Muscle Strength Testing using Wearable Wireless Sensors

Debadri Mukherjee
Centre for Speckled Computing
School of Informatics
University of Edinburgh
D.Mukherjee@sms.ed.ac.uk

DK Arvind
Centre for Speckled Computing
School of Informatics
University of Edinburgh
dka@inf.ed.ac.uk

ABSTRACT

Manual muscle testing and its variants have a long history of use for classifying muscle strengths. For the first time, inexpensive wearable wireless sensors combined with machine learning techniques are used to classify different levels of muscle strength, which addresses some limitations of the manual method. A mean accuracy of 93% was obtained across ten subjects using gyroscope and accelerometer data in classifying four distinct levels of strengths of the biceps brachii muscle when performing muscle contraction under appropriate load. This was reduced by 2% for accelerometer-only data, thus offering a potentially inexpensive and viable solution for testing muscle strength.

Keywords

Muscle strength testing; Orient specks; Classification; *k*-nearest neighbours.

1. INTRODUCTION

Muscle strength can be reduced for a number of reasons such as the result of a stroke, Parkinson's disease or bone fracture. Patients would be required to undergo a regime of rehabilitation to improve their muscle strength. Muscle testing is important to calibrate the strength and track improvements in movement during the treatment. Accurate assessment of muscle strength is therefore central in determining the course of treatment by physiotherapists. The two common techniques for muscle testing are Manual Muscle Testing (MMT) and Handheld Dynamometry (HHD). MMT is a popular choice amongst physiotherapists where the equipment cost of the dynamometer is an issue. However, MMT has been criticised for its seeming lack of objectivity as the physiotherapist's strength is said to influence the assessment [1]. It has been suggested that MMT is more reliable when assessing heavy impairment, but equipment-based testing is preferred in the cases of moderate or mild levels of impairment [1].

This paper proposes a wireless sensor-based approach to test muscle strength accurately and objectively. The method uses an Orient speck [2], which combines a tri-axial gyroscope, accelerometer and magnetometer with wireless networking capabilities. The sensor data is analysed and supervised machine learning methods are applied to classify four distinct levels of muscle strength of the biceps brachii. Ten healthy subjects simulated these four levels of muscle strength by performing the muscle contraction under the application of an appropriate load. We present results obtained using a *k*-nearest neighbour classifier model, firstly on individual datasets corresponding to the subjects, and then on an aggregated dataset, in order to address issues of subjectivity. The features (four) in the dataset(s) were constructed using the gyroscope and accelerometer data. We show further that it is possible to use only the accelerometer data, with a small reduction (2%) in the classification performance.

2. BACKGROUND AND RELATED WORK

In MMT, the examiner (therapist) assigns a numerical (integer) grade between 0 and 5 to a muscle, based on the results of muscle-specific test procedures using the criteria listed in Table 1:

Table 1: Criteria for assigning different grades in MMT.

| Grade	Description	Criterion
0	No Activity	Muscle is completely inert on palpitation or visual inspection.
1	Trace Activity	Detection of contractile activity but no resulting movement.
2	Poor	Full range of motion (ROM) only in a horizontal plane.
3	Fair	Full range of motion (ROM) against the effect of gravity.
4	Good	Full ROM against gravity and test position is held under minimum resistance
5	Normal	Full ROM against gravity and test position is held under maximum resistance.

