A Feature Set Evaluation for Activity Recognition with Body-Worn Inertial Sensors

Syed Agha Muhammad[1], Bernd Niklas Klein[2], Kristof Van Laerhoven[1], and Klaus David[2]

[1] TU Darmstadt, Darmstadt, Germany
{muhammad,kristof}@ess.tu-darmstadt.de
[2] University of Kassel, Kassel, Germany
{niklas.klein,klaus.david}@comtec.eecs.uni-kassel.de

Abstract. The automatic and unobtrusive identification of user activities is a challenging goal in human behavior analysis. The physical activity that a user exhibits can be used as contextual data, which can inform applications that reside in public spaces. In this paper, we focus on wearable inertial sensors to recognize physical activities. Feature set evaluation for 5 typical activities is performed by measuring accuracy for combinations of 6 often-used features on a set of 11 well-known classifiers. To verify significance of this analysis, a t-test evaluation was performed for every combination of these feature subsets. We identify an easy-to-compute feature set, which has given us significant results and at the same time utilizes a minimum of resources.

1 Introduction

Physical activity can be defined as "any bodily moment produced by skeletal muscles that result in energy expenditure" [1]. For decades, activity recognition has been a topic of fundamental interest for practitioners and scientists. The need to know how humans behave and act in different situations and contexts has been suggested to be of use for doctors, psychologists and good health professionals [12]. Activity recognition systems could play an important role in ubiquitous computing scenarios. Providing services to the users based on their location or activity is an active research area. A well-designed system in a public space benefits from an understanding of its users' states and their environment.

In addition to the context aware systems, activity recognition systems found their relevance in the field of health-care related applications [14]. In principle, activity recognition can be used for social benefit, especially in human-centric applications such as elderylcare and health-care. Activity recognition systems also can assist people to stay physically fit and maintain their health by designing assessment tools [13].

In this paper, we focus on the feature set evaluation on inertial data for such activity recognition. Different feature sets are created of the time domain

features. In a first evaluation, they are used with popular machine learning algorithms to find their impact on accuracy. Normally, the importance of features depends upon the activity to be detected. For activities such as running, walking, climbing, descending and related gait, frequency information from 3-axis accelerometer is important. Different feature subsets are evaluated to find their accuracy and significance. We focus on easy-to-compute feature sets, which utilise minimal resources and produce results which are acceptable in terms of accuracy. At the end, we evaluate the results with a two-tailed t-test to find the significance of the feature subsets.

The remaining part of the paper is organised as follows: In Section II, we discus the related work. In Section III, we discus the experimental setup and methodology, which includes hardware devices, number of users, number of tests, extracted features, and classification classifiers used. In Section IV, evaluation results and the accuracies of the feature sets are discussed. We introduce a new feature "Meantilt", which was included into the feature sets, and focus on finding suitable features for activity recognition, which would provide us the highest accuracy results. We investigate also whether some features have a prominent role or certain features are negligible for activity recognition. In Section V, two-tailed t-test is discussed to find out whether the differences between the recognition results using all the features and with certain features removed are significant. In Section VI, the main conclusions of this paper are summed up.

2 Related Work

Human activity recognition is an important field and this fact has been acknowledged by the rich content of literature available in this area [2],[3],[4],[5],[6],[9],[7], [11]. Activities such as walking, standing, sitting, climbing stairs and descending stairs naturally impart themselves to recognition using acceleration sensors, since these activities are clearly defined by the motion and relative positions of the body parts. Small and cheap sensors can be easily integrated into accessories such in garments or mobile phones to recognize these activities as described in [2], [3], and [4].

In the past, some research has focused on feature selection in the field of activity recognition. The authors of [15] have a combination of discriminative and generative classifiers. With eight different sensors, 651 different features were extracted. The authors have used AdaBoost to automatically select the best features and to learn an ensemble of static classifiers to recognize different activities. Second, the classification margins from the static classifiers were used to compute the posterior probabilities, which are then used as inputs into HMM models to capture the temporal regularities and smoothness of activities.

