset of three upper-limb exercises, which includes 900 motions from 15 post-stroke survivors, ground truth scores from primary and secondary therapists, and motor impaired scores of 15 post-stroke survivors with Fugl Meyer Assessment (FMA).

Using this dataset, we evaluate our approach of assessing the quality of a movement. First, we show that a threshold model with binary classifications performs better or equally good with multi-class classification or regression approaches using various machine learning algorithms (i.e. Decision Tree, Linear Regression, Support Vector Machine, Neural Network, Long Short Term Memory Network). It promises more scalable development of an automated assessment than multi-class or regression approaches that require expensive data collection from therapists and post-stroke survivors.

Second, our approach can achieve decent agreement levels with the primary therapist (0.82, 0.80, and 0.75 F1-scores on three exercises), which is equally good or better than the secondary therapist. In addition, the computed scores of our approach has a strong correlation with FMA scores. Our approach can consistently assess clinically relevant, quantitative performance scores without repetitively requiring multiple hours of discussion between therapists.

The contributions of this paper are as follows:

- We introduce an automated assessment method that utilizes a threshold model with binary classifications to qualitatively and quantitatively assess task-oriented rehabilitation exercises.
- We present the experimental results from 15 post-stroke survivors and two therapists on three upper-limb exercises across three performance components, which validate the effectiveness of our approach and clinical relevance with Fugl Meyer Assessment (FMA).
of an exercise movement into three levels (e.g. Bad, Good, Excellent) [45].

This gesture-recognition approach provides the number of correct movements to motivate a patient’s participation and follow a patient’s progress. However, it has limitation to represent diverse levels of patient’s progress. We cannot differentiate the performance of two patients who have the equivalent number of incorrect movements but with different degrees of incorrectness. Moreover, it would continuously indicate the incorrectness until patient’s full recovery. It may de-motivate a patient at some point. Thus, this paper mainly focuses on exploring the feasibility of quantitatively assessing patient’s exercise performance.

Several research works have shown the usefulness of kinematic variables to represent an objective assessment of upper limb motor performance. Using acted-out dataset from healthy subjects, Zhao et al. described a potential benefit of joint angles and hand positions to evaluate patient’s recovery [49]. Ozturk et al. demonstrated that the feasibility of using the speed and joint angle measurements to differentiate three stroke patients from two healthy subjects [29]. Murphy et al. identified that the measurement of compensatory trunk and arm movements can be utilized to discriminate moderate and mild arm impairment using ‘reaching and drinking’ exercise [27]. Patterson et al. demonstrated that kinematic variables (e.g. peak velocity, trunk displacement, etc) may be feasible and useful to measure functional ability [32]. However, limited works address how these kinematic variables can be exploited to quantitatively assess post-stroke survivor’s exercise performance and functional ability scores.

This leads to the following research questions:

- RQ1. How can we develop a system that automatically assesses exercise performance using kinematic variables?
- RQ2. How closely do computed scores of an automated method align with therapist’s assessment?
- RQ3. Do computed scores of an automated method have any clinical relevance?

3 EXPERIMENTAL DESIGNS OF TECHNOLOGY PROB FOR ASSESSING EXERCISES

The goal of this paper is not to show comprehensive functionalities of a computer-assisted rehabilitation system. Instead, this paper focuses on presenting a method of assessing the quality of stroke rehabilitation to collect patient’s performance data and further demonstrating how well it can assess compare to therapists as a technology probe [17, 18].

A therapist utilizes observation-based tests to assess the motor ability of a post-stroke survivor. We analyzed these existing functional assessment tests for stroke rehabilitation to identify therapist’s assessment strategies. After having iterative discussion with therapists, we specified our experimental designs to demonstrate a feasibility of quantitatively assessing stroke rehabilitation exercises. In the following subsections, we describe functional assessment tools for stroke rehabilitation and our experimental designs (i.e. performance components, therapist’s scoring guidelines, three upper limb exercises).