Muscle testing has been applied in practice for around one hundred years [3], and can be traced back to the pioneering work of Wilhelmine G. Wright [4] and Robert W. Lovett [5]. This paper reports the first use of wireless inertial sensors to classify muscle strength and addresses the known deficiencies of manual muscle testing. Previous work using wireless sensors in this domain has focused mainly on fatigue detection [6][7].
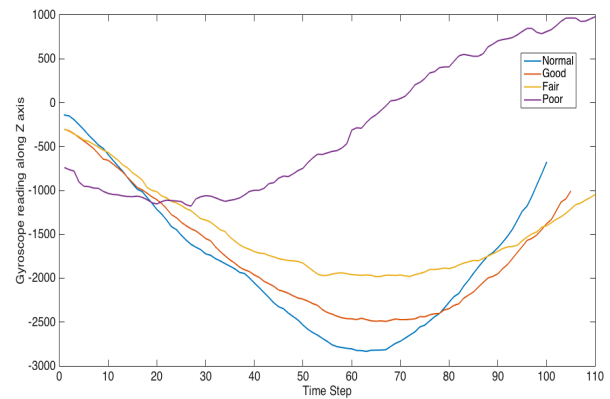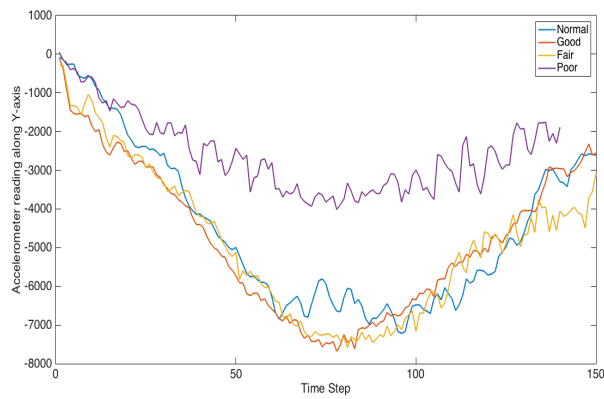
Figure 2: Sample Plots of Accelerometer data along Y-axis (left) and Gyroscope data along Z-axis (right) against time steps during the simulation of different muscle qualities.

3. METHODOLOGY

3.1 Data Collection
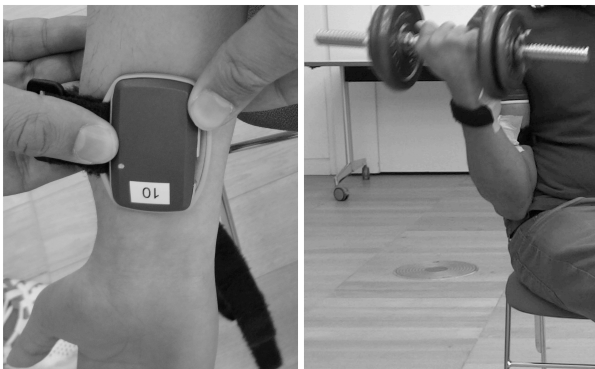
Figure 1: Placement of the Orient speck (left) during elbow flexion (right).

This work has focused on the different strength (weakness) levels of the biceps (bicep brachii), and therefore the motion studied was elbow flexion against gravity (Figure 1, right). The ten able-bodied subjects (5 female and 5 male) were in turn instructed to: "Sit up straight with the (dexterous) arm stretched out by your side and fingers pointing to the ground. The palm should be facing out. Close your grip (or hold the weight). Perform (or attempt) a smooth elbow flexion ending with your forearm touching your shoulder." An Orient speck was placed on the forearm (Figure 1, left) of the subject as close to the palm as possible, which is the point with the greatest degree of freedom with the hand behaving as a Class III lever and the elbow acting as the fulcrum. The four distinct muscle strength levels simulated by each subject have been described in Table 2, which correspond to MMT grades ranging between 2+ and 5 (Table 1).

Each subject demonstrated five instances of elbow flexion for each level (20 in total for each subject), by picking a suitable resistance in the form of weights - R_0, R_1, R_2 and R_3 – depending upon their individual strengths, which would allow them to match the description of muscle strength levels N, G, F and P, respectively:

$$R_0 = 0 \rightarrow \text{No resistance, and } R_1 < R_2 < R_3.$$

Table 2: Description of the four levels of muscle strength.