Forward-Backward sequential search methods for feature selection has been suggested by [16], [17]. In [16] features such as mean, standard deviation, correlation (x, y axes), mean crossing, as well as heart rate mean were tested with forward-backward search, which is a well-known feature selection algorithm. With this procedure, a subset of best (giving the best classification result) fea-

tures can be determined for the final analysis. Later, a multilayer perception model and KNN were applied to classify the activity.

The authors of [9] propose to use a simple measure of cluster precision to evaluate the best features for discriminative activities. It shows, how the selection of different features can improve the recognition rates. Fast Fourier Transform (FFT) features had the highest cluster precision, with different components and window lengths required for different activities. The authors also conclude that variance has consistently high precision values acceleration data with most of the activities.

3 Experimental Setup and Methodology

This section presents the experimental setup, the feature extraction methods and the classifiers chain used in this paper.

3.1 Experimental Setup

Sun SPOT wireless sensor nodes [10] were used to perform the experiments discussed in this paper. After several experiments, we found that the position which suits our set of activities the best was the thigh. We performed ten experiments on six test subjects. Four out of six subjects participated twice in the test. The test data was taken for 350 minutes with each test covering an interval of 35 minutes. Each subject was requested to carry the Sun SPOT in the trousers' pocket and to perform the movements, which are sitting, standing, walking, climbing stairs, and descending stairs. Subjects were told to perform the sequence of activities but not specifically how to do them. The movements were annotated using the Nokia N800 tablet. This provided the annotation of the movements, which will be used as class information for the training and testing the classifiers. The recorded data and annotations were synchronized afterwords.

3.2 Feature Extraction

In order to detect activity information using classification algorithms, the raw data must be pre-processed to extract the more useful information. This process is called feature extraction. Popular computed time-domain features used in activity recognition include mean [5][6][7], variance or standard deviation [5][6], energy [5][6][7], entropy [5], and correlation between axes. In most of the cases, energy and entropy are calculated using the frequency domain [6] [7] [13].

In frequency domain features, first we have to transform the window of signal data into the frequency domain using the fourier transform. Normally, the output of FFT gives us the set of coefficients [6] which represents the distribution of the signal energy and amplitude of the frequency component of the signal. However, FFT requires multiple components to discriminate different activities. Hence it will increase computation and is not suitable for real time applications. As the

time domain features can be easily extracted in real time, they are more popular in many practical acceleration activity recognition systems [15].

The used features include mean, variance and standard deviation of the acceleration data. We have also included energy and entropy of the FFT. All combinations of possible feature subsets were used for evaluation. Apart from the above mentioned features, we have introduced another feature to our feature set, called Meantilt [3] along axes. It was calculated as shown in equation 1.

$$Meantilt_j = \sqrt{\left(\sum_{i=j}^{j+W} x_i\right)^2 + \left(\sum_{i=j}^{j+W} y_i\right)^2 + \left(\sum_{i=j}^{j+W} z_i\right)^2 / W} \qquad (1)$$

Where x, y and z represents the tilt along the respective axis, W stands for window length. Similarly, j indicates the window overlapping percentage, which is 50% for the window length of 32. Experiments have shown that our introduced features has produced better results, and at the same time it is easy to calculate.

Features were extracted from inertial data using a sliding window approach with a window size of 32 with 16 overlapping between the consecutive windows. Window overlapping with 50% have shown success in previous works [6][7][11]. At the sampling rate of 32Hz, each window represents data for 1 second. After labeled data was acquired, the features were extracted, to which the classification algorithms were applied.

3.3 Classification Chain

We have used the wide range of classification algorithms available in the Weka Toolkit[4]. The base-level classifiers are decision tree (DT), support vector machine (SVM) or sequential minimal optimization (SMO) in weka toolkit, K-nearest neighbours (KNN), and naive bayes (NB). So-called meta-level classifiers use bagging, boosting, and voting on the top of these classifiers. The classification was carried out using 10 fold cross validation. Figure 1 shows the data flow cycle of the complete process, with all features and classifiers.

