IMPROVING ACCELEROMETER BASED ACTIVITY RECOGNITION

Maja Žbogar[1], Hristijan Gjoreski[1, 2], Simon Kozina[1, 2], Mitja Luštrek[1]
[1]Department of Intelligent Systems, Jozef Stefan Institute
[2]Jozef Stefan International Postgraduate School,
Jamova 39, 1000 Ljubljana, Slovenia
e-mail: maya.zbogar@gmail.com, {hristijan.gjoreski, simon.kozina, mitja.lustrek}@ijs.si

ABSTRACT

This paper presents the findings of a research on how to improve activity recognition from data captured with chest mounted accelerometer. Several methods were applied to achieve this purpose, including: simple smoothing technique, hidden Markov models (HMM), extraction of frequency domain features, principal component analysis (PCA) and dynamic time warping (DTW). The paper describes each of the methods and presents the achieved results of our research. The results indicate that these techniques, used separately or in combination, can increase recognition accuracy.

1 INTRODUCTION

Well-performing activity recognition systems have many potential usages. Being able to precisely determine human activity at a certain point in time is potentially useful in many domains. Such systems can help in solving the problem of elderly care [1, 8, 9] and are also valuable for other more mundane usages like: sports, gaming and entertainment industry [2]. Besides being effective, it is also very important, that activity recognition systems strive to be economically feasible for wide spread usage [1]. One of the relatively cheap and widely available technologies, that allow the recognition of activities, is easily accessible wearable accelerometers. However, the person wearing the sensors should also feel comfortable; therefore it is important to keep the number of sensors down to minimum.

For these reasons, we studied different ways of improving the basic machine learning (ML) activity recognition, using data captured with a single chest mounted accelerometer. We tried to achieve greater accuracy by applying various techniques such as simple smoothing technique, hidden Markov models, extraction of frequency domain features, principal component analysis and dynamic time warping. Results show that some of these techniques can potentially improve accuracy, especially if they are combined.

2 CONFIDENCE DATA SET

Our initial data set was created as a part of Confidence project, which is aimed at creating a ubiquitous health care system to support independent elderly living [1]. The recorded activities were: sitting, sitting on the ground, on all fours, lying, standing, walking and transitional activities consisting of going up and going down. We used data captured with a three-axial accelerometer mounted on the test persons' chest. A special test scenario, containing all activities, was created. The scenario was recorded by 11 young, healthy volunteers (7 males and 4 females). It was repeated 5 times by each person, resulting in 55 recordings. Classification accuracy was determined by using leave one person out cross validation technique.

2.1 ML Baseline approach: Random Forest

In order to evaluate our improvement, we first established a baseline approach. The raw sensor data was preprocessed with low-pass filter, which reduced the problem of noise in the collected data. After smoothing the raw sensor data, we applied an overlapping sliding window technique. Inside of every window frame, we computed time domain features such as: standard deviation, mean value and root mean square. The machine learning analysis was made using the application program interface of the software toolkit WEKA. Several algorithms were tested and the Random Forest (RF) was the algorithm that yielded the best results in preliminary tests. RF is an ensemble of decision trees in which the final decision is made by majority vote of the tree models [1]. Baseline classification accuracy was 72.2%. Initially we identified two main problems with this approach. First problem is spurious transitions between activities. Second problem is the inability to distinguish between certain groups of activities such as standing and sitting. Our goal was to address these problems.

2.1 Simple modus classifier output smoothing technique

We first tried to tackle the problem of spurious transitions between classified activities. Results of our ground level RF classification algorithm contained a great amount of erratic transitions from one class label to another. These fast, momentary transitions are impossible to occur in normal day to day activities.

We tried to reduce the problem with a simple technique of changing the activity labels, based on the surrounding labels. Sliding windows of different lengths were used to correct the recognized activity of the ground level RF classifier. Corrected activity was the majority activity within the observed window. This simple technique presents a variety of other possibilities on how to implement this type of a level two pseudo meta-classifier. Different parameters, for instance length of surrounding window, can be considered. Another version of the same method was also implemented. RF's probability distribution was summed up for each predicted activity within time window. The activity with maximum sum was selected as the corrected activity. However, this variation did not produce any noticeable improvement.

