Automatic Classification of Physical Exercises from Wearable Sensors using Small Dataset from Non-Laboratory Settings

Alok Kumar Chowdhury, Aleksandr Farseev, Prithwi Raj Chakraborty, Dian Tjondronegoro, *Sr. Member, IEEE*, and Vinod Chandran, *Sr. Member, IEEE*

Abstract— **Effective classification of physical exercises allows individuals to assess their levels of physical activity and functional ability for maintaining physical fitness and help reduce risks of chronic diseases. This paper investigates and compares classification techniques for detecting physical exercise in real-world contexts that often only supports a small training dataset. The system combines heart rate with other exercise-related features, such as distance, duration, calories, etc. The experiment uses a dataset of 40 realistic (uncontrolled) sessions from 22 individuals wearing wearable sensors while performing different exercises, including walking, aerobics, running, indoor cycling, and weight training. Based on a 5-fold cross validation approach, AdaBoost demonstrated the highest (87.25%) classification accuracy compared to other classifiers, including support vector machine, neural network, and binary decision tree when used individually. When fused together at the decision level using majority-voting techniques, these classifiers achieved higher accuracy (89.25%) than that of individual applications.**

I. Introduction

While the recent advent of technologies makes people's life easier and enjoyable, it also changes the lifestyle of people to sedentary - which is now considered as a health risk for people in both developed and developing countries. On the contrary, physical exercise has a good impact on people's well-being and World Health Organization (WHO) recommends adults of age 18-64 to perform at least 150 minutes of moderate-intensity or 75 minutes of vigorous-intensity physical exercise weekly. Also, patients with certain diseases may require following certain well-defined exercises. Hence, Automatic recognition of physical activities can be useful for people to learn their lifestyle and set long-term goals.

Automatic human activity recognition research has been approached in two different ways – one approach is video observation based and another is based on wearable sensor data. Despite video observation based techniques [1] have remarkable success in this area, users' may become unwilling to be constantly kept under a video surveillance. Activity recognition based on wearable sensors has gained much attention due to the development of sensor technologies, which increasingly become more efficient and wearable.

Human activity recognition can utilize multi-modal sensors data, such as accelerometer, physiological, location and environment related data [2]. The increasing use of non-invasive and pervasive sensors is expected to be more capable of capturing realistic data [3], compared to previous work that requires users to wear several accelerometers in different positions of their body [4] or recording devices that are uncomfortable in real-time [5]. For example, Berchtold et al. [6] presented human activity detection system using only a mobile phone's built-in tri-axial accelerometer. In addition to accelerometer, physiological sensors data (e.g. heart-rate) could potentially reflect on a person's internal body states, therefore hypothetically should have a greater role in human activity recognition. Most existing work that uses physiological cues along with accelerometer has not been able to accurately correlate these data with activity [4, 5]. Some works have recommended that physiological sensor data or vital sign can be exploited for better recognition accuracy [7-9]. In this study, we use data from multi-sensors that capture both physiological and contextual information.

Designing machine learning based classifiers for human activity detection has been previously investigated in some recent work. Some studies showed decision tree classifiers as the best performer among base- and meta-level classification techniques [8], while some has found that boosting learning algorithm can outperform decision tree classifiers [10]. Single accelerometer based activity detection using decision tree, neural networks and support vector machine classification based techniques have reported high performance [11].

An optimal fusion of multiple classifiers is expected to improve the overall classification performance, especially when each classifier has good performance and the range of the classifiers have sufficient diversity. However, the appropriate selection of classifiers and the performances of individual and fused classifiers are yet to be explored in real world context of physical exercise classification which may provide small dataset.

To sum up, the existing challenges for accurately detecting activity using sensors data are: 1) data collection using non-invasive sensors; 2) feature extraction and classification, especially on smaller dataset; 3) less accuracy in realistic or uncontrolled settings compared to lab based experiments and 4) efficient multi-source multi-modal data fusion [9]. This paper will present results from a study to investigate the feasibility of four state-of-the-art classification techniques - including AdaBoost, support vector machine, neural network, binary decision tree, and their fusion - for effective recognition of realistic (non-laboratory based) physical exercise activities, including walking, aerobics,

Alok Kumar Chowdhury, Prithwi Raj Chakraborty, and Vinod Chandran are with Queensland University of Technology, Science and Engineering Faculty, Brisbane, Australia. (email: alok.chowdhury@qut.edu.au, p1.chakraborty@qut.edu.au, vinod.chandran@ieee.org)

Aleksandr Farseev is with National University of Singapore, LMS lab, School of computing, Singapore. (e-mail: farseev@u.nus.edu).

