Future of Information and Communication Conference (FICC) 2018
5-6 April 2018 | Singapore

A Generic Multi-modal Dynamic Gesture Recognition System using Machine Learning

Gautham Krishna G$^{\alpha}$, Karthik Subramanian Nathan$^{\beta}$, Yogesh Kumar B$^{\gamma}$, Ankith A Prabhu$^{\delta}$, Ajay Kannan$^{\pi}$, Vineeth Vijayaraghavan$^{\epsilon}$
Research Assistant$^{\alpha\gamma}$, Undergraduate Student$^{\beta\delta\pi}$, Director - Research & Outreach, Solarillion Foundation, Chennai, India$^{\epsilon}$
Solarillion Foundation$^{\alpha\gamma\epsilon}$, College of Engineering, Guindy$^{\beta\pi}$, SRM University$^{\delta}$
Email: (gautham.krishna$^{\alpha}$, nathankarthik$^{\beta}$, yogesh.bkumar$^{\gamma}$, ankithprabhu$^{\delta}$, ajaykannan$^{\pi}$, vineethv$^{\epsilon}$)@ieee.org

***Abstract*—Human computer interaction facilitates intelligent communication between humans and computers, in which gesture recognition plays a prominent role. This paper proposes a machine learning system to identify dynamic gestures using tri-axial acceleration data acquired from two public datasets. These datasets, uWave and Sony, were acquired using accelerometers embedded in Wii remotes and smartwatches, respectively. A dynamic gesture signed by the user is characterized by a generic set of features extracted across time and frequency domains. The system was analyzed from an end-user perspective and was modelled to operate in three modes. The modes of operation determine the subsets of data to be used for training and testing the system. From an initial set of seven classifiers, three were chosen to evaluate each dataset across all modes rendering the system towards mode-neutrality and dataset-independence. The proposed system is able to classify gestures performed at varying speeds with minimum preprocessing, making it computationally efficient. Moreover, this system was found to run on a low-cost embedded platform – Raspberry Pi Zero (USD 5), making it economically viable.**

***Keywords*—Gesture recognition; accelerometers; feature extraction; machine learning algorithms**

I. Introduction

Gesture recognition can be defined as the perception of non-verbal communication through an interface that identifies gestures using mathematical, probabilistic and statistical methods. The field of gesture recognition has been experiencing a rapid growth amidst increased interests shown by researchers in the industry. The goal of current research has been the quick and accurate classification of gestures with minimalistic computation, whilst being economically feasible. Gesture recognition can find use in various tasks such as developing aids for the audio-vocally impaired using sign language interpretation, virtual gaming and smart home environments.

Modern gesture recognition systems can be divided into two broad categories - vision based and motion based systems. The vision based system proposed by Chen et al. in [1] uses digital cameras, and that proposed by Biswas et al. in [2] uses infrared cameras to track the movement of the user. For accurate classification of the gestures, these systems require proper lighting, delicate and expensive hardware and computationally intensive algorithms. On the other hand, motion based systems use data acquired from sensors like accelerometer, gyroscope and flex sensor to identify the gestures being performed by the user. Of late, most gesture recognition systems designed for effective interaction utilize accelerometers for cheaper and accurate data collection.

Accelerometer-based hand gesture recognition systems deal with either static or dynamic gestures as mentioned in [3]. Static gestures can be uniquely characterized by identifying their start and end points, while dynamic gestures require the entire data sequence of a gesture sample to be considered. Constructing a dynamic gesture recognition system that is compatible with any user is difficult, as the manner in which the same gesture is performed varies from user-to-user. This variation arises because of the disparate speeds of the dynamic gestures signed by users. To tackle this problem, a common set of features which represent the dynamic nature of the gestures across various users should be selected.

The authors of this paper propose a gesture recognition system using a generic feature set, implemented on two public datasets using accelerometers - uWave and Sony, as shown in [4] and [5], respectively. This system has been trained and tested across various classifiers and modes, giving equal importance to both accuracy and classification time, unlike most conventional systems. This results in a computationally efficient model for the classification of dynamic gestures which is compatible with low-cost systems.