4 Evaluation of the Feature Sets

Table 1 shows the accuracy results when certain features were removed from the full set of feature (S10), for all classifiers and feature sets [5]. Meta-level classifiers

[3] for the detailed discussion of tilt calculation from accelerometers please visit, http://www.sunspotworld.com/docs/AppNotes/AccelerometerAppNote.pdf, last visited at 28 July 2011.

[4] http://www.cs.waikato.ac.nz/ml/weka/, last visited at 27th July 2011.

[5] S1= all features; S2= mean, standard deviation, energy, and entropy; S3= mean, standard deviation, variance, and entropy; S4= mean, energy, and entropy; S5= mean, standard deviation, variance; S6= mean, standard deviation, and entropy; S7= mean, and variance; S8= mean, and energy; S9= mean, standard deviation, variance and Meantilt; S10= all features and Meantilt.

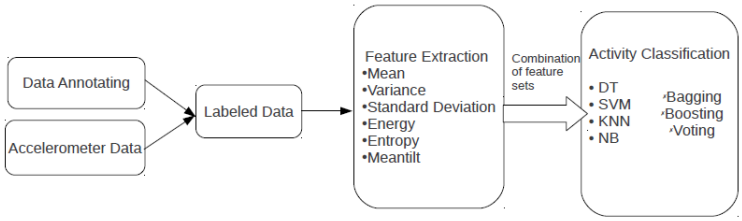

Fig. 1. Overview of how data was classified using a wide variety of feature sets and classifiers.

have the better overall accuracy compared to base-level classifiers, but they tend to be slower.

Algorithm	S1	S2	S3	S4	**S5**	S6	S7	S8	**S9**	**S10**
J48(DT)	83.28	84.27	84.75	82.5	**82.05**	82.73	81.71	81.01	**84.59**	**84.94**
Naive Bayes(NB)	80.62	80.44	78.80	81.22	**77.42**	79.48	78.42	79.67	**82.95**	**83.28**
SVM	81.90	84.46	67.06	77.26	**69.76**	83.45	64.54	63.46	**83.28**	**84.03**
Bagging(DT)	85.54	84.83	85.31	83.22	**82.76**	83.45	81.54	80.95	**85.42**	**85.98**
Bagging(KNN)	80.06	79.35	80.58	77.50	**78.05**	77.09	75.64	73.48	**80.06**	**83.77**
Bagging(SVM)	78.61	77.42	69.98	77.26	**69.77**	75.09	64.91	63.50	**78.61**	**84.06**
Boosting(DT)	84.18	82.98	83.40	81.74	**80.38**	80.58	78.55	77.74	**84.18**	**85.33**
Boosting(KNN)	80.0	79.29	80.45	77.44	**77.96**	77.02	74.57	74.36	**80.0**	**83.65**
Boosting(SVM)	78.61	77.75	69.61	77.26	**65.44**	75.95	66.5	63.70	**78.61**	**84.03**
Voting(KNN&SVM)	80.0	79.29	80.45	77.44	**77.96**	77.02	75.51	73.47	**80.0**	**83.65**
Voting(DT&KNN)	80.71	80.05	80.64	78.09	**78.1**	77.56	67.56	73.84	**80.71**	**83.35**

Table 1. Comparison between the different feature sets, denoted by S1,..., S10.

For base-level classifiers, DT has the best recognition accuracy between 81.01 % and 84.94%. For meta-level classifiers, bagging with DT has achieved the best recognition accuracy, between 80.95% and 85.98%. Boosting with DT has achieved accuracies between 77.74% and 85.33%.

