A System for Accelerometer-Based Gesture Classification Using Artificial Neural Networks

Robert M. Stephenson, *Member, IEEE*, Ganesh R. Naik, *Senior Member, IEEE* and
Rifai Chai, *Member, IEEE*

***Abstract*— A great many people suffer from neurological movement disorders that render typical hardware interface devices ineffective. A need exists for a universal interface device that can be trained to accept a wide range of inputs across varying types and severities of movement disorders. In this regard, this paper details the design, testing and optimization of an accelerometer-based gesture identification system. A Bluetooth-enabled IMU mounted on the wrist provides hand motion trajectory information to a local terminal. Several techniques are applied to decrease the intra-class variance and reduce classifier complexity including filtering, segmentation and temporal scaling. Datasets consisted of 520 training samples, 260 validation samples and a further 520 testing samples. A multi-layer feed forward artificial neural network (ML-FFNN) was used to classify the input space into 26 different classes. Initial system accuracy, using arbitrary hyperparameters was 77.69% with final optimized accuracy at 99.42%.**

I. Introduction

Human interface devices (HIDs) form a layer of technological abstraction which translates user intent into practical control signals. HIDs can be categorized by the way users interact with them (voice, touch, movement, etc.) and rated on their aptitude at capturing a user's true intention.

Typical HIDs are often unsuitable for users with disabilities. A keyboard, for example, requires consistent fine motor control. Over 20% of the population suffers from some form of neurological movement disorder [1]. Parkinson's Disease, affecting over 1% of the population over 50, is manifested by characteristic tremors, bradykinesia and akinesia (rendering a typical keyboard inadequate). The consequences of both ischemic and hemorrhagic strokes can range from mild motor function impairment to complete contralateral paralysis. While in many cases stroke victims can regain some of the lost motor functions through rehabilitation, it is a difficult process with varying results [2].

While some specialized HIDs exist to improve the quality of life of sufferers, they often target a specific characteristic and lack adaptability across differing disorders or even levels of severity within the same disorder. There exists a need for a universal solution that can be trained on an ad hoc basis to apply specifically to a user's condition.

A gesture is defined as a movement of part of the body, usually the hands, to express an idea or meaning. To track gestures, two core components are required: limb trajectory recording (motion tracking) and classification. Typical HIDs place the responsibility of the user for operational efficacy so a gesture tracking system that can instead adapt to the user would be an ideal solution. Absolute localization within the user's environment is unnecessary (we are only interested in the relative motion of the hand) rendering most unwieldy and expensive motion tracking solutions as excessive and inadequate for user mobility.

Previous attempts at accelerometer and EMG based gesture classification have been made using Fischer Discriminant Analysis (FDA) [3] and Bayesian Networks (BN) [4, 5] with reported accuracies between 85-96%. One study [5] explores the fusion of both accelerometer and EMG sensors, but reports only minor improvements over pure EMG.

Gestures cover a large area of possibilities ranging from small directive movements to elaborate sign language expressions. One study [6] provides a framework for the classification of Chinese Sign Language (CSL) using a hidden Markov model (HMM) while another [7] classifies Greek Sign Language based on intrinsic mode entropy (IMEn) and Mahalanobis distance (both still requiring EMG however).

This paper explores the design, testing and optimization of a purely accelerometer based gesture identification system with classification performed by an artificial neural network. Large amounts of custom gesture classes, which can be cumbersome for the subject to keep track of, are easily substituted with a pre-existing alphabet of symbols that can be virtually traced. Each unique character provides a motion trajectory path for the user to follow. Thus, this study uses the English alphabet as the basis for its gesture catalogue.

Robert M. Stephenson is with the Centre for Autonomous Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Broadway NSW 2007, Australia (Robert.Stephenson@student.uts.edu.au).

Ganesh R. Naik and Rifai Chai are with Centre for Health Technologies, Faculty of Engineering and Information Technology, University of Technology Sydney, Broadway NSW 2007, Australia (Ganesh.Naik@uts.edu.au, Rifai.Chai@uts.edu.au).