The rest of the paper is organized as follows. Section II presents the related work in the area of Gesture Recognition. Section III states about the problem statement of the paper. Sections IV & V deal with the datasets and pre-processing used in building the model. Section VI showcases the features extracted from the datasets. Section VII discusses about the different modes provided to the end-user. Section VIII describes the model used and the experiments performed. Section IX enumerates the results analyzed in the paper. Section X finally concludes the paper.

II. Related Work

Since the inception of gesture recognition systems, there has been a plethora of research in this domain using accelerometers. Specifically tri-axial accelerometers have been in the spotlight recently owing to their low-cost and low-power requirements in conjunction with their miniature sizes, making them ideal for embedding into the consumer electronic platform. Previous gesture recognition systems have also used sensors such as flex sensors and gyroscope, but they have their own shortcomings. The glove based system proposed by Zimmerman et al. in [6] utilizes flex sensors, which requires intensive calibration on all sensors. The use of the these sensors increases the cost of the system and also makes the system physically cumbersome. These shortcomings make

978-1-5386-2056-4/18/$31.00

inertial sensors like accelerometers and gyroscope, a better alternative. In this paper, datasets employing accelerometers were preferred over gyroscopes, as processing the data from gyroscopes results in a higher computational burden.

Contemporary gesture recognition systems employ Dynamic Time Warping (DTW) algorithms for classification of gestures. For each user, Liu et al. [4] employs DTW to compute the look-up table (template) for each gesture, but it is not representative of all users in the dataset, thereby not generalizing for the user-independent paradigm. To achieve a generic look-up table for each gesture that represents multiple users, the proposed system in [7] uses the concept of idealizing a template wherein, the gestures exhibiting the least cost when internal DTW is performed is chosen as the look-up table. Furthermore, DTW is performed again for gesture classification while testing, making the model computationally very expensive to be used in a low-cost embedded platform. The accelerometer-based gesture recognition system proposed in [8] uses continuous Hidden Markov Models (HMMs), but their computational complexity is commensurate with the size of the feature vectors which increase rapidly. In addition, choosing the optimum number of states is difficult in multi-user temporal sequences, thereby increasing the complexity of estimating probability functions.

Moreover, the length of the time series acceleration values of a gesture sample is made equal and quantization is performed in the system proposed in [4]. This makes the data points either lossy or redundant. However, the paper proposed utilizes the data points as provided in the datasets without any windowing or alteration, thereby decreasing the computational costs whilst not compromising on efficiency.

In the system employed in [9], Helmi et al. have selected a set of features that has been implemented on a dataset that contains gestures performed by a single user, making them gesture-dependent and not taking into account the concept of generic features encompassing multiple users. The need for a generic set of features which capture the similarities between gestures, motivated the authors of this paper to implement a feature set that can be applied to any system.

III. Problem Statement

Previous works in haptic-based gesture recognition systems employed computationally expensive algorithms for identification of gestures with instrumented and multifarious sensors, making them reliant on specialized hardware. Moreover, most gesture recognition systems extract features that are model-dependent, and fail to provide the user a choice between accuracy and classification time. Thus a need for a flexible gesture recognition system that identifies dynamic gestures accurately with minimalistic hardware, along with a generic feature set arises.

To overcome this problem, this paper presents a machine-learning based dynamic gesture recognition system that utilizes accelerometers along with a generic set of features that can be implemented across any model with adequate gesture samples. The system provides the end-user the option to choose between accuracy and classification time, thereby giving equal importance to both.