Dian Tjondronegoro is with Southern Cross University, School of Business and Tourism, Australia. (e-mail: dian.tjondronegoro@scu.edu.au).

running, indoor cycling, and weight training. The experiments use data from 40 exercise sessions of 22 individuals wearing non-invasive and wearable sensors, which capture physiological, location and environment related data. The proposed framework is based on the combination of multiple classifiers and their fusion using majority voting for activity classification.

II. METHOD

This paper focuses on the use of several classification techniques to recognize physical exercises, comparing the effectiveness of using individual or fused classifiers at the decision level. Figure 1 depicts the proposed framework.

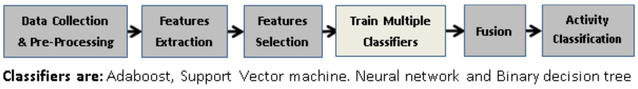
Data Collection & Pre-Processing
Features Extraction
Features Selection
Train Multiple Classifiers
Fusion
Activity Classification
Classifiers are: Adaboost, Support Vector machine, Neural network and Binary decision tree

Figure 1. Flow diagram of the proposed framework which was developed using Matlab Simulation Tool

A. Data Collection and Pre-processing

We've utilized the subset of the NUS-SENSE [12] dataset, which was provided in fully-anonymized form by NExT Research Center[1] during the Web Science & Big Data Analytics Sumer School 2014[2]. The data was collected from Endomondo[3] user profiles who has publicly released their workouts on Twitter in the time interval between Nov 10, 2014 and Nov 20, 2014. The above suggests the non-applicability of the code of ethics to this research. The physiological sensor data was collected from wearable devices, while the location data was obtained from GPS-enabled mobile phones. The weather information was gathered from AccuWeather weather forecast service. Table I shows the number of collected activities, environment type (indoor/outdoor) and the activity types. In total, 40 activities of 22 individuals from different demographic groups (males and females from 20 to 40 years old) were collected. The duration of each measurement was from half an hour to two hours.

TABLE I. TOTAL NUMBER OF USERS PER ACTIVITY

Activity	Environment Type	Total Activities
Walking	Outdoor	4
Aerobics	Indoor	3
Running	Outdoor	20
Indoor Cycling	Indoor	3
Weight Training	Indoor	10

In the pre-processing step, unique-value imputation [13] was applied on the dataset where each missing values were replaced by an arbitrary unique value (0). Jittering noise presented in heart-rate data was scaled down by a using moving-average filter of span 4 samples.

B. Features Extraction

The goal of feature extraction is to find derived values from the raw data intended to be non-redundant, informative and also usable for machine learning. This work identified total fifteen (15) features from the raw data and six (6) out of them were derived from heart rate.

A Physiological signal such as heart-rate features is a good indicator for physical activity detection [8]. We computed some overall time domain features like mean and standard deviation for each activity using all heart rate samples for that activity. Adopting a similar approach [14], for each activity, our work also calculated several time-varying features of heart-rate such as derivatives, gradient, energy, and variance. Obtained heart rate data was time stamped in beats-per-minute and a sliding window (with a span of 9 seconds) was used to compute the above temporal features. For computing derivative, we divided the total absolute difference between the neighboring heart rate samples by the difference of time stamped with first and last samples within the sliding window. The variance was computed by taking statistical variance of the heart rate samples with the sliding window using standard MATLAB function. We computed a series of directional values for increasing order of the heart rate sample (within the sliding window) and took the average of those directional values as gradient features. Energy from low frequency band (0.04-0.15 Hz) was computed passing all the heart rate samples through a band-pass filter. Finally, the average of each of these time-varying features is taken as features for classification. Apart from heart-rate features, the other location and environment related granular features were gathered: distance travelled, duration of exercise, calories spent, average speed, amount of hydration, minimum altitude, maximum altitude, wind speed, and weather type (e.g. sunny, cloudy, partly cloudy, night etc.). To make the feature set appropriate for training and classification, numeric unique values were assigned for the features which had non-numeric values. For example, we used numeric values 1, 2, 3, 4, 5, 6, 7 and 8 for weather types 'Flurries', 'Intermittent clouds night', 'Partly cloudy night', 'Clear night', 'Cloudy', 'Mostly Cloudy', 'Partly sunny', and 'Sunny' respectively.