II. Methodology

A. System Overview

Trajectory data is first recorded using an inertial measurement unit (IMU) on the subject's wrist (II.B). A combination of filtering, segmentation and temporal scaling (II.D) is performed to control variance, noise and dimensionality (the effect of which is examined in III.A). The remaining vector is classified by the trained ML-FFNN (II.E) the output of which can be used for any number of control applications.

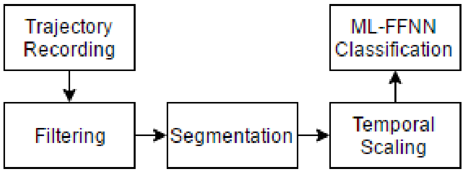

Figure 1. System block diagram

B. Hardware

The core of the hardware device is a 9-DOF inertial measurement unit (IMU). Only the accelerometer and gyrometer components are used which provide sensitivity in the order of 0.061mg/LSB and 8.75mbps/LSB respectively (enough to detect even the smallest of involuntary movements). The IMU data stream is transmitted wirelessly via an SPI Bluetooth interface to a nearby terminal. The device is powered by a small LiPo cell connected to a DC-boost circuit (see Fig 2.). Each component is sewn onto a medical-grade flexible wristband.

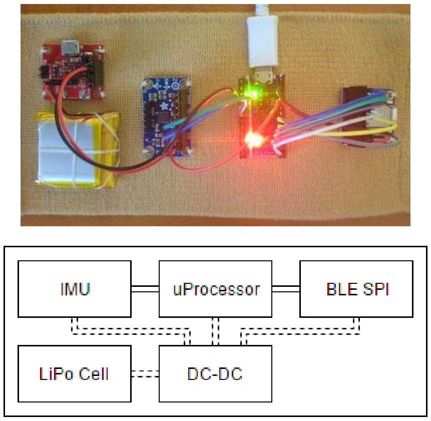

Figure 2. Wristband hardware component layout

C. Data Collection

The IMU was attached to the dorsal wrist of a healthy 24-year-old male subject's dominant arm. A gesture is initiated with the subject's hand at an arbitrary position approximately 20-50cm in front of the chest with the index finger extended as a guide. This is designated as the rest position. The subject then draws a symbol in an imaginary planar surface perpendicular to their line of sight (see Fig 3.) before returning to the rest position. The subject was instructed to make the gesture as naturally as possible within the 3.5s window i.e. without procedural bias or inclination.

Each gesture was recorded 50 times for a total of 1300 samples (560 for training, 260 for validation and a further 560 for testing). Each sample contained 6 waveforms (one for each acc./gyr. axis) recorded at 57Hz resulting in approximately 1300 points per sample.

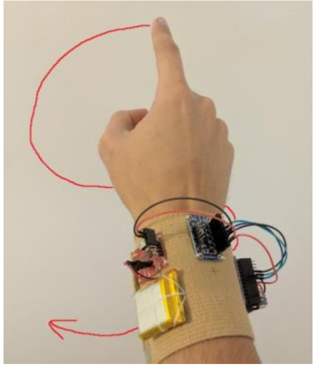

Figure 3. Gesture formation for the letter s

D. Pre-Processing and Dimensionality Reduction

Steps are taken to reduce the number of redundant or unknown variables in each sample. The orientation of the subject within their gravitational frame of reference (present as a bias on a sensor axis) is irrelevant and is thus removed:

$$y_{ji} = y_{ji} - \frac{\sum_{i=1}^{n} y_{ji}}{n} \qquad (1)$$

where y_{ji} is the i^{th} point of the j^{th} waveform and n is the number of waveforms ($n = 6$). Involuntary muscle movement and high frequency noise is removed with a rolling average:

$$y_{ji} = \frac{y_{ji-1}+y_{ji}+y_{ji+1}}{3} \qquad (2)$$

Inconsistencies introduced by the subject consequentially locate the voluntary motion segments in each recording anywhere within the 3.5s window. Without segmentation, the high variance would likely result in poor classification. Different wave features were experimentally tested for identifying active segments, but the highest accuracy was achieved by looking at rolling average variance:

$$\left(\frac{y_{ji-1}+y_{ji}+y_{ji+1}}{3} - \frac{\sum_{i=1}^{n} y_{ji}}{n}\right)^2 > k \qquad (3)$$

where k is a threshold value (0.3 yielded highest detection accuracy in trials). The first i^{th} point to validate inequality (3) represents the start of the active region. The end of the active region is identified by the same algorithm fed points in reverse order.