Class	Name	Description
N	Normal	Full and smooth ROM against gravity.
G	Good	Full ROM against gravity with a little difficulty.
F	Fair	Full ROM against gravity with considerable difficulty.
P	Poor	Partial (incomplete) ROM against gravity.

3.2 Data Analysis

A representative example of the accelerometer readings along the Y-axis and gyroscope readings along Z-axis is presented in Figure 2. There is a central region common to both the plots, which has a characteristic 'U' shape.

In case of the accelerometer data (Figure 2, left), there are undulations displayed as changes in the slope of the graph due to subsequent increase and decrease (or vice versa) in the acceleration due to gravity. During this full range of motion, the alignment of the Y- axis of the accelerometer changes from 90° to 0°, and then from 0° to 90° with respect to the vertical, i.e., direction of the gravitational force. Therefore, at the starting position of this motion, the gravitational force acting upon the Y-axis of the accelerometer is 0 as it is perpendicular to the vertical; it reaches its peak at the half-way point when the forearm is perpendicular to the upper arm along the elbow joint, as in this case the Y-axis of the accelerometer is along the vertical. In general, we observe different degrees of change in the slope at points in the plot (Figure 2, left) for the four muscle qualities. The direction of the slope changes more frequently and the value of the slopes is greater (on average) when simulating greater muscle weakness. On increasing the resistance, the subject faces greater difficulty in completing the motion and this is manifested in jerky movements with frequent accelerations and decelerations. The slope at the individual points of the accelerometer readings also changes direction more frequently and the rate of change of accelerometer readings (the slopes at individual points) is also greater on increasing the resistance, which corresponds to simulating greater level of muscle weakness. Therefore, the behaviour of the slope is an important factor for characterising different levels of muscle strength.

The gyroscope data along the Z-axis (Figure 2, right), also reveals information of relevance. An important observation was that the value of the minima (global) of the curves is inversely proportional to the muscle strength.

3.3 Features for Classification

Average smoothness of acceleration (JC): This was measured as the mean of the absolute value of the derivatives in the accelerometer readings at each time step for the motion. This metric, the jerk cost, was computed as follows:

$$JC = \frac{1}{n}\sum_{i=1}^{n}\left|accl_i^{read} - accl_{i-1}^{read}\right| \quad (1)$$

Directional changes in the slope (SC): The number of changes in direction of the derivative of the accelerometer readings in subsequent time steps where there was a change in acceleration was recorded. This was then divided by the number of time steps so that the value remained independent of the duration of motion.

Minimum value of the accelerometer readings (MA): In order to minimise the impact of noise on this feature, the five lowest accelerometer readings in the motion were identified and this feature was computed as its average.

Accelerometer reading at the end point (DP): This recorded the accelerometer reading at the last time step in the motion. In order to guard against the effect of noise, DP was defined as the mean of the last 5 accelerometer readings.

Range of Motion Indicator (ROM): The features MA and DP had the potential to detect partial range of motion based on the analysis of the data: the former for partial range of motion below $90°$, and the latter for cases beyond $90°$. They were first normalised and then combined into a binary feature, ROM, which served as an indicator for full range of motion, and was set as follows:

$$ROM = \begin{cases} 1 & DP > 0.5 \text{ or } MA < 0.5 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The rationale for these conditions was that the normalised value of MA should be high and that of DP should be low on completing the full range of motion. Setting the threshold at a value of 0.5 also guarded against noise in the data. The replacement of the two features, MA and DP, with ROM ensured that we did not overfit to the noise present in these features.

Minimum value of the gyroscope reading (MinG): As the name suggests, the feature noted the minimum value across the gyroscope readings in the motion inspired by the specific observations made in Figure 2 (right). It was calculated exactly the same way as MA to guard against noise in the sensor data.

In summary, four features were constructed for the classification task. Three (SC, JC and ROM) of them were calculated using the accelerometer readings only, whereas the gyroscope readings were used to determine MinG. All of these features were normalised prior to their use in classification.