3.1 Functional Assessment Tests of Stroke Rehabilitation

The Fugl Meyer Assessment (FMA) [41] and the Wolf Motor Function Test (WMFT) [46] are frequently used to determine the motor ability of adult post-stroke survivors aged over 18 years old. The FMA examines the functional use of both upper and lower extremities through monitoring selected movement patterns. For each pattern, a therapist assigns the quality of movement on a 3-point ordinal scale (0 - 2): 0 = ‘cannot preform’, 1 = ‘partially perform’, and 2 = ‘fully performs’. The evaluation of upper extremity includes 33 tasks with the maximum of 66 points. The WMFT requires a therapist to measure time and assign a Functional Ability (FA) score to each of the 17 functional tasks on a 6-point ordinal scale (0-5). The scoring guidelines of the WMFT describe a single, combined statement that suggests considering the fluidity, precision of movements, and the existence of compensatory movements (e.g. the extent to which the head and trunk are maintained in normal alignment).

These clinical tests include a series of functional movements, where a therapist observes a targeted joint motion and assigns a numerical score to each functional movement using the scoring guidelines of a selected performance test. The summation of performance scores from functional movements represents post-stroke survivor’s functional ability.

3.2 Performance Components of Task-Oriented Stroke Rehabilitation Exercises

We identified commonly used factors to assess a functional ability of stroke patients from popular stroke rehabilitation assessment tools (i.e. the FMA and the WMFT) and related work on computer-assisted stroke rehabilitation monitoring systems. After the discussion with therapists, we specify three performance components to provide more detailed assessment and feedback instead of having a single overall score.

Table 1 summarizes the descriptions of three performance components: ‘Range of Motion (ROM)’, ‘Smoothness’, and ‘Compensation’. This abstracted categorization of three performance components can be represented with various kinematic factors of a motion for stroke rehabilitation. We describe the definition of three performance components along with the citations of related works, whose monitoring features can belong to our specified performance components, and therapist’s scoring guidelines.

One primary performance component is to evaluate whether a post-stroke survivor can achieve a particular motion pattern or task. The ‘ROM’ component represents the amount of an active movement with a specific joint. The ‘ROM’ component can be represented by joint angles [6, 16, 29, 41, 49, 50] or joint trajectory positions [5, 45, 46, 49].

Another performance component is to check the existence of jerky motion patterns or compensated movements. The ‘Smoothness’ component represents the degree of trembling and irregular motion movements of joints. The ‘Smoothness’ component can be indicated by velocity related features [5, 32, 40, 41, 46].

The ‘Compensation’ component monitors if a post-stroke survivor performs compensatory movements to achieve the target...
Table 1: Three Performance Components and Therapist’s Scoring Guidelines of Stroke Rehabilitation Exercises

<table>
<thead>
<tr>
<th>Performance Components</th>
<th>Descriptions</th>
<th>Related Work</th>
<th>Binary Labels</th>
<th>Score</th>
<th>Therapist’s Guidelines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range of Motion (ROM)</td>
<td>The amount of movement around a specific joint</td>
<td>[5, 6, 16, 29, 41, 45, 46, 49, 50]</td>
<td>Incorrect</td>
<td>0</td>
<td>Does not or barely involve any movement</td>
</tr>
<tr>
<td></td>
<td>Smoothness</td>
<td>The degree of trembling movement</td>
<td>Correct</td>
<td>1</td>
<td>Attempt is made but limited to be normal</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Correct</td>
<td>2</td>
<td>Movement appears to be normal</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Incorrect</td>
<td>0</td>
<td>Excessive tremor or not smooth coordination</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Incorrect</td>
<td>0</td>
<td>Movement influenced by tremor</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Correct</td>
<td>2</td>
<td>Smoothly coordinated movement</td>
</tr>
<tr>
<td>Compensation</td>
<td>The involvement of compensation to perform a motion</td>
<td>[5, 32]</td>
<td>Incorrect</td>
<td>0</td>
<td>Noticeable compensation in more than two joints</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Correct</td>
<td>1</td>
<td>Noticeable compensation in a joint</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Correct</td>
<td>2</td>
<td>Does not involve any compensations</td>
</tr>
</tbody>
</table>

Postures. Specifically, during upper limb movements, therapists focus on the occurrence of the following compensatory movements: leaning or swaying torso movements or elevated shoulder joints [5, 32, 46] as shown in Figure 2b.