The varying duration of each motion required to produce a gesture means the number of data points within the active region also varies. The static number of classifier inputs i means the active region needs to be remapped onto a fixed number of points. Simple linear interpolation allows projection of an arbitrary number of points to a destination value p_d:

$$y_{ji*} = y_a + (i^* - a) \times \left(\frac{y_b - y_a}{b-a}\right) \qquad (4)$$

where i^* is the index of the source point i multiplied by the ratio $\frac{p_s}{p_d}$ and a and b are the lower and upper bounds of i^* respectively. Data is normalized at this point.

D. *Classification*

A multi-layer feed forward artificial neural network (ML-FFNN) was used to classify processed samples:

$$\mathbf{z_k}(\boldsymbol{x},\boldsymbol{w}) = f\left(\sum_{i=1}^{n} \boldsymbol{w_{kj}} \cdot f\left(\sum_{i=1}^{n} \boldsymbol{w_{ji}} \cdot \boldsymbol{x_i}\right)\right) \quad (5)$$

where the activation function is bipolar logistic $f(v) = \frac{1-e^{-v}}{1+e^{-v}}$. Supervised training was done through back-propagated stochastic gradient descent:

$$\Delta w_{kj} = -\eta \cdot \frac{\partial E}{\partial w_{ki}} \quad (6)$$

where E is the Least Squared Error (LSE) cost function:

$$E(d,z) = \frac{1}{2}\sum_{k=1}^{o}(d_k - k_k)^2 \quad (7)$$

Training was cross-validated at the end of each cycle and stopped at the global E_v minimum. Weights are initialized with random values before training.

III. Results

A. *Variance*

Inter and intra class variability is examined in one study [8] as a means of identifying muscle activation patterns in EMG data. In this study, the variance is used as a tool to quantitatively measure the effect of pre-processing and dimensionality reduction on classification accuracy. High inter-class variance and low intra-class variance reduces the necessary complexity of the classifier which in turn reduces overall training times. The variance of each point y_{ji} from the global class mean $\bar{y}_j$ is averaged over p points in waveform j then averaged across all waveforms in the sample to provide a single measure of variation for each gesture:

$$\sigma_{avg} = \frac{\sum_{j=1}^{n}\left(\frac{\sum_{i=1}^{p}\left(y_{ji}-\bar{y}_j\right)^2}{p}\right)}{n} \quad (8)$$

The averaged variance σ_{avg} between samples after processing (II.D) is shown in Table I. The diagonal represents the variation between samples of the same class. It should be noted that the values along the diagonal are 1-2 orders of magnitude lower than that of differing classes (evidence of effective pre-processing).

TABLE I. Average Inter-Class Variance with Intra-Class Variance on Diagonal (Only a to f Shown)

	a	***b***	***c***	***d***	***e***	***f***
a	4	63	352	16	378	88
b	63	10	294	72	346	140
c	352	294	6	355	199	235
d	16	72	355	4	296	112
e	378	346	199	296	12	310
f	88	140	235	112	310	8

B. *Initial Findings*

The network was arbitrarily configured with the following parameters: *i*=300, *j*=20 and *k*=26$j = 20$ where *i* and *k* representing the input vector dimensions and the number of classes respectively. Using this configuration, the ML-FFNN correctly classified 404 of the 520 test gestures (77.69%).

Table II shows a portion of the classification results before optimization. The left column is the true sample class and the top row is the classifier output. Correct classifications are shown along the diagonal. Using the configuration above, the ML-FFNN was unable to correctly classify the letters *a*, *b*, *d*, *e*, *h* and *n*.