IV. Datasets

TABLE I. Datasets Characterized by their Respective Attributes

Dataset	Users (U)	Gestures (N_G)	Samples per Gesture (S_G)	Days (N_D)	N_{GS}
uWave (D_u)	8	8	10	7	4480
Sony (D_S)	8	20	20	-	3200

For this paper, the authors have chosen two public gesture datasets, viz. uWave dataset (D_u) and Sony dataset (D_S). The datasets were selected owing to their large user campaign of 8 users, with a multitude of gesture samples for a variety of gestures. D_u encompasses an 8-gesture vocabulary with 560 gesture samples per dataset over a period of 7 days, while D_S consists of a collection of 20 gestures with 160 gesture samples each. This shows the diverse nature of the datasets. Both the datasets are characterized by U users signing N_G gestures with S_G samples per gesture, over N_D days with the total number of gesture samples being N_{GS} per dataset, as shown in Table I.

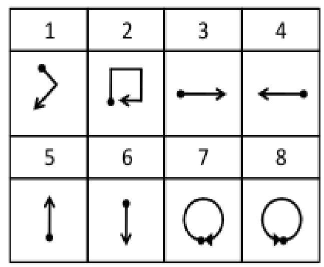

Fig. 1. uWave gestures.

D_u comprises of 3-D accelerations (g-values) that were recorded using a Wii Remote. The start of a gesture is indicated by pressing the ‘A’ button on the Wii Remote and the end is detected by releasing the button as mentioned in [4]. D_u consists of four gestures which have 1-D motion while the remaining gestures have 2-D motion, as depicted in Fig. 1.

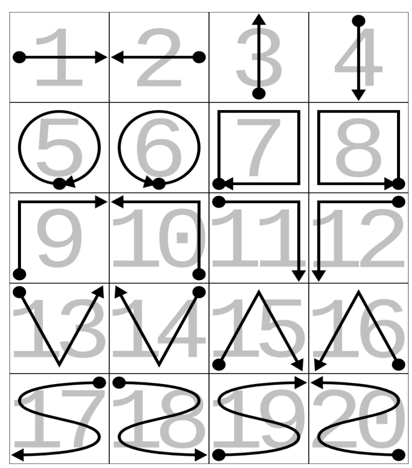

Fig. 2. Sony gestures.

D_S was recorded using a tri-axial accelerometer of a first generation Sony Smartwatch which was worn on the right wrist of the user. Each gesture instance was performed by tapping the smartwatch screen to indicate the start and end of

TABLE II. CONFUSION MATRIX FOR U_M MODE ON D_S; BROWN: GESTURE SIGNED, YELLOW: GESTURE CLASSIFIED, GREEN: CORRECT CLASSIFICATIONS, RED: INCORRECT CLASSIFICATIONS

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	2.5	0	0	97.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	97.5	0	0	0	0	0	0	2.5	0	0	0	0	0	0
8	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	95	5	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	2.5	97.5	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	97.5	2.5	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	97.5	2.5	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	97.5	2.5	0	0	0	0
16	0	0	2.5	0	2.5	0	0	0	0	0	0	0	0	0	2.5	92.5	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

the gesture. The data recorded from the smartwatch consists of timestamps from different clock sources of an Android device, along with the acceleration (g-values) measured across each axis, as mentioned in [5]. In D_S, there are four gestures which have motions in 1-D while the remaining gestures have motions in 2-D as illustrated in Fig. 2.

V. DATASET PREPROCESSING

To generalize the datasets, we eliminate the dependency on time from the datasets, which consist of timestamps along with the raw g-values. The set of g-values per gesture sample (γ), each with a cardinality of n_γ is given by,

$$\gamma = \{g_x^i,\ g_y^i,\ g_z^i\}, \forall\ i \in [1,\ n_\gamma] \quad where,\ g_x^i,\ g_y^i,\ g_z^i\ are\ g-values\ of\ the\ accelerometer \tag{1}$$

On account of the dynamic nature and varying speeds at which different gestures are signed by the users, n_γ varies between different gesture samples. From (1), an overall dataset (G) can be defined by,

$$G = \bigcup_{i=1}^{N_{GS}} \gamma^i \tag{2}$$

Equation (2) is representative of both D_S and D_u datasets. No pre-processing other than elimination of timestamps was done to both the datasets as the g-values in γ would be altered, thereby making the datasets lossy.