C. Feature Selection

Not all extracted features were selected for classification; correlation of features with the exercises was investigated to find good candidate features. Features, having the absolute correlation coefficient greater than 0.15, were selected for classification.

Due to collecting in the real context, some features contained a large number of missing values. As too much imputation on a single feature may lead to biased classification accuracy, we removed those features from the final feature set. For example, average speed feature of our data had over 80% missing values and not selected for classification. After feature selection process, following twelve (12) features were finally chosen for classification.

- **Heart rate features:** mean, standard deviation, derivatives, gradient, energy, and variance.
- **Other features:** distance travelled, duration of exercise, calories spend, amount of hydration, minimum altitude, and maximum altitude.

[1] next.comp.nus.edu.sg
[2] webscience.org/2014/10/16/web-sci-big-data-analytics-summer-school
[3] www.endomondo.com

D. Training & Classification

This study used the static selection of classifiers, where four widely used classification techniques including AdaBoost, Support vector machine (SVM), Neural network (NN), Binary decision tree (BDT) were evaluated; and also fused them together for physical exercise classification task. These classifiers are considered as state-of-the-art classifiers, which improves the diversity and in turn also the accuracy of the classification task.

Human activity recognition works in supervised fashion because such system should return a label. Let the training data is T= {(x_1, y_1), (x_2, y_2), …, (x_m, y_m). Where each x_i is the set of features and y_i is the corresponding label for each feature set for training. In our case, each x_i is the set of twelve features and $y_i \in \{1,2,3,4,5\}$, where 1 to 5 are numeric class labels for physical exercises: walking, aerobics, running, indoor cycling, and weight training respectively. The classification problem for physical activity or exercise is essentially a multi-class classification problem.

AdaBoost -We used AdaBoost.M2, multi-class classifier, which combines the outputs of several weak learners into a weighted sum and also tweaks subsequent weak learners based on the misclassification by the previous classifiers. AdaBoost is chosen because it often referred as best out-of-the-box classifiers and also shows less prone to the overfitting problem. We investigated 20, 100, 200, 1000 cycles as learning cycles of AdaBoost, but finally set it to 200 cycles which was optimum in our case considering both classification accuracy and processing time.

SVM - Support vector machine was used to classify the five classes effectively. Generally, SVM classifiers can find the best separator between the two classes only. The two-class SVM was adapted during this research to enable multi classification. Adapted SVM function works in a fashion that firstly it classifies one class against all other classes and then it classifies another classes verses remaining classes and so on. Linear classification function was chosen as the kernel function. Before applying SVM classification techniques, we applied principle component analysis (PCA) on the features (x) and a set of linearly uncorrelated variables (x'), i.e. principle components, were obtained. Then, instead of using actual features (x), the principle components (x') were used for training and testing – which led to better classification performance for SVM.

NN- A feed-forward neural network was used in this research. Neural networks have been widely used in the field of computer vision and speech processing and can solve many problems that are difficult to solve using heuristic rules – which lead us to choose it as a classifier in our problem. The inputs of the NN classifier were 12 features and outputs were 5 physical exercises. Hence, the neural network consists of 12 neurons in the input layer and 5 neurons in the output layer. The number of neurons in the hidden layer was set to 5 neurons. Using the backpropagation training algorithm, the networks were trained repeatedly until the root mean square error reach to an optimum level.

BDT - Binary decision tree uses hierarchical approaches where it divides the problem of activity recognition into smaller sub- problems. During training, an optimum decision tree was generated from the training dataset using the cross-validation within the training data. The obtained decision tree was then used for classification tasks.