TABLE II. Class Confusion Matrix (Only a to l Shown)

	a	***b***	***c***	***d***	***e***	***f***	***g***	***h***	***i***	***j***	***k***	***l***
a	**0**	0	0	0	0	0	0	0	0	0	2	9
b	0	**8**	0	0	0	0	0	0	0	0	0	0
c	0	0	**20**	0	0	0	0	0	0	0	0	0
d	0	0	0	**0**	0	0	0	0	0	0	0	0
e	0	0	0	0	**1**	0	0	0	0	0	0	0
f	0	0	0	0	0	**20**	0	0	0	0	0	0
g	0	0	0	0	0	0	**20**	0	0	0	0	0
h	0	0	0	0	0	0	0	**0**	0	0	0	0
i	0	0	0	0	0	0	0	0	**20**	0	0	0
j	0	0	0	0	0	0	0	0	0	**20**	0	0
k	0	0	0	0	0	0	0	0	0	0	**17**	0
l	0	0	0	0	0	0	0	0	0	0	0	**20**

C. *Optimization*

The number of output nodes k remains constant. The number of input nodes i is controlled through the interpolation described in (4) which determines the final dimensions of the input vector. The number of hidden neurons j is progressively increased starting from the minimum number required to separate the input space into m regions [9]:

$$j_{min} = \log_2 \frac{k^2+k+2}{2} \quad (9)$$

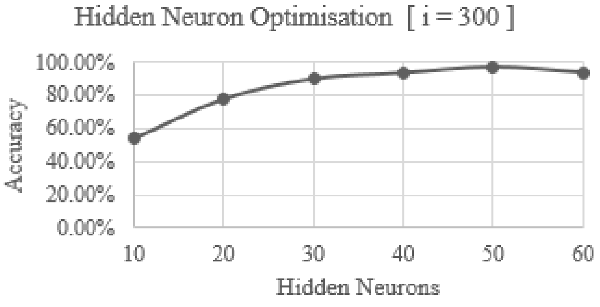

Figure 4. Hidden neuron optimization for $i = 300$

A steady rise in classifier accuracy is shown in Fig 4. as j is varied from 10 to 60 for a fixed number of input nodes $i = 300$. Accuracy peaks at 97.12% with $j = 50$ before starting to decline. To confirm this trend and optimize the number of input nodes, the process is repeated for different values of i (see Table III.).

It can be seen from Table III. that accuracy consistently peaks at 50 hidden neurons, independent of the number of input nodes. Highest overall accuracy was achieved with $i = 120$.

TABLE III. INPUT VS. HIDDEN NEURON CLASSIFICATION ACCURACY

	i				
j	**300**	**240**	**180**	**120**	**60**
10	54.04%	50.00%	56.15%	62.31%	70.00%
20	77.69%	77.50%	75.19%	84.62%	82.12%
30	89.81%	97.50%	95.77%	98.85%	97.69%
40	93.46%	95.77%	95.38%	96.54%	96.15%
50	**97.12%**	**97.69%**	**97.31%**	**99.42%**	**98.46%**
60	93.85%	96.15%	96.73%	96.35%	98.46%

The optimized network parameters ($i = 120$, $j = 50$ and $k = 26$) allowed the ML-FFNN to correctly identify 517 of the 520 test gestures yielding a final accuracy of 99.42%. Table IV. shows a portion of the optimized classifier output.

TABLE IV. OPTIMISED CLASS CONFUSION MATRIX (ONLY *A* TO *L* SHOWN)

y	***a***	***b***	***c***	***d***	***e***	***f***	***g***	***h***	***i***	***j***	***k***	***L***
a	**20**	0	0	0	0	0	0	0	0	0	0	0
b	0	**19**	0	0	0	0	0	0	0	0	0	0
c	0	0	**20**	0	0	0	0	0	0	0	0	0
d	0	0	0	**20**	0	0	0	0	0	0	0	0
e	0	0	0	0	**20**	0	0	0	0	0	0	0
f	0	0	0	0	0	**20**	0	0	0	0	0	0
g	0	0	0	0	0	0	**20**	0	0	0	0	0
h	0	0	0	0	0	0	0	**20**	0	0	0	0
i	0	0	0	0	0	0	0	0	**20**	0	0	0
j	0	0	0	0	0	0	0	0	0	**20**	0	0
k	0	0	0	0	0	0	0	0	0	0	**20**	0
l	0	0	0	0	0	0	0	0	0	0	0	**20**

The misclassifications of b as p and n as h can be linked to the similarity in their trajectories with the most significant difference being the duration of the initial stroke. This is further supported by the variance of these pairs: $b_p = 26$ and $n_h = 19$. Variance between other classes is generally an order of magnitude higher e.g. k_f=183 as shown in Table I.