VI. FEATURE EXTRACTION

A. Feature Characterization

The proposed system uses the features mentioned in Table III, which have been already utilized in previous accelerometer-based studies. From [10], the significance of Minimum, Maximum, Mean, Skew, Kurtosis and Cross correlation have been shown for Activity Recognition, which is a superset of Gesture Recognition. The inter-axial Pearson product-moment (PM) correlation coefficients as a feature has been signified in [11]. The Spectral Energy as a feature has been illustrated in [12]. The aforementioned features were iteratively eliminated in various domains until the efficiency of the model was the highest.

TABLE III. FEATURE CHARACTERIZATION IN TIME AND FREQUENCY DOMAINS

Features\Domain	Time	Frequency	
		FFT	HT
Mean	✓ (T^1)	✗	✓ (H^1)
Skew	✓ (T^2)	✗	✓ (H^2)
Kurtosis	✓ (T^3)	✗	✗
PM correlation coefficients	✓ (T^4)	✗	✗
Cross correlation	✓ (T^5)	✗	✗
Energy	✗	✓ (F^1)	✓ (H^3)
Minimum	✗	✗	✓ (H^4)
Maximum	✗	✗	✓ (H^5)

B. Domain Characterization

The set of extracted features were then applied in both time and frequency domains. The features that were extracted in the time domain are Mean, Skew, Kurtosis, PM correlation coefficients and Cross correlation. These features along each axis of the gesture samples of a dataset (G) in time domain constitute the TF vector, as given in (3).

$$TF = \bigcup_{j=1}^{5} \{T_x^j,\ T_y^j,\ T_z^j\} \tag{3}$$

The time-domain sequences were converted to frequency domain using Fast Fourier Transform (FFT) and only Energy, as mentioned in [12] was used, while the other features from FFT did not provide any significant improvements. The feature vector (FF) along each axis of the gesture samples of a dataset (G) is represented by (4).

$$FF = \{F_x^1,\ F_y^1,\ F_z^1\} \tag{4}$$

TABLE IV. CONFUSION MATRIX FOR U_I MODE ON D_S; BROWN: GESTURE SIGNED, YELLOW: GESTURE CLASSIFIED, GREEN: CORRECT CLASSIFICATIONS, RED: INCORRECT CLASSIFICATIONS

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	95	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	65	15	0	0	5	0	0	0	0	0	15	0	0	0	0	0
6	0	0	0	0	0	95	5	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	90	10	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0
9	5	0	5	0	0	0	0	0	85	5	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	10	90	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	95	5	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	5	95	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	90	10	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	80	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95	0	0	5
18	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	5	90	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	70

From [13], it can be seen that the Hilbert Transform (HT) hypothesized as a feature was successful in providing a competitive recognition rate for camera-based handwriting gestures. This novel approach was used in the accelerometer-based gesture recognition model proposed by the authors.

Hilbert transform is a linear transformation used in signal processing, that accepts as input a temporal signal, and produces an analytic signal. The analytic signal consists of a real and an imaginary part as explained in [13], where the negative half of the frequency spectrum is zeroed out. The real part represents the input signal and the imaginary part (y) is representative of the Hilbert transformed signal, which is phase shifted by $\pm 90°$. For an input signal ($x(t)$), the analytic signal ($x_a(t)$) after Hilbert transform is given by (5).

$$\begin{aligned} x_a = F^{-1}(F(x)2U) = x + iy, \\ where,\ F - Fourier\ transform, \\ U -\ Unit\ step\ function, \\ y - Hilbert\ transform\ of\ x. \end{aligned} \tag{5}$$

This approach of using Hilbert Transform for feature extraction is applied across all gesture samples, and the features - Mean, Skew, Minimum, Maximum and Energy are calculated as shown in Table III. The features Mean and Skew, which have been used in TF, and Energy which is calculated in FF are also used as Hilbert features since reusing them here yields a better performance for classifying different gestures. These Hilbert transformed features along each axis on a dataset (G) make up the HF vector, which is given by Equation 6.