Results of this sliding window smoothing technique showed a slight improvement in accuracy. Classification accuracy increased to 73.8%, which is 1.6 percentage points better than baseline accuracy (Table 1). Although there was an increase in accuracy, this method suffers from a problem of error propagation. In case of miss correcting previous activity labels, we can cause further errors and create sequences of mislabeled activities. Furthermore, it also cannot address the problem of confusion between activities, where chest is in the same vertical or horizontal position, e.g. sitting and standing. These two postures are characterized by chest being in vertical upright position. Easy way to avoid this problem is to combine both types of activities. However, we can argue that this is not a solution per se, but rather a mere elimination of the problem itself.

2.2 Hidden Markov models (HMM)

The problem characterized by spurious class label transitions, lies in the fact, that machine learning algorithms like RF, fail to take into account the continuity of the processes such as human activity. They discretely classify each instance in isolation and assume there is no connection between previous and following instance.

A common way to address this problem, and consequently reduce spurious activity transitions, is to add temporal dependence component by using hidden Markov models (HMM) [3]. HMM observes Markov property, which states that current system state is dependent only on the previous state of the system. The model consists of a number of hidden states and associated transition probabilities between these hidden states. The hidden states emit events with certain emission probability, and these events are observed by the outside observer [4]. Our hidden system states were the true class label sequences, which were unknown. Observed states were RF's predicted class labels. We generated test sequences on our training data. After generating the sequences, we used Viterbi dynamic programming algorithm, which is used to generate the most likely sequence of hidden states given an observation sequence of events [4]. Output of Viterbi algorithm was used to correct the initial RF's erratic predictions.

Using HMM to solve practical problems brings many questions. First question is how to build the initial model. One way is to use Baum-Welch algorithm. Using this approach the obtained classification accuracy slightly increased from 72.2% to 73.9%. We also tried building the Markov model manually, by learning the emission and transition probabilities directly from our training sets. Latter option seems to perform slightly better. As seen in Table 1, classification accuracy increased to 75 %, which is 2.8 percentage points better than our ground classifier. Possible reason, why this later option outperformed the first, is that we were actually learning the distribution of activities in our test scenarios. Building model directly on the training data also brings another potential problem of possible deterioration in performance on new data, since the ratio of our daily life activities and likelihood of transitions between them is not the same. Another question is the selection of the length of training sequences.

To a large degree, we can contribute the improvement in classification accuracy, to the elimination of spurious transitions. Results show slight improvement in accuracy, but the problems of error propagation and inability to distinguish between upright positions like sitting and standing, remain. Next step was an attempt to solve these problems, by improving the classifier on the first level. This effort was based on the notion, that, if we get better results with our ground classifier, second level classifier and other techniques, have a bigger chance of performing in a better manner. To achieve this, our next step was to try to improve our feature set.

	Accuracy
Only Random Forest	72.2%
Simple modus smoothing technique	73.8%
Hidden Markov Model	**75.0%**

Table 1: *Comparison of classification accuracy achieved using different techniques.*

3 CHIRON DATA SET

Second part of the research was conducted on Chiron project dataset. Chiron project is an European project whose final goal is to develop reference architecture for personal elderly care [6]. The data set, among other data sources, also contains inertial force measurements captured with three-axial accelerometers. In this section new attributes are derived from the data in order to improve the classification accuracy on the first level.

3.1 Baseline accuracy: Time domain feature set

As in our first experiment, baseline accuracy is evaluated. Raw accelerometer data was preprocessed with low-pass filter. RF was used as a classification algorithm. For the purpose of signal segmentation and time domain feature computation, several lengths of overlapping sliding windows were used. Feature set contained same time domain features as the ones used in previous section. Due to the fact, that data consists of several different scenarios, we preselected some of them. Selected data contained our target activities: sitting, running, walking, standing, lying, on all fours, standing up and going down. With the exception of adding running activity and removing sitting on the ground activity, target activities are same as in the previous section. Accuracy of ground level classifier, using time domain features, is presented in the first row of the Table 2. Using only information extracted in the time domain, classification accuracy varies from 69.0% up to 72.2%, depending on the length of time window.