The results acquired using **S10**, **S1**, **S2**, **S3** and **S9** have the best overall results. The elimination of one and two feature **(S1, S2, and S3)** does not affect the overall results. The elimination of three features (S4, S5 and S6) results in overall decline of the algorithm accuracy. Entropy and energy are derived from the FFT domain, in cases where they are eliminated simultaneously, the performances of certain algorithms have declined. They are typically useful features for certain activities. For instance if the user is climbing stairs, the mean and standard deviation will lie in the same region as descending stairs, but energy will change which will help to recognize the activity. Similarly, eliminating variance and standard deviation has also produced weaker overall results. These features help to differentiate between daily activities. With the elimination of two features, there is a decline of overall 2% in the performance of every algorithm. The results overall are still acceptable, but the performance of some algorithms

such as SVM has drastically gone down. When we eliminated four features S7, and S8, the performances of the algorithms have degraded further. It results in the lowest accuracies, but consumes minimum resources, resulting from the fact that all the accuracies are calculated using two features.

S10 has the best accuracy results, but if we compare the results produced using **S9**, they do not have a large difference. The differences between the two are hardly more than 1%, which might be insignificant. The amount of resources used for **S10** will be higher. By including Meantilt as a feature, the accuracy reaches the same range as **S10**, but it is easy to calculate and consumes minimal resources. It has improved the accuracy of classifiers. In terms of easy-to-compute features, **S9** and **S5** have acceptable results and at the same time consume minimum resources. By skimming the results acquired using these subsets, one will analyze that **S9** has an accuracy improvement of at least 2% for every classifier and in some cases even more. For bagging with SVM, there is an accuracy improvement of more than 8%. Similarly for NB, there is the accuracy improvement of more than 5% between the two sets, but the biggest difference was observed for SVM as a base-level classifier and boosting with SVM, which have the improvement of 13%.

We thus prefer **S9** over the other feature sets at this point because it has the second highest results in terms of accuracy after **S10**, and at the same time consumes significantly less resources. The inclusion of entropy and energy results in a slightly improved accuracy, but they have higher computation cost and consume memory [6][7], which makes them less suitable for real time applications. For example, if we have a data logged over a longitudinal period of time and one uses FFT to compute the coefficients, it will take a huge time and consume a lot of memory, compared with the time domain features, which are easy to calculate and consumes less memory.

From the above discussion, we have observed that, after adding extra features, the processing time will increase, but it will also help to increase accuracy. Certain feature sets will help to recogniz certain activities or classes of activities, but a full feature set could be useful in detecting a wide range of activities. A classifier or model could then automatically select the feature subset that is most suited for a given task. There is a tradeoff between accuracy, and processing and memory consumption. But if we can achieve satisfactory results using easy-to-compute features, then there is no point in using complex features.

With every feature set, bagging and boosting with DT and DT as a base-level classifiers have the best accuracy results, followed by NB with 79.48 % accuracies. In Figure 2, the black line shows the average accuracy for each of the classifiers and accuracy results for the feature sets. The bars in Figure 2 show the results for selected feature sets.

5 Significance t-test Evaluation

The evaluations were repeated by performing a two-tailed t-test to find out if the differences between the recognition results using all the features and with certain

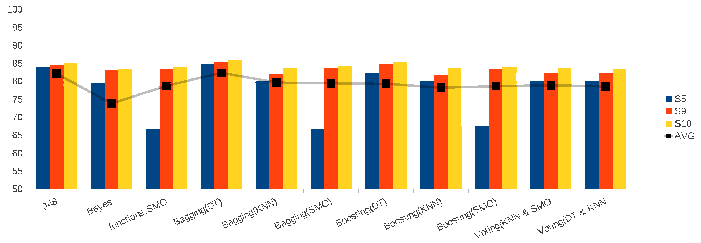

Fig. 2. Accuracy results for different feature sets and average results for all classifiers.

features removed are significant, and to investigate whether some features are negligible.

5.1 Hypothesis

Our null hypothesis is stated as, "There is no significant difference between the accuracy of the machine learning algorithms, when certain features were removed". Our alternative hypothesis is stated as, "There is a significant difference between the accuracy of the machine learning algorithms, when certain features were removed." We have performed a two-tailed t-test with $alpha$=0.1, which mean hat, $alpha$ for each side will be 0.05. The significance value is t=±1.73. To reject our null hypothesis, the value of t must be either significantly higher or lower than the significance value.