IV. CONCLUSION

This paper has presented the design, testing and optimization of an accelerometer-based gesture identification system. While research has been conducted on the use of lower dimensional waveform features for classification, such as Mean Absolute Value [5, 6] or Fast Fourier Transform [10], this study has shown that a combination of procedural time-domain techniques including filtering, segmentation and temporal scaling can also be effective. Despite this result, it should be noted that further training and testing is required for subjects with movement disorders to be able to properly gauge the potential for this system. Different processing techniques may be required with the introduction of tremors (such as additional filtering algorithms and revisions to the segmentation process). The gesture window may have to be increased for subjects with bradykinesia while the gesture catalogue would have to be revised for simpler or less intricate movements.

The decision to use a single test subject in this study is justified by the original purpose of having a system trained by the end user to address the need for specific conditional adaptation. This study can be expanded to observe the robustness to a variety of users for single-user training sets and the accuracy of the system for users with naturally higher intra-class variance.

The device is wireless making it theoretically possible to translate the software running on the local terminal to a smartphone worn by the user, opening the door to control-based applications for immobile users. Further research could also lead to real-time sign language translation and hand-eye coordination programs.

REFERENCES

[1] W. G. Ondo and R. Young, "Gait and Movement Disorders," *American Academy of Neurology*, pp. 1-21, 2013.

[2] M. F. Walker, K. S. Sunnerhagen, and R. J. Fisher, "Evidence-based community stroke rehabilitation," *Stroke,* vol. 44, pp. 293-297, 2013.

[3] J. K. Oh, C. Sung-Jung, B. Won-Chul, C. Wook, C. Eunseok, Y. Jing, C. Joonkee, and K. Dong Yoon, "Inertial sensor based recognition of 3-D character gestures with an ensemble classifiers," in *Ninth International Workshop on Frontiers in Handwriting Recognition*, 2004, pp. 112-117.

[4] C. Sung-Jung, O. Jong Koo, B. Won-Chul, C. Wook, C. Eunseok, J. Yang, C. Joonkee, and K. Dong Yoon, "Magic wand: a hand-drawn gesture input device in 3-D space with inertial sensors," in *Ninth International Workshop on Frontiers in Handwriting Recognition*, 2004, pp. 106-111.

[5] X. Chen, X. Zhang, Z. Y. Zhao, J. H. Yang, V. Lantz, and K. Q. Wang, "Hand Gesture Recognition Research Based on Surface EMG Sensors and 2D-accelerometers," in *2007 11th IEEE International Symposium on Wearable Computers*, 2007, pp. 11-14.

[6] X. Zhang, X. Chen, Y. Li, V. Lantz, K. Wang, and J. Yang, "A Framework for Hand Gesture Recognition Based on Accelerometer and EMG Sensors," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans,* vol. 41, pp. 1064-1076, 2011.

[7] V. E. Kosmidou and L. J. Hadjileontiadis, "Intrinsic mode entropy: An enhanced classification means for automated Greek Sign Language gesture recognition," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2008, pp. 5057-5060.

[8] S. W. Lee, K. M. Wilson, B. A. Lock, and D. G. Kamper, "Subject-Specific Myoelectric Pattern Classification of Functional Hand Movements for Stroke Survivors," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 19, pp. 558-566, 2011.

[9] H. B. Demuth, M. H. Beale, O. De Jess, and M. T. Hagan, *Neural network design*: Martin Hagan, 2014.

[10] D. Zhuxin, U. C. Wejinya, Z. Shengli, S. Qing, and W. J. Li, "Real-time written-character recognition using MEMS motion sensors: Calibration and experimental results," in *2008 IEEE International Conference on Robotics and Biomimetics*, 2009, pp. 687-691.