$$HF = \bigcup_{j=1}^{5} \{H_x^j,\ H_y^j,\ H_z^j\} \tag{6}$$

The FeatureSet (FS) for a single dataset (G) is formed from (3), (4) and (6) by appending all the feature vectors - TF, FF and HF calculated across a dataset (G), as shown in (7).

$$FS = \bigcup_{j=1}^{N_{GS}} \{TF \cup FF \cup HF\}^j \tag{7}$$

VII. END-USER MODELLING

This paper enables the end-user to select any one of three proposed modes of operation - User Dependent, Mixed User and User Independent. The User Dependent (U_D) mode is an estimator of how well the system performs when the train-test split is between the gestures of a single user. Mixed User (U_M) is representative of the complete set of gestures of all participants. The User Independent (U_I) mode employs a stratified k-fold cross validation technique which corresponds to training on a number of users and testing on the rest.

VIII. EXPERIMENT

The three modes U_D, U_M and U_I, as explained in Section VII, were first trained and tested on an Intel Core i5 CPU @ 2.20 GHz, operating on Ubuntu 16.04. The proposed system was initially implemented using seven classification algorithms that were identified from [14], [15] and [16] – Extremely Randomized Trees (Extra Trees), Random Forests, Gradient Boosting, Bagging, Decision Trees, Naive Bayes and Ridge Classifier.

Upon further analysis, the seven classifiers were found to run on a low-cost Raspberry Pi Zero, and the accuracies and time taken for a gesture sample to be classified are catalogued for the three modes across both datasets, as shown in Table V. From this set of seven classifiers, the Extra Trees (ET), Gradient Boosting (GB) and Ridge Classifier (RC) were chosen, based on the inferences from Section IX-A. Each of the three modes is evaluated using the three classifiers individually, and the efficiencies are noted and analyzed for both the datasets D_u and D_S in Sections IX-B and IX-C, respectively.

IX. RESULTS

A. Evaluation of Classifiers

The classifiers chosen in Section VIII were further analyzed for each dataset based on their computational characteristics.

TABLE V. AVERAGE ACCURACIES (ACC) FOR THE DATASETS D_u AND D_S AND TIME TAKEN (IN SECONDS) FOR CLASSIFICATION OF A SINGLE GESTURE SAMPLE ACROSS ALL MODES; GREEN: HIGHEST EFFICIENCIES IN ALL MODES, YELLOW: LEAST CLASSIFICATION TIMES TAKEN FOR A GESTURE SAMPLE IN ALL MODES

	uWave (D_u)						Sony (D_s)					
Classifier\Mode	User Dependent (U_D)		Mixed User (U_M)		User Independent (U_D)		User Dependent (U_D)		Mixed User (U_M)		User Independent (U_D)	
	Acc	Time	Acc	Time	Acc	Time	Acc	Time	Acc	Time	Acc	Time
Extra Trees	97.76	0.6287	97.85	0.6873	82.49	0.6853	95.88	0.6038	98.63	0.6538	75.1	0.669
Random Forest	97.41	0.6991	95.45	0.73	77.91	0.6995	95.25	0.6887	97.13	0.7041	70.13	0.686
Gradient Boosting	93.75	0.0072	94.38	0.0076	75.64	0.0078	90.5	0.0055	95.5	0.0057	66.41	0.0057
Bagging	92.74	0.1526	94.19	0.1527	76.64	0.1528	93.5	0.1673	93.37	0.1527	62.44	0.1529
Decision Trees	89.55	0.0015	84.11	0.0053	66.73	0.0015	84.13	0.0014	86.25	0.0051	50.9	0.0015
Naive Bayes	91.16	0.0273	71.96	0.0292	64.66	0.0273	91.38	0.0118	65.5	0.0116	54.35	0.0117
Ridge Classifier	97.5	0.0013	83.84	0.0013	74.64	0.0013	94.13	0.0013	77.75	0.0014	61.59	0.0013

It can be inferred that ET yields the highest accuracy in all modes for both datasets D_u and D_S, among all the three chosen classifiers. Apart from being more accurate than Random Forest, ET is also computationally less expensive as stated in [17]. It can also observed that the time taken for classification of a gesture sample is the least in RC. GB, whilst yielding a significantly lower classification time than ET, provides reasonable accuracies and hence was chosen.