Tables 2, and 3 show the results, for which there was a significant difference in the result. The dash (-) sign represents that there was no significant difference between the two data sets. We have compared the data sets values of the feature sets with each other. Initially we have compared the data sets values from S1 with the remaining sets till S8. After that we compared S2 with the remaining sets and so on. It can be observed that SVM is effected most by the variations in feature subsets, DT as a base-level classifier did not show any significant differences with the change of feature sets.

When we compared S1 with remaining sets till S8, highest significant values were observed for S7 and S8. Similarly there was not significant differences between S1 and S2. For data sets S4, S5 and S6 the differences are moderate. Similarly for S2 maximum significance was observed with S7 and S8, while for S4 significance was minimum. By comparing S7 and S8 with the other sets, the highest significance was observed, proving the fact that the accuracy differences between these sets and other are significant.

In few feature sets, there were only one or two classifiers, which were effected by the variation of feature sets. For example in Table 1, the accuracy differences between S9 and S10 are almost 1% or in some cases even more than 1% for every algorithm, which seems like a differences, but after performing the significance test, we found that there is no significant difference between them. Similarly there are no significant differences between between S1 and S2, S2 and S6 etc.

	Algorithm	S3	S4	S5	S6	S7	S8		S3	S4	S5	S7	S8		S4	S6	S7	S8
	SVM	7.47	7.68	6.35	-	7.008	-		3.78	3.10	3.93	3.84	7.12		3.70	3.80	-	3.71
	Bagging(DT)	-	-	-	-	3.27	2.43		-	-	-	-	1.83		-	-	-	2.16
	Bagging(KNN)	-	-	-	-	3.46	3.13		-	-	-	-	2.57		1.84	-	1.78	3.16
S1	**Bagging(SVM)**	4.61	-	4.73	-	4.57	3.81	S2	3.77	-	3.92	3.84	-	S3	-	3.71	3.71	-
	Boosting(DT)	-	-	1.81	2.16	3.43	4.16		-	-	-	2.08	2.81		-	-	2.48	3.20
	Boosting(KNN)	-	-	-	-	-	2.95		-	-	-	-	2.30		-	-		2.91
	Boosting(SVM)	4.29	-	4.34	4.07	4.70	-		3.77	-	3.69	4.02	-		3.45	3.55	-	3.3
	Voting(KNN&SVM)	-	-	-	-	-	3.11		-	-	-	-	2.56		-	-	1.78	3.10
	Voting(DT& KNN)	-	-	-	1.95	2.17	3.61		-	-	-	7.53	3.05		-	-	1.97	3.27

Table 2. Evaluation results of activity recognition with certain features removed.

	Algorithm	S5	S6	S7	S8		S6	S7	S8		S7	S8		**S10**
	SVM	2.93	3.24	3.74	3.85		3.92	3.89	4.12		3.83	4.18		-
	Bagging(KNN)	-	-	-	-		-	-	3.02		-	2.17		-
	Bagging(SVM)	3.91	-	3.75	-		-	3.81	3.02		3.84	3.84		-
S4	Boosting(DT)	-	-	-	1.96	S5	2.80	-	2.17	S6	-	-	**S9**	-
	Boosting(KNN)	-	-	-	-		-	-	2.62		-	-		-
	Boosting(SVM)	3.50	-	3.82	-		3.60	3.56	3.35		3.93	3.97		-
	Voting(KNN&SVM)	-	-	-	-		-	-	2.85		-	2.15		-
	Voting(DT& KNN)	-	-	-	-		-	-	2.76		2.49	2.29		-

Table 3. Results achieved using S2 and S5 compared with the other feature sets.

6 Conclusions

This paper's contributions are twofold: A new feature for activity recognition on inertial data, 'Meantilt', is presented. Our experimental results have shown that the introduced feature has improved the accuracies of classifiers by slightly more than 2%. Our second contribution lies in the evaluation and significance testing of all combinations of feature sets. We have evaluated our feature sets with different base-level and meta-level classifiers to find out how the accuracy and significance of these feature sets are affected.