Figure 2: (a) Motion of unaffected side (b) Compensated shoulder and trunk joints due to limited motion on a wrist

3.3 Three Upper Limb Exercises

This paper utilizes three upper limb stroke rehabilitation exercises (Figure 3) to validate the proposed approach. These exercises represent examples of task-oriented movements for stroke rehabilitation. In the Figure 3, the ‘Initial’ indicates the initial position of an exercise and the ‘Target’ describes the desired position of an exercise. Exercise 1 is ‘Bring a Cup to the Mouth’, in which a subject has to hold a cup and raise it to the mouth as if drinking water. Exercise 2 is ‘Switch a Light On’, where a patient pretends touching a light switch on the wall with shoulder forward flexion movement. Exercise 3 is ‘Move a Cane Forward’, which aims to practicing the usage of a cane while performing elbow extension at the seated position.

Figure 3: (a) Exercise 1 (b) Exercise 2 (c) Exercise 3

A therapist prescribes to repeat few exercises with major muscle groups during in-home rehabilitation [2, 33]. Based on the discussion with therapists, three exercise are selected due to their correspondence with major motion patterns: Exercise 1 to elbow flexion, Exercise 2 to shoulder flexion, and Exercise 3 to elbow extension.

In addition, previous studies have applied similar movement patterns to evaluate the usefulness of kinematic variables: ‘reaching and drinking from a glass’ task [27], ‘reaching’ task [32], and ‘forward reaching’ movement [29]. These three exercises can be also mapped into movements in clinically validated assessment tests: the Fugl Meyer Assessment (FMA) and Wolf Motor Function Test (WMFT). Exercise 1 is related to the ‘Elbow Flexion’ movement of the FMA and the ‘Lift can’ task of the WMFT. Exercise 2 is similar to the ‘Shoulder Flexion’ movement of the FMA and the ‘Hand to a box’ task of the WMFT. Exercise 3 is relevant to the ‘Elbow Extension’ of the FMA and the ‘Extend Elbow’ task of the WMFT.

Note that the purpose of this study is not to replace the existing functional ability tests but to demonstrate the feasibility of quantifying exercise performance score. This paper focuses on exploring the applicability of the proposed method to multiple patients using three exercises.

4 LEARNING A MODEL TO ASSESS THE QUALITY OF A MOVEMENT

An assessment of in-home rehabilitation exercises is to predict a categorical performance score (e.g. ‘0: no movement perform’, ‘1: limited movement’, ‘2: normal movement’), which can be considered as an intermediate problem between classification and regression. We can cast all categorical labels into real values and apply standard regression or multi-class classification techniques. Most popular approaches are called as threshold models [14, 22], which assume that categorical responses has a latent variable on the real line. They aim to learning a function, \( f(x) \) that predicts the values of a latent variable [14] and estimating a categorical response with a set of threshold in the range of \( \text{f(x)} \).

Prior work showed that well-tuned binary classification approaches can be transformed into good ranking algorithms [12, 22]. The confidence of a binary classifier can be considered as an ordering preference [13, 23]. Furthermore, Li and Lin showed theoretical and empirical analysis that this problem can be transformed into binary classifications [22]. Inspired by prior work, we utilize a threshold model with binary classifiers for assessing the quality of a movement.
This section describes how we extract kinematic features and which ability, a joint trajectory involves a number of valley shapes. His/her average speed or jerk with limited functional ability is expected to be lower than health subjects. A zero-crossing ratio indicates the period during which the speed changes from positive to negative or vice versa. As we assess stroke rehabilitation exercise in terms of five frames to reduce noise of acquiring joint positions from a Kinect similar to [44].