Based on the users' preferences, a trade-off can be made between efficiency and classification time of a gesture sample among the three classifiers. To evaluate the performance of this model in all the modes across both datasets, evaluation metrics - cross-validation scores and confusion matrices were taken for the results obtained from the ET classifier, as it has the highest efficiency and an acceptable classification time.

TABLE VI. ACCURACIES OF 8 USERS FOR BOTH DATASETS D_u AND D_S IN USER DEPENDENT MODE (U_D) USING EXTRA TREES CLASSIFIER

Data\User	1	2	3	4	5	6	7	8	Avg
D_u	96.42	93.5	98.57	97.14	99.28	99.28	97.88	100	97.76
D_S	98	92	94	93	98	100	95	97	95.88

B. uWave Dataset Analysis

In the U_D mode on D_u, each user's data was divided into a randomized 75%-25% train-test split across all 8 gestures. The average recognition rate achieved over all 8 users was found to be 97.76%, using ET. The individual users' efficiencies for the same can be observed in Table VI.

The classification model implemented using U_M mode on D_u with a randomized 75%-25% train-test split, provides an accuracy of 97.85%, and its confusion matrix can be showcased in Table VII.

TABLE VII. CONFUSION MATRIX FOR U_M MODE ON D_u; BROWN: GESTURE SIGNED, YELLOW: GESTURE CLASSIFIED, GREEN: CORRECT CLASSIFICATIONS, RED: INCORRECT CLASSIFICATIONS

	1	2	3	4	5	6	7	8
1	99.29	0.71	0	0	0	0	0	0
2	0.71	99.29	0	0	0	0	0	0
3	0.71	0	97.86	1.43	0	0	0	0
4	0	0	2.14	97.86	0	0	0	0
5	0	0	0	0	97.14	2.86	0	0
6	0	0	0	0	0	100	0	0
7	1.43	1.43	0	0	0	0	95	3.57
8	0	0.71	0	0	0	0	2.86	96.43

It was seen that on applying U_I mode on D_u, an average efficiency (average Leave-One-User-Out Cross-Validation score) of 82.49% was observed, while [4] has achieved 75.4%. The confusion matrix for the last user as test, trained upon the first seven users, which yields an accuracy of 92.14% has been illustrated in Table VIII.

TABLE VIII. CONFUSION MATRIX FOR U_I MODE ON D_u; BROWN: GESTURE SIGNED, YELLOW: GESTURE CLASSIFIED, GREEN: CORRECT CLASSIFICATIONS, RED: INCORRECT CLASSIFICATIONS

	1	2	3	4	5	6	7	8
1	95.71	4.29	0	0	0	0	0	0
2	0	100	0	0	0	0	0	0
3	0	0	82.86	17.14	0	0	0	0
4	0	0	0	100	0	0	0	0
5	0	0	0	0	100	0	0	0
6	0	4.29	0	0	0	95.71	0	0
7	0	1.43	0	0	0	0	70	28.57
8	0	1.43	0	0	0	0	5.71	92.86

Table IX shows that RC has the least classification time along with an accuracy of 97.5%, which is marginally lesser than ET, thereby effectively capturing the similarities between gesture samples signed by the same user, i.e, in U_D mode. GB provided accuracies and computational times which are intermediary between ET and RC across all three modes, thereby giving the user a choice to reduce classification time by a significant margin, while having a reasonable accuracy.