Within this evaluation and significance test, we have focused and selected sets of easy-to-compute features, which consume minimal resources while providing acceptable results. The most promising set of features was found to contain mean, standard deviation, variance and Meantilt. We have concluded that meta-level classifiers provide the best accuracy results compared to base-level classifiers. Bagging or boosting with a decision tree, and decision tree as a base level classifier have the best accuracy results for almost all the feature sets. Adding other features increased the accuracy results slightly with 1%, but comes at a considerably higher processing cost.

References

1. Caspersen, C. J.; Powell, K.E.; Christensen: Physical Activity, Exercise, and Physical Fitness: Definitions and Distinctions for Health-Related Research, Public Health Rep. 100 126 31, 1985
2. Makikawa, A.; Asajima, S.; Shibuya, K.; Tokue, R.; Shinohara, H.: Portable Physical Activity Monitoring System for the Evaluation of Activity of the Aged in Daily Life,

in : Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, IEEE, Housto, USA, 2002
3. Maurer, U.; Smailagic, A.; Siewiorek, D.P.; Deisher. M.: Activity Recognition and Monitoring Using Multiple Sensors on Different Body Positions, in: Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks, IEEE, Washington, USA, 2006
4. Chun, Z.; Weihua, S.: Human Daily Activity Recognition in Robot-assisted Living Using Multi-sensor Fusion, in: 2009 IEEE International Conference on Robotics and Automation Kobe International Conference Center, Kobe, Japan, May 12-17, 2009
5. Wang, S.; Yang, J.; Chen, N.; Chen, X.; Zhang, Q.: Human Activity Recognition with User-Free Accelerometers in the Sensor Networks, IEEE Int. Conf. Neural Networks and Brain, vol. 2, pp.12121217, 2005
6. Bao, L.; Intille, S.S.: Activity Recognition from User-annotated Acceleration Data, Pervasive 2004, vol. 300, pp.1-17, Apr. 2004
7. Ravi, N.; Dandekar, N.; Mysore, P.; Littman, M.L.: Activity Recognition from Accelerometer Data, Proceedings of the Seventeenth Innovative Applications of Artificial Intelligence Conference, pp. 15411546, 2005
8. Sa-kwang, S.; Jaewon, J.; Soojun, P.: A Phone for Human Activity Recognition Using Triaxial Acceleration Sensor, in: International Conference on Consumer Electronics, Las Vegas, USA, 2008
9. Huynh, D.: Human Activity Recognition with Wearable Sensors, PHD Thesis, 2008
10. Sun Microsystems,http://www.sunspotworld.com/docs/Red/Tutorial/Tutorial.html, last visited 1st May 2011
11. Lau, S. L.; Knig, I.; David, K.; Parandian, B.; Carius-Dssel, C. & Schultz, M.: Supporting Patient Monitoring using Activity Recognition with a Smartphone, The Seventh International Symposium on Wireless Communication Systems (ISWCS'10). 2010
12. Pltz, T.: How To Do Good Research In Activity Recognition, Position paper, Newcastle, UK, 2010
13. Tapia, Al.: Real-Time Recognition of Physical Activities and Their Intensities Using Wireless Accelerometers and a Heart Rate Monitor, in: Wearable Computers 11th IEEE International, 2007
14. Preece, Al.: Activity Identification using Body-Mounted Sensors- a Review of Classification Techniques, Physiological Measurement, vol. 30, no. 4, pp. R1-R33, April 2009.
15. Danny, W.; Matthai, P.; Tanzeem, C.: Unsupervised Activity Recognition using Automatically Mined Common Sense, In the Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI 2005), July 2005, Pittsburg, PA
16. Pirttikangas, S.; Fujinami, K.; Nakajima, T.: Feature selection and activity recognition from wearable sensors, Lectures Notes Computer Sciience, vol. 4239, p. 516, 2006
17. Fukunaga, K.: Introduction to statistical pattern recognition (2nd ed.), San Diego, CA, USA: Aca- demic Press Professional, 1990