We process and normalize joint positions to reduce individual physical variabilities. The list of preprocessed and normalized features are described with their equations in Table 2. In addition to the common notations, we use a superscript to specify the statistics (i.e. $\max$ for the maximum and $\avg$ for the average). $f_t$ denotes a type of feature (i.e. $sp$ for speed, $ac$ for acceleration, and $jk$ for jerk as defined in Table 2). $I$ is an indicator function.

The joint angle ($j_a(t)$) feature computes an angle among three joints. The relative trajectory ($r_t$) computes how far a selected joint is moved away from the basis joint using the Euclidean distance. We specify a head joint as the basis joint and elbow and wrist joints as selected joints for upper limb exercises. The projected trajectory ($p_t$) describes the absolute distance between one selected joint and the other selected joint in a specific coordinate. Using the relative trajectory features, we also compute the following quantities of a motion: speed ($sp_t$), acceleration ($ac_t$), and jerk ($jk_t$).

In addition, we define several normalized features to compensate individual’s physical variability. The normalized relative trajectory ($nrt_t$) and normalized projected trajectory ($npt_t$) describe the change of a trajectory feature from an initial position. These normalized features can be considered as the normalization with subject’s physical conditions, because an initial position is dependent on subject’s physical characteristics (e.g. the length of limbs). It can utilize to segment a starting and ending frames of an exercise.

To normalize the speed related features, we utilize an average and maximum of speed or jerk until a selected frame. Normalized speed or jerk ($nspt$ or $njk_t$) is the division of an average speed or jerk by the maximum speed or jerk value. If a subject has limited functional ability, a joint trajectory involves a number of valley shapes. His/her average speed or jerk with limited functional ability is expected to be smaller than that of health subjects. He/she is expected to have small value of normalized speed or jerk [40]. Mean Arrest Period Ratio (MAPR) represents the portion of frames when speed exceeds a percentage (10%) of the maximum speed [40]. As a subject reaches a target position without unnecessary stops, speed profiles will stay less near zero. We expect subjects with limited functional abilities will have more stationary movements and higher values of MAPR than health subjects. A zero-crossing ratio indicates the period of a movement, in which a sign of acceleration or jerk changes. Subjects with limited functional abilities will have stationary or trembling movements, which involve a number of valley shapes in speed profiles. Thus, post-stroke survivors with limited functional abilities expect to have higher zero-crossing ratio of acceleration or jerk than healthy subjects.

Given the list of preprocessed and normalized features, we describe the list of extracted features to train classifiers of performance components in Table 3.

The features of the ’ROM’ component include joint angles ($j_a(t)$), normalized relative trajectories ($nrt_t$), and projected trajectories ($npt_t$). The features of $j_a(t)$ compute elbow and shoulder joint angles. The features of $nrt_t$ measure the distance of wrist and elbow joints with respect to the head joint. The features of $npt_t$ compute the distance of wrist with respect to the head and shoulder joints in X, Y, Z axis.

In addition, we represent the degree of smoothness with various speed-related features (i.e. speed ($sp_t$), acceleration ($ac_t$), jerk ($jk_t$), MAPR ($mapr_t$), zero-crossing ratios ($zc_t$), and normalized speed and jerk ($nspt_t$ and $njk_t$) [40]). As this work includes only upper arm movements, we only present the features of the upper limb (e.g. shoulder, elbow, wrist).
This section describes our approach of developing models to assess the quality of a movement. We first apply standard (a) multi-class classification and (b) regression approaches as the baselines and then compare with the (c) ‘BinToMulti’, proposed approach, threshold models with binary classifications (Figure 4). We utilize ‘Scikit-learn’ [34] and ‘PyTorch’ [31] libraries to implement Decision Trees (DTs), Linear Regression (LR), Support Vector Machine (SVM), Neural Networks (NNs), and Long Short Term Memory (LSTM) Network.