E. Fusion of classifiers

After training and testing using all four classifiers individually, we also combined them to achieve the combined accuracy. We applied Majority Voting Algorithm which is simplest but effective [15] to fuse the classifier's results. The class, which is mostly identified by the individual classifiers for a test data, is selected as the final output. If there is a tie among the classifiers then our fusion chose the output of AdaBoost as the final class because of its better performance.

III. EXPERIMENTAL RESULTS

Classification results of *n*-class classification problem can be organized in a confusion matrix M_{nxn} where each cell (M_{ij}) is the numbers of instances of class *i* that were actually classified as class *j*. From the confusion matrix the value of true-positive, true-negative, false-positive and false-negative can be calculated. Then, we used classification accuracy as a performance metric, which can be given by the following equation.

$$Classification\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100\%$$

During the train and test accuracy, the 5-fold cross validation was carried out. In 5-fold cross-validation, the original sample is randomly partitioned into 5 equal size sets. Then, a single set is used for testing and all other sets are used for training. Thus every set is used exactly once for testing. Finally, the performance metrics of each fold are averaged to produce a single estimated performance metric for the classifier. We applied random sampling of training sets due to small dataset and uneven distribution of available data for each class of activity. This is a common problem in real-world data collection that often cannot acquire a uniform number of large samples per activity.

Table II, shows the classification accuracy of each fold under different classification techniques. Out of the four classifiers, AdaBoost provides the best classification accuracy (87.25%) and worst result is given by neural network (76.25%). Binary decision tree also shows outstanding 84.25% classification accuracy, which is almost close to AdaBoost. SVM provides 77% classification accuracy. We also tried to combine all four classifiers together using simple majority voting algorithm to learn the combined result. Classification accuracy of combined classifiers was 89.25%, which outperforms the performances of all individual classifiers. We also removed NN from fusion but it brings fusion performance down to 86.75%. Overall, Table II demonstrates that based on the average scores, fusion of classifiers has outperformed any of the single classifier.

Additionally, we also compute the precision and recall to compute the F1-Score, reported in Table III. This was used to understand the performances of classifiers for each class. The results showed that combined classifiers (fusion) showed better F1-score for most classes compared to that of others. Among the single classifiers, AdaBoost performed well in the case of small and uneven dataset compared to the other three classifiers. Neural network cannot classify when very few numbers of training samples were available, which suggest that the minimum viable training sample for activity classification is 20 to avoid either over- or under- fitting. This is further supported with the patterns that can be observed across all classifiers: running (20 samples) consistently showed a higher F1-score compared to weight training (10 samples).

Across all classifiers, walking has the lowest performance compared to the other activities as it may lack of distinguishable features in the small dataset. On the other hand, aerobics, running and indoor cycling all have high performance (particularly compared to weight training) when AdaBoost, SVM and BDT was used. This suggests that future data collection needs to maintain a more distributed balance of samples to avoid over-fit or under-fit.

TABLE II. CLASSIFICATION ACCURACY OF EACH FOLD-SET

Set #	Classification Accuracy (%)				
	AdaBoost	**SVM**	**NN**	**BDT**	**Fusion**
#1	80.00	75.00	87.50	81.25	87.50
#2	87.50	75.00	68.75	81.25	87.50
#3	87.50	85.00	81.25	90.00	90.00
#4	87.50	68.75	75.00	87.50	87.50
#5	93.75	81.25	68.75	81.25	93.75
Total (avg)	**87.25**	**77.00**	**76.25**	**84.25**	**89.25**

TABLE III. F1 SCORE FOR EACH CLASS UNDER DIFFERENT CLASSIFICATION

Class/ Physical Exercise	F1 Score				
	AdaBoost	**SVM**	**NN**	**BDT**	**Fusion**
Walking	0.29	0.18	0	0	**0.23**
Aerobics	1.00	0.83	0	0.83	**1.00**
Running	0.82	0.68	0.75	0.80	**0.91**
Indoor Cycling	0.53	0.50	0	0.29	**0.75**
Weight Training	0.49	0.40	0.36	0.59	**0.73**

IV. CONCLUSION & FUTURE WORK

This study used realistic sensor data of 40 uncontrolled sessions from 22 individuals. Dataset was small in size and also contained missing values due to collecting in real contexts. After applying pre-processing on the raw data, we used total 12 features for classification models training where 6 features were extracted from the heart-rate signal. Then, we investigated four classification techniques on the features and also combine the classifier's output to effectively recognize physical activities. The results of the proposed exercise classification, after applying 5-fold cross validation, were encouraging as the system has successfully classified most of the physical exercises despite having a small training dataset. AdaBoost showed better classification accuracy compared to other classifier, whereas fusion of multiple classifiers using majority voting has outperformed the performances of individual classifiers.