3.2 Frequency domain features

In the next step, we tried to enhance the attribute set in the feature vector, consequently increasing accuracy and improving distinction between upright activities. The feature vector includes important cues for distinguishing various activities [2]. Our goal was not to rely solely on time domain features, but to also try to gather addition information from features in the frequency domain. To compute meaningful features in frequency domain over a time window, algorithm needs higher data sampling frequency dataset. This is the reason we used Chiron dataset with 20 Hz sampling frequency. Features focusing on periodic structure of the signal in the frequency domain are also commonly used in other activity recognition studies [2].

We used coefficients derived from Fast Fourier Transforms. Features such as magnitude, spectral energy and maximum magnitude index, were computed for all three accelerometer axis. First we measured accuracy using only frequency domain features. Results are shown in Table 2. Measured accuracy in accordance to our expectations, depends of the length of the overlapping sliding window and ranges from 68.0% up to 76.7%. This time the lower bound is one percentage point less than in the case of using time domain features, but the upper bound is four percentage points better. We also experimented with different combinations of frequency domain feature subsets, but they did not yield significantly better results

3.3 Principal component analysis (PCA)

Principal component analysis (PCA) is a well known and widely used statistical analysis method to transform high dimensional data into a lower dimensional space [2]. It is popular in number of areas ranging from neuroscience to graphics and image compression [5].

Our idea was to extract the information available from the raw signal. By using PCA, we were hoping to utilize the characteristics of the signal. Accelerometer signal consists of x, y and z axis inertial force measurement. If one, for example takes a 64 sample window, this result in 192 attributes, which is a rather big amount of features. Furthermore one can perform PCA on different things. We can take raw signal values, acceleration angles, differences between current acceleration angle and previous angle and other signal characteristics. In order to reduce the number of features we used PCA. Results in terms of classification accuracy of our experiment using PCA are shown in Table 2. We used 3 eigenvectors computed on raw acceleration data. PCA by itself seems to perform worse, than using either only time domain or only frequency domain attributes.

	Sliding window size 16	**Sliding window size 24**	**Sliding window size 32**	**Sliding window size 64**
Time domain features	69.0%	67.6%	72.7%	70.6%
Frequency domain features	68.0%	71.9%	73.0%	76.7%
PCA features	62.8%	63.7%	63.8%	63.0%
Time and frequency domain features	73.4%	73.2%	**76.6%**	71.8%
PCA, time and frequency domain features	**76.8%**	**78.1%**	75.6%	**77.6%**

Table 2: *Table of accuracy of different feature sets and different overlapping sliding window sizes.*

3.4 Comparison of different feature sets

By examining Figure 1, one could draw a conclusion, that using only baseline time domain features (blue line), performs worse than using additional features (red and green line).

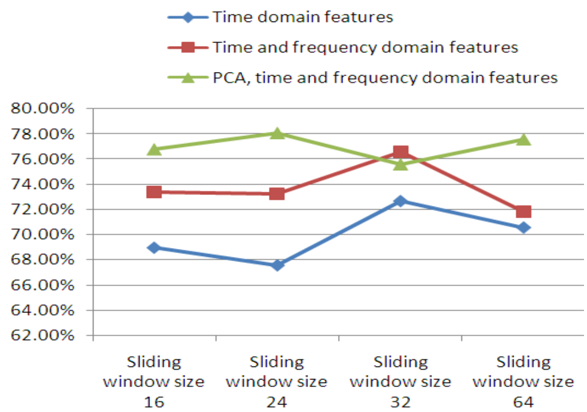

Figure 1: *Comparison of classification accuracy of different feature sets.*

Using all of the available features including frequency, time domain and principal components (green line), at first sight appears to perform better than other feature set combinations, because it achieves greater accuracy for the most of the considered sliding window sizes. It is hard say, if that is really the case, due to the dependence on preprocessing window size and limited number of test subjects. In an event that this approach would prove to work better, there is still a problem of computational complexity, which increases in accordance to growing number of features. Furthermore, there is also another important thing to consider, since certain other classification algorithms tend to show greater sensitivity to excessive usage of redundant features.