For DTs, we utilize Classification and Regression Trees (CART) [1] to build a prune trees. For LR models, we apply either L1 or L2 regularization to avoid overfitting. For SVMs, we apply either linear or RBF kernels with penalty parameter, C = 1.0. For NNs, we explore various architectures (i.e. one to three layers with 16, 32, 64, 128, 256, 512 hidden units) and with adaptive learning rate with various initial learning rates (i.e. 0.005, 0.001, 0.01, 0.1). To explore the usefulness of applying a sequential model, we implement LSTM networks. As a therapist assesses the quality of a motion after observing patient’s entire motion, LSTM networks have many-to-one architecture (Figure 5). We apply 0.5 drop-out to LSTM layers, explore one to three LSTM layers with 128, 256, 512 hidden units, and also apply three fully connected layers (the same hidden unit size with LSTM layers) to generate an output. For training NNs and LSTM networks, we apply ‘AdamOptimizer’ with mini-batch size of 5 and epoch = 1. Cross-entropy loss is utilized for classification and mean-square-error loss for regression approach.

We train classifiers of individual performance components using the features in Table 3. For each exercise trial, we compute five statistics (i.e. max, min, range, average, and standard deviation) of features at each time-stamp. Sequential models utilize the feature matrix over the entire time-stamps. Non-sequential models (i.e. DTs, LRs, SVMs, and NNs) apply feature vectors at the last time-stamp that summarizes the entire motion with statistics of features.

Table 3: The list of extracted features for modeling classifiers of performance components

<table>
<thead>
<tr>
<th>Performance Components</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROM</td>
<td>(ja_t(b_p, sh, eb), ja_t(sh, eb, wr)) and (nrt_t(b, s)) where (b = \text{hd} = \text{hd}) and (s \in {\text{eb}, \text{wr}}) and (npt_t(j, j, c)) where ((j, j, j) = \text{hd, hd, wr}) for (c \in C)</td>
</tr>
<tr>
<td>Smoothness</td>
<td>(sp(j, ac(j), f_c(j)), nsp(j, ac(j), f_c(j))), (maps(j, ac(j), f_c(j)), nmaps(j, ac(j), f_c(j))) where (j \in {\text{elbow, wrist}})</td>
</tr>
<tr>
<td>Compensation</td>
<td>((j)\text{ss}(sh_{init}, sh_{final}, sh), ja(hp, sh, eb)) (dpt_t(sh_{init}, hd, c), dpt_t(ss_{init}, ss_{final}, c), dpt_t(sh_{init}, sh, c)) for (c \in C)</td>
</tr>
</tbody>
</table>

For the ‘Compensation’ component, we compute joint angles \(ja_t\) and projected trajectories \(dpt_t\) to distinguish a compensated movement. The features of \(ja_t\) calculate the tilted angle of a spine, the elevated angle of a shoulder, and shoulder abduction angle. The features of \(dpt_t\) measure the distance between the initial and current joint positions of head, spine, and shoulder joints in X, Y, Z axis.

4.2 Models to Assess the Quality of a Movement

This section describes our approach of developing models to assess the quality of a movement. We first apply standard (a) multi-class classification and (b) regression approaches as the baselines and then compare with the (c) ‘BinToMulti’, proposed approach, threshold models with binary classifications (Figure 4). We utilize ‘Scikit-learn’ [34] and ‘PyTorch’ [31] libraries to implement Decision Trees (DTs), Linear Regression (LR), Support Vector Machine (SVM), Neural Networks (NNs), and Long Short Term Memory (LSTM) Network.

For DTs, we utilize Classification and Regression Trees (CART) [1] to build a prune trees. For LR models, we apply either L1 or L2 regularization to avoid overfitting. For SVMs, we apply either linear or RBF kernels with penalty parameter, C = 1.0. For NNs, we explore various architectures (i.e. one to three layers with 16, 32, 64, 128, 256, 512 hidden units) and with adaptive learning rate with various initial learning rates (i.e. 0.005, 0.001, 0.01, 0.1). To explore the usefulness of applying a sequential model, we implement LSTM networks. As a therapist assesses the quality of a motion after observing patient’s entire motion, LSTM networks have many-to-one architecture (Figure 5). We apply 0.5 drop-out to LSTM layers, explore one to three LSTM layers with 128, 256, 512 hidden units, and also apply three fully connected layers (the same hidden unit size with LSTM layers) to generate an output. For training NNs and LSTM networks, we apply ‘AdamOptimizer’ with mini-batch size of 5 and epoch = 1. Cross-entropy loss is utilized for classification and mean-square-error loss for regression approach.