TABLE IX. ACCURACIES (ACC) AND TIME TAKEN (IN SECONDS) FOR CLASSIFICATION OF A SINGLE GESTURE SAMPLE FOR D_u USING THE 3 SELECTED CLASSIFIERS ACROSS ALL MODES

Classifier\Mode	U_D		U_M		U_I	
	Acc	Time	Acc	Time	Acc	Time
ET	97.76	0.6287	97.85	0.6873	82.49	0.6853
GB	93.75	0.0072	94.38	0.0076	75.64	0.0078
RC	97.5	0.0013	83.84	0.0013	74.64	0.0013

C. Sony Dataset Analysis

For all eight participants, the average accuracy for D_S in U_D mode was observed to be 95.88% using ET, where each user's data was divided into a 75%-25% train-test split in random, as done in D_u. The efficiencies across all individual users can be observed in Table VI.

TABLE X. ACCURACIES (ACC) AND TIME TAKEN (IN SECONDS) FOR CLASSIFICATION OF A SINGLE GESTURE SAMPLE FOR D_S USING THE 3 SELECTED CLASSIFIERS ACROSS ALL MODES

Classifier\Mode	U_D		U_M		U_I	
	Acc	Time	Acc	Time	Acc	Time
ET	95.88	0.6038	98.63	0.6538	75.1	0.669
GB	90.5	0.0055	95.5	0.0057	66.41	0.0057
RC	94.13	0.0013	77.75	0.0014	61.59	0.0013

In the U_M mode, the proposed model with a 75%-25% train-test split at random on D_S, yields an efficiency of 98.625%. Table II shows the confusion matrix for the U_M mode across all 20 gestures.

TABLE XI. BEST ACCURACIES (ACC) AND LEAST TIMES TAKEN (IN SECONDS) FOR CLASSIFICATION OF A SINGLE GESTURE SAMPLE ACROSS ALL MODES FOR BOTH DATASETS D_u AND D_S

Dataset\Mode	U_D		U_M		U_I	
	Acc (ET)	Time (RC)	Acc (ET)	Time (RC)	Acc (ET)	Time (RC)
D_u	97.76	0.0013	97.85	0.0013	82.49	0.0013
D_S	95.88	0.0013	98.63	0.0014	75.1	0.0013

An average 8-fold cross validation score (average recognition rate across all users) of 75.093% was observed in the U_I mode of D_S. The confusion matrix trained upon the first seven users with the last user as test, which yields an accuracy of 91.75%, is shown in Table IV.

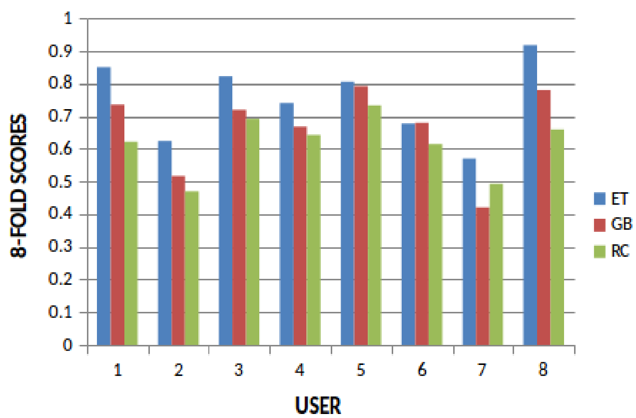

Fig. 3. 8-fold Cross-validation Scores for all Users in D_S.

Fig. 3 shows the behavior of all users as test (Leave-One-User-Out Cross-Validation scores) in the U_I mode on D_S. The large number of misclassifications in the second and seventh user is indicative of both users having not performed the gestures in a manner similar to the rest, thereby reducing the overall efficiency across all classifiers.

Owing to the generic FeatureSet (FS) implemented, the results observed in Table X for D_S are analogous to the results observed in D_u for all three chosen classifiers across all the three modes.

X. CONCLUSION

A Gesture Recognition system with the capability to operate in any of the three modes – User Dependent, Mixed User and User Independent, with a set of generic features was designed, validated and tested in this paper. The proposed system was tested on two public accelerometer-based gesture datasets – uWave and Sony. Table XI showcases the best classifier for each category - Efficiency and Classification Time for a gesture sample, across all three modes of both the datasets. As can be seen in Table XI, Extremely Randomized Trees was observed to perform the best in terms of accuracy, while Ridge Classifier provided the least classification time which was around 500 times faster than ET for a gesture sample, irrespective of the mode or dataset.