3.4 Dynamic time warping (DTW)

Our basic inspiration for using dynamic time warping (DTW) was to find similarities in the shapes of the accelerometer signal. DTW is a much more robust distance measure for time series, than Euclidean distance measure. It allows similar shapes to match, even if they are out of phase in time axis [7]. This distance measure is widely used in many areas like medicine, science and finance [7].

We implemented an activity recognition system that matched every single test instance, which we would like to classify, against all other instances in our training set and classified the unknown activity to the majority class of the closest n activities. We evaluated our implementation on a small subset of test instances, but it did not perform as well as other previously mentioned techniques and achieved a rather poor classification accuracy of 51.2%. Small subset of instances was chosen, because we stumbled against the problem of high computational time complexity. Issue was mainly caused, because our activity recognition system was designed to match every single test instance, against all other instances in our training set and classified the unknown activity to the majority class of the closest n activities. This implementation also suffers from a similar problem as the lazy k nearest neighbors' (KNN) classifier, since it is always difficult to know how many closest instances one should use in the classification process. Potentially better idea would be to select a subset of exemplary activity of certain target class and compare our unknown activity only against this smaller subset. By implementing our system in this way, the DTW classifier could work faster. The question of how to actually select this smaller subset could present one of the starting points for further research.

6 CONCLUSIONS

It is hard to say with some convincing degree of confidence, which of the used techniques works best. In spite of our endeavors to improve activity recognition, we did not manage to achieve all the expected improvements. One of the problems is the side effect of using different data sets. For this reason results cannot be rightfully compared. Furthermore, research was conducted only on two different data sets. As a consequence of this small verification scale, our findings cannot be generalized to other datasets and it is impossible to say, how well they could perform outside the experiment setting, i.e. in real-time activity recognition setting. It is also very difficult to compare our results with other research findings, since different researches examine different sets of target activities. Furthermore target activities can be defined differently. For instance, the sitting activity can be defined in many different ways e.g.: sitting on the ground, sitting on a chair, sitting with legs crossed etc. This problem of transparency and lack of standard definitions is unfortunately widely present in the field of activity recognition research. However, we can conclude with an optimistic notion that talking small steps is required in order to make a big leap forward. Same principle of gradual progress also applies to making better research and improved activity recognition systems.

Acknowledgements. This work was partly supported by the Slovenian Research Agency under the Research Programme P2-0209 Artificial Intelligence and Intelligent Systems and partly from the European Community's Framework Programme FP7/2007-2013 under grant agreement No.214986.

References

[1] H. Gjoreski. Adaptive human activity recognition and fall detection using wearable sensors. *Msc Thesis*. Jozef Stefan International Postgraduate School. Ljubljana, Slovenia. 2011.

[2] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, P. Haviga. Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. *ICACS*. The Netherlands. 2010.

[3] B. Kaluža. Reducing spurious activity transitions in a sequence of movement. *Proceedings of the Eighteenth International Electrotechnical and Computer Science Conference – ERK 2009*, vol. B, pp. 163-166, Portorož, Slovenia, September 2009.

[4] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEE*. Vol. 77. pp. 590-599. 1999.

[5] L. Smith. A tutorial on principal component analysis. www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf. 2002.

[6] H. Gjoreski. Risk assessment model for congestive heart failure. *4th Jozef Stefan International postgraduate school students' conference*. Ljubljana, Slovenia. 2012.

[7] E. Keogh, C. A. Ratanamahatana. Exact indexing of dynamic time warping. *Proceedings of the 26th Int'l. Conference on Very Large Data Bases*. pp. 406-417. Hong Kong. 2004.

[8] The European FP7 Confidence Project. Available: http://www.confidence-eu.org

[9] CHIRON project JU ARTEMIS. Available: http://www.chiron-project.eu/