We train classifiers of individual performance components using the features in Table 3. For each exercise trial, we compute five statistics (i.e. max, min, range, average, and standard deviation) of features at each time-stamp. Sequential models utilize the feature matrix over the entire time-stamps. Non-sequential models (i.e. DTs, LRs, SVMs, and NNs) apply feature vectors at the last time-stamp that summarizes the entire motion with statistics of features.
\( f(x) \) describes the output, confidence score of binary classifications.

\[
\alpha(x) = \sum_{i=1}^{K-1} \| f(x, i) > 0 \| \tag{1}
\]

where \( f(x, i) = f(x) - \theta_i, \theta_i = \frac{1}{K} \| a \| \) is defined to be 1 if a holds and 0 otherwise.

## 5 VALIDATING MODELS TO ASSESS THE QUALITY OF A MOVEMENT

This section describes our experiment to validate models to assess the quality of a movement using the data set from 11 healthy subjects, 15 post-stroke survivors, and two therapists. First, we demonstrate how well the assessment of models can be aligned with a therapist. Second, we show that quantified performance scores of proposed assessment method can have strong correlation with the scores of Fugl-Meyer Assessment, clinically validated motor impairment tool.

### 5.1 Data Collection

We collected an exercise movement dataset using a Kinect v2 sensor with a therapist. Second, we show that quantified performance scores of proposed assessment method can have strong correlation with the scores of Fugl-Meyer Assessment, clinically validated motor impairment tool.

### 5.1.1 Participants

Both healthy subjects and post-stroke survivors contributed to the dataset of three upper limb exercises (Figure 3) after signing the consent form. 11 healthy subjects (10 males and 1 female) with an average and standard deviation of 32.3 ± 5.81 years were recruited to collect unaffected movements. Healthy subjects were instructed to perform 15 repetitions for each exercise.

In addition, 15 post-stroke survivors (13 males and 2 females) with an average and standard deviation age of 63 ± 11.43 years participated in the data collection. A post-stroke survivor performed 10 trials with both affected and unaffected sides for each exercise. Thus, we have 465 trials for each upper limb exercise: 315 trials of unaffected movements and 150 trials of affected movements. The starting and ending frames of exercise movements are manually annotated for the experiment. In addition to recording exercise motions, a therapist assessed post-stroke survivor’s functional ability of using the Fugl Meyer Assessment (FMA). The profiles of 15 post-stroke participants are summarized in Table 4.

### 5.1.2 Ground Truth Scores

We recruited two stroke rehabilitation therapists: the primary and secondary therapists. The primary therapist managed the recruitment of participants in Table 4 and evaluated their functional ability with Fugl Meyer Assessment (FMA). The secondary therapist has an experience in stroke rehabilitation, but no prior interactions with recruited participants.

Two therapists conducted two evaluation phases to generate ground truth scores (Figure 7). In the first phase, each therapist individually watched the videos of participant’s motions and assigned scores of each performance components using the scoring guidelines (Table 1). In the second phase, two therapists discussed their evaluation strategies for an hour by using one sample trial and then individually assigned scores again.

### 5.2 Level of Agreement with Therapist’s Ground Truth

As a post-stroke survivor mainly interacts with a single therapist for the consistent assessment and guidance [11, 39], our study mainly explores how well our proposed method can be calibrated with the primary therapist’s assessment and applicable to multiple patients.