3.4 Classifier Model

The implementation of *k*-Nearest Neighbour (*k*-NN) classifier in the Weka tool [10] (known as '*iBk*' [9]) was used to perform the classification task using the features constructed from the sensor data. As a non-parametric classifier, *k*-NN does not require strong assumptions about the underlying data distributions. These models are known to be more flexible in comparison to their parametric counterparts [8]. This aspect ensured that a lower bias was introduced by not making subject-specific assumptions about the data and consequently made the analysis of the results more coherent. This model was computationally inexpensive as well, thanks to the low dimensionality of the data.

4. RESULTS

The mode of evaluation was set to 10-fold cross validation for all the cases as it is a robust way of assessing classification performance. The area under the receiver operating characteristic (ROC) curve and the F-measure along with classification accuracy (Table 3). The ROC Area reflects the classification performance without regard to class distribution and error costs [9]. A high value of F-measure (harmonic mean of precision and recall) indicates high values of both Recall and Precision that are useful measures of performance.

4.1 Individual Results

Table 3: Classification Results using four features.

Subj.	Instances	k	Accuracy	F-measure	ROC Area
1	19 (N:5,G:4,F:5,P:5)	3	100%	1	1
2	18 (N:5,G:4,F:4,P:5)	3	94.44%	0.94	0.99
3	18 (N:4,G:5,F:5,P:4)	3	100%	1	1
4	17 (N:3,G:4,F:5,P:5)	1	94.12%	0.94	0.96
5	17 (N:3,G:4,F:5,P:5)	1	82.35%	0.82	0.82
6	19 (N:5,G:4,F:5,P:5)	3	94.74%	0.95	0.96
7	17 (N:3,G:5,F:4,P:5)	1	82.35%	0.82	0.90
8	19 (N:4,G:5,F:5,P:5)	3	94.74%	0.95	1
9	20 (N:5,G:5,F:5,P:5)	3	95%	0.95	1
10	18 (N:3,G:5,F:5,P:5)	1	88.89%	0.88	0.89

The results for the ten subjects are presented in Table 3: the total number of instances in the dataset and their distribution across the 4 classes, Normal (N), Good (G), Fair (F) and Poor (P), are detailed in the second column. Four features were used for the classification task, namely, SC, JC, ROM and MinG. The value of parameter *k* for the *k*–NN classifier used in each case is reported in the third column. The F-Measure and ROC Area reported in each case are the weighted averages across the classes weighed by the number of instances in the class. A mean classification accuracy of $92.6 \pm 6.3\%$ was obtained in this case.

4.1.1 Assessing the contribution of the Gyroscope

The gyroscope data was used to construct only one feature in the dataset – MinG, and its impact on the classification performance was evaluated by removing this feature and measuring the classification performance based on the other three features. A mean accuracy of $90.8 \pm 8.7\%$ across the ten subjects was obtained. A detailed study of the results revealed that using only

the accelerometer data improved the classification performance for three subjects and did not impact the results for another three. However, the mean accuracy decreased slightly (1.78%) when the gyroscope data was not used. This is a price worth paying given that the gyroscope is usually an order of magnitude more expensive in terms of its price and power consumption. Therefore the rest of the results in this paper do not use the 'MinG' feature.
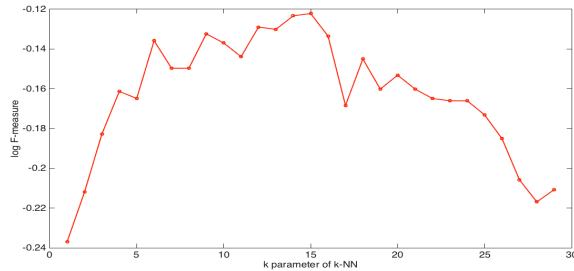
4.2 Classification on the Full Dataset

Figure 3: Plot of log F-Measure (y-axis) against *k*.