The end-users are given the flexibility to choose any combination of modes and classifiers according to their requirements. This system was implemented on a Raspberry Pi Zero priced at 5 USD making it a low-cost alternative.

ACKNOWLEDGMENT

The authors would like to thank Solarillion Foundation for its support and funding of the research work carried out.

REFERENCES

[1] Q. Chen, N. D. Georganas and E. M. Petriu, "Real-time Vision-based Hand Gesture Recognition Using Haar-like Features," 2007 IEEE Instrumentation & Measurement Technology Conference IMTC 2007, Warsaw, 2007, pp. 1-6.

[2] K. K. Biswas and S. K. Basu, "Gesture recognition using Microsoft Kinect," The 5th International Conference on Automation, Robotics and Applications, Wellington, 2011, pp. 100-103.

[3] S. Mitra and T. Acharya, "Gesture Recognition: A Survey," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 37, no. 3, pp. 311-324, May 2007.

[4] J. Liu, Z. Wang, L. Zhong, J. Wickramasuriya and V. Vasudevan, "uWave: Accelerometer-based personalized gesture recognition and its applications," 2009 IEEE International Conference on Pervasive Computing and Communications, Galveston, TX, 2009, pp. 1-9.

[5] SmartWatch Gestures Dataset, Technologies of Vision, Fondazione Bruno Kessler, [Online]
Available: https://tev.fbk.eu/technologies/smartwatch-gestures-dataset

[6] T.G. Zimmerman, J. Lanier, C. Blanchard, S. Bryson, and Y. Harvill, "Hand Gesture Interface Device," Proc SIGCHI/GI Conf. Human Factors in Computing Systems and Graphics Interface, pp 189-192, 1986.

[7] Shah Muhammed Abid Hussain and A. B. M. Harun-ur Rashid, "User independent hand gesture recognition by accelerated DTW," 2012 International Conference on Informatics, Electronics & Vision (ICIEV), Dhaka, 2012, pp. 1033-1037.

[8] T. Pylvninen, "Accelerometer Based Gesture Recognition Using Continuous HMMs.," in IbPRIA (1), 2005, vol. 3522, pp. 639646. 492

[9] N. Helmi and M. Helmi, "Applying a neuro-fuzzy classifier for gesture-based control using a single wrist-mounted accelerometer," 2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation - (CIRA), Daejeon, 2009, pp. 216-221.

[10] K. Altun, B. Barshan, and O. Tunel, "Comparative study on classifying human activities with miniature inertial and magnetic sensors.," Pattern Recognition, vol. 43, no. 10, pp. 36053620, 2010.

[11] L. Jing, Y. Zhou, Z. Cheng, and J. Wang, "A Recognition Method for One-Stroke Finger Gestures Using a MEMS 3D Accelerometer.," IEICE Transactions, vol. 94D, no. 5, pp. 10621072, 2011.

[12] J. Wu, G. Pan, D. Zhang, G. Qi, and S. Li, "Gesture Recognition with a 3-D Accelerometer.," in UIC, 2009, vol. 5585, pp. 2538.

[13] H. Ishida, T. Takahashi, I. Ide, and H. Murase, "A Hilbert warping method for handwriting gesture recognition.," Pattern Recognition, vol. 43, no. 8, pp. 27992806, 2010.

[14] L. Tencer, M. Reznkov, and M. Cheriet, "Evaluation of techniques for signature classification from accelerometer and gyroscope data.," in ICDAR, 2015, pp. 10661070.

[15] W. Aswolinskiy, R.F. Reinhart, and J.J. Steil. "Impact of regularization on the model space for time series classification." In Machine Learning Reports, pages 4956, 2015.

[16] Y. Yang and Y. Yu, "A hand gestures recognition approach combined attribute bagging with symmetrical uncertainty," 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, Sichuan, 2012, pp. 2551-2554.

[17] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Machine Learning, vol. 63, no. 1, pp. 342, Jul. 2006.