The overall procedure to measure agreement levels with the primary therapist is described in Figure 7. We first measured the agreement between primary and secondary therapists during two evaluation phases, which can provide an idea of how well secondary therapist can align without or with one-hour discussion. We then computed the agreement levels between primary therapist and computer-assisted approaches (i.e. proposed and alternative approaches). Even if alternative approaches (i.e. multi-class classification and regression) are not shown in Figure 7, we also computed their agreement levels using numerical ground truth of primary
therapist’s 2nd evaluation. We then compared agreement levels of the proposed approach with alternative approaches and secondary therapist to analyze any benefits of the proposed method.

Figure 7: Procedure to Compute Agreement with Primary Therapist

For the evaluation of computer-assisted approaches, we apply Leave-One-Subject-Out (LOSO) cross validation on post-stroke survivors, which trains a model with data from all subjects except one post-stroke survivor and test with data from the left-out post-stroke survivor.

According to the experiments of binary classification (Table 5), Decision Trees (DTs) outperform other algorithms for ‘E2-ROM’ (max-depth=3) and ‘E3-Smooth’ (max-depth=4) and Neural Networks (NNs) outperform others algorithms for the rest of exercise performance components: ‘E1-ROM’ with hidden layers (128, 128) and initial learning rate (1e-3), ‘E1-Smooth’ with hidden layers (128, 128, 128) and initial learning rate (1e-3), ‘E1-Comp’ with hidden layers (512, 512) with initial learning rate (1e-1), ‘E2-Smooth’ with hidden layers (512, 512) and initial learning rate (1e-1), ‘E2-Comp’ with hidden layers (256, 256, 256) and initial learning rate (1e-1), ‘E3-ROM’ with hidden layers (128, 128) and initial learning rate (1e-1), ‘E3-Comp’ with hidden layers (256, 256, 256) and initial learning rate (1e-3).

Using these algorithms and parameters, we present the agreement levels of computer-assisted approaches (i.e. ‘Multi-Class’ classification, ‘Regression’, and the proposed, ‘BinToMulti’) in Table 6. The highest F1-scores and lowest MSE of computer-assisted approaches are highlighted in a bold font. In addition, Table 6 presents the agreement levels of secondary therapist’s evaluation, in which the highest F-scores and lowest MSE are highlighted in an italics font.

5.3 Relationship with Fugl Meyer Scores

We analyze the relationship between the sum of computed scores using the proposed method and scores of the Fugl Meyer Assessment (FMA) using a linear regression. The linear model has 0.95 R-squared value (p < 0.001), which indicate a strong linear relationship between computed scores of the proposed method and the FMA scores of 15 post-stroke survivors (Figure 8). Using this linear regression model and the computed scores of exercises, we can estimate the Fugl Meyer score of a post-stroke survivor.

6 DISCUSSION

Our approach provides a method to automatically assess stroke rehabilitation using kinematic features and machine learning algorithms. Several studies have demonstrated the usefulness of kinematic variables to describe motion functional abilities [27, 29, 32, 49]. Furthermore, Olesh et al. described the feasibility of measuring the movement impairment with linear regression between two principal components of four joint angles and qualitative scores from eight subjects [28]. However, their study is limited to individual joint analysis and not applicable for complex task-oriented exercises. The current study involves the experimental designs for assessing task-oriented exercises (i.e. ‘ROM’, ‘Smoothness’ ‘Compensation’ performance components) with more kinematic variables and demonstrates the effectiveness of the proposed automated assessment.
Exercise 2

(a) mse

Table 5: Binary Classification Results

<table>
<thead>
<tr>
<th>Exercise 1</th>
<th>Exercise 2</th>
<th>Exercise 3</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROM</td>
<td>0.9700</td>
<td>0.8318</td>
<td>0.8746</td>
</tr>
<tr>
<td>BinToMulti</td>
<td>0.9472</td>
<td>0.6478</td>
<td>0.4321</td>
</tr>
</tbody>
</table>

Figure 9: Scatter Plots of (1) F1 Scores and MSE, in which an origin (0,0) indicates perfect agreement with ground truth scores, and (2) 1st Evaluation (Red), 2nd Evaluation (Green), Final Evaluation (Blue), and Proposed Approach (Magenta) of Secondary Therapy.