This paper has shown reliable classification of exercises from small-sized realistic sample. Our future research will investigate the most effective features and sensors for classifying a wide variety of physical activities, which will benefit from both physiological and contextual data. We will also collect more wearable sensor data by involving individuals from different groups (e.g. age, gender) and apply our proposed multi-classification techniques to validate it on a larger dataset [12].

REFERENCES

[1] J. Candamo, M. Shreve, D. B. Goldgof, D. B. Sapper, and R. Kasturi, "Understanding transit scenes: A survey on human behavior-recognition algorithms," *Intelligent Transportation Systems, IEEE Transactions on,* vol. 11, no. 1, pp. 206-224, 2010.

[2] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *Communications Surveys & Tutorials, IEEE,* vol. 15, no. 3, pp. 1192-1209, 2013.

[3] G. Bieber, J. Voskamp, and B. Urban, "Activity recognition for everyday life on mobile phones," in *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments*: Springer, 2009, pp. 289-296.

[4] E. M. Tapia *et al.*, "Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor," in *Wearable Computers, 2007 11th IEEE International Symposium on*, 2007, pp. 37-40: IEEE.

[5] J. Pärkkä, M. Ermes, P. Korpipää, J. Mäntyjärvi, J. Peltola, and I. Korhonen, "Activity classification using realistic data from wearable sensors," *Information Technology in Biomedicine, IEEE Transactions on,* vol. 10, no. 1, pp. 119-128, 2006.

[6] M. Berchtold, M. Budde, D. Gordon, H. R. Schmidtke, and M. Beigl, "Actiserv: Activity recognition service for mobile phones," in *Wearable Computers (ISWC), 2010 International Symposium on*, 2010, pp. 1-8: IEEE.

[7] O. D. Lara, A. J. Pérez, M. A. Labrador, and J. D. Posada, "Centinela: A human activity recognition system based on acceleration and vital sign data," *Pervasive and mobile computing,* vol. 8, no. 5, pp. 717-729, 2012.

[8] A. Reiss and D. Stricker, "Introducing a modular activity monitoring system," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, 2011, pp. 5621-5624: IEEE.

[9] A. Farseev and T.-S. Chua, "Tweet can be fit: Integrating data from wearable sensors and multiple social networks for wellness profile learning," *ACM Transactions on Information Systems (TOIS),* vol. 35, no. 4, p. 42, 2017.

[10] A. Reiss, G. Hendeby, and D. Stricker, "A competitive approach for human activity recognition on smartphones," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2013), 24-26 April, Bruges, Belgium*, 2013, pp. 455-460: ESANN.

[11] I. C. Gyllensten and A. G. Bonomi, "Identifying types of physical activity with a single accelerometer: evaluating laboratory-trained algorithms in daily life," *Biomedical Engineering, IEEE Transactions on,* vol. 58, no. 9, pp. 2656-2663, 2011.

[12] A. Farseev and T.-S. Chua, "TweetFit: Fusing Multiple Social Media and Sensor Data for Wellness Profile Learning," in *AAAI*, 2017, pp. 95-101.

[13] M. Saar-Tsechansky and F. Provost, "Handling missing values when applying classification models," *Journal of Machine Learning Research,* vol. 8, pp. 1625–1657, 2007.

[14] P. R. Chakraborty, L. Zhang, D. Tjondronegoro, and V. Chandran, "Using Viewer's Facial Expression and Heart Rate for Sports Video Highlights Detection," in *Proceedings of International Conference on Multimedia Retrieval* Shanghai, China, 2015: ACM.

[15] L. Lam and C. Y. Suen, "Application of majority voting to pattern recognition: an analysis of its behavior and performance," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on,* vol. 27, no. 5, pp. 553-568, 1997.