The lack of objectivity has been one of the main criticisms of MMT. In order to address this issue, results are presented in this subsection for classification on the 'full' dataset formed by aggregating the data from the 10 subjects, resulting in 182 instances. A systematic search of the state space was conducted to set the value of k for the k-NN classifier. Figure 3 shows the effect of k on performance and justifies the choice of $k = 15$. We obtained a classification accuracy of 88.46%. The weighted F-measure and weighted ROC Area were 0.89 and 0.96, respectively.

Table 4: Confusion Matrix

classified ➔	N	G	F	P
N	38	2	0	0
G	0	32	13	0
F	0	6	42	0
P	0	0	0	49

5. CONCLUSION

The results have highlighted the potential of using the Orient specks for testing muscle strength objectively and accurately. We were able to classify the four levels of muscle strength (weakness) with approximately 93% accuracy across the 10 subjects when using both gyroscope and accelerometer data. The use of accelerometer data alone resulted in a modest dip of only 2% in the classification performance. Thus, a 3-axis accelerometer with wireless networking capabilities could be used as a substitute. The objectivity of the approach was validated by performing classification on the full dataset (combined data of all ten subjects with varying muscularity). An accuracy of 88.5%, just 2.5% lower compared to the mean accuracy of the individual results, demonstrated that the model did not over-fit to the individual subjects.

The majority of the misclassification errors were in distinguishing between 'Good' and 'Fair' muscle strength levels. Some of the errors in distinguishing between 'Good' and 'Normal' may be attributed to noise in the data introduced by the subjects simulating weaknesses which is unlikely to be present in patients during clinical trials. Despite the presence of noise, it is encouraging to note that there were no errors in classifying 'Poor'

muscle strength and all of the errors made were in distinguishing between subsequent levels of muscle strength, say between 'Good' and 'Fair'. We were able to determine perfectly whether full range of motion was completed and the errors in distinguishing between two classes occurred when these shared a boundary.

In conclusion, the work presents a possible solution to muscle strength testing, which addresses the issues of MMT, but at the same time is inexpensive and less complicated compared to approaches such as equipment based testing. In the future, this approach may be part of a self-management strategy for the patient to follow the exercise regime at home and assess improvements objectively and share a digital audit trail of progress with the remote physiotherapist.

6. ACKNOWLEDGEMENTS

7. REFERENCES
[1] Helen Hislop, Dale Avers, and Marybeth Brown. Daniels and Worthingham's Muscle Testing: Techniques of Manual Examination and Performance Testing. Saunders, 9th edition, 2013.

[2] A.D. Young, M. J. Ling, D. K. Arvind, ``Orient-2 - A Wireless Real- time Posture Tracking System using Local Orientation Estimation'', 4th Int. Workshop on Embedded Networked Sensors, ACM.

[3] Nancy Berryman Reese. Muscle and Sensory Testing. Elsevier Health Sciences, 3rd edition, 2013.

[4] Wilhelmine G. Wright. Muscle training in the treatment of Infantile Paralysis. Boston Medical Surgery J., 1912.

[5] Robert W. Lovett. Treatment of infantile paralysis. Preliminary Report. JAMA, 1915.

[6] Jian Zhang, Thurmon E. Lockhart, and Rahul Soangra. Classifying Lower Extremity Muscle Fatigue during Walking using Machine Learning and Inertial Sensors. Annals of biomedical engineering, 2014.

[7] Kang-Ming Chang, Shin-Hong Liu, and Xuan-Han Wu. A Wireless sEMG Recording System and its Application to Muscle Fatigue Detection. Sensors, 2012.

[8] Kevin P. Murphy. Machine Learning: A Probabilistic Perspective. MIT Press, 2012.

[9] Ian H. Witten, Eibe Frank and Mark A. Hall. "Data Mining: Practical Machine Learning Tools and Techniques". The Morgan Kaufmann Series in Data Management Systems. Third Edition, 2011.

[10] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009). The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.