We first empirically show that our proposed “BinToMulti” method can perform equal or better than other approaches (i.e., multi-class classification and regression) by quantifying the performance of our proposed approach.

Learning to Assess the Quality of Stroke Rehabilitation Exercises IUI '19, March 17–20, 2019, Marina del Rey, CA, USA
problem of learning a quantification function into sub-problems, set of binary classification and utilize confidence scores of binary classification for better exploitation of samples and performance improvement. Our approach may be utilized as a way to address imbalance samples in a healthcare application (including computer-assisted stroke rehabilitation) [8], which has a costly process of collecting data samples due to widely distributed subjects. However, it is not safe to assume that this approach can be applicable to any data-sets. Additional theoretical and empirical analysis are necessary to achieve generalizability of this approach.

In addition, our approach is feasible to reproduce primary therapist’s assessment with the decent levels of agreement (Table 6) and estimate a clinically validated functional ability score (Figure 8).

Even if the primary and secondary therapists engaged in one-hour discussion to share and develop common strategies of evaluation, this one-hour discussion is not sufficient to completely reduce therapist’s subjective interpretation on the clinical functional assessment tool. One-hour discussion does not necessarily improve the agreement level on the ‘Smoothness’ and ‘Compensation’ components. For the further improvement on the agreement level between the primary and secondary therapists, they are required to conduct an additional expensive evaluation process, where they have to meet and analyze the disagreed trials one-by-one iteratively until the convergence.

Without having additional expensive evaluation process, our approach can achieve equally good and better agreement with the primary therapist than the secondary therapist without or with one hour discussion. Our proposed method can serve as a low-cost method that reproduces therapist’s assessment with decent agreement to consistently collect patient’s performance data with clinical relevance.

7 CONCLUSION AND FUTURE WORK

In this paper, we proposed a method that utilizes threshold models with binary classifications to assess performance of stroke rehabilitation exercises. This work contributed the empirical study that validates the effectiveness of the proposed approach. For the validation, we collected a dataset with 11 healthy subjects and 15 post-stroke survivors performing three task-oriented exercises and two therapists (i.e. primary and secondary) generating ground truth scores and assessing the functional ability scores of post-stroke survivors with Fugl Meyer Assessment (FMA). We empirically demonstrated that the proposed method can achieve equally good or better level of agreement with the primary therapist than other approaches (i.e. multi-class classification and regression) or the secondary therapist. In addition, we showed that our approach can estimate a clinical FMA score with the sum of computed scores.

Although this study shows a feasibility of the proposed method that can reproduce therapist’s assessment with a decent agreement level (0.84, 0.82, and 0.79 F1-scores on three exercises) and assess the quality of a movement, it is still challenging to reach the perfect agreement level with computer-assisted approaches. According to our experiments, applying a complex model (i.e. LSTM models) does not necessary guarantee to improve the agreement levels with the therapist. In future, we would investigate a learning technique to derive a personalized assessment model and then compare it with an user-agnostic assessment model. Moreover, we would explore a way to generate explanations of an automated assessment method and derive a human-in-the-loop system, which can updates an assessment model based on feedback of a therapist to achieve a better agreement level and usability.

ACKNOWLEDGMENTS

The authors would like to thank Fabio Pereira and Carolina Vieira for discussing the experimental designs and collecting and evaluating the dataset. The authors would also like to thank all participants in the study and the anonymous reviewers for their valuable comments and helpful suggestions. This work is supported by the FCT through the Augmented Human Assistance project (CMUP-ERI/HC1/0046/2013) and SFHR/BD/113694/2015.

REFERENCES

Learning to Assess the Quality of Stroke Rehabilitation Exercises

<table>
<thead>
<tr>
<th>Year</th>
<th>Author(s)</th>
<th>Title and Journal Details</th>
</tr>
</thead>
</table>