Machine Learning Project
Analyzing runners profiles
First Report

Jessa Bekker (s0215494)
Koen De Cock (s0214395)

November 2013

1 Introduction

This report describes how we will address the tasks of interfering on what surface runners run and whether or not the runners are trained.

To address the task of runner classification, four main questions need to be answered. 1) Which features will be used for learning, 2) how those features will be derived from the raw data, 3) which machine learning technique to use and 4)how to evaluate the classifier. Our (partial) answers to these questions are addressed in sections 1 to 4.

2 Which Features

To apply the learning algorithm to the examples, features have to be selected that will represent the data. The idea is to select features that will be useful to make the difference between the classes.

The paper of Preece et al.[3] compares different feature extraction methods for the Classification of Dynamic Activities From Accelerometer Data. Their main conclusions are that time/frequency features significantly outperform wavelet features. Especially using the magnitude of the first five components of the FFT power spectrum as features gave good results. The paper however distinguishes different types of activities in contrast to our task: different types of running. The latter might be more subtle, so results could be different in our case. Components of the FFT power spectrum might however still be interesting to look at.

When running on different terrains, the force applied by the feet will differ. Hence adding the maximal force during a step can be added as a feature.

The maximum, mean and minimum speed may point out a difference between trained or untrained runners and terrain. The step-frequency may also be a useful piece of information. Both the average frequency and the variation in frequency can be used. The expectation is that trained runners will run more rhythmically than untrained runners and thus have less variation in their step-frequency.

The step length may point out towards some specific terrain or running behaviour.

3 Derivation of Features

The acceleration can be taken as a measure of force. However, the raw acceleration data will need to be smoothed to reduce the noise[5]. This can be done by taking the average of a number of samples.

To extract the speed from the acceleration data, dead reckoning has to be used. This induces a cumulative error, which can be reduced with a number of optimizations:

- Electrical noise can be reduced by averaging over a number of samples.[5]
- Mechanical noise during period of no movement can lead to instabilities. This can be reduced with a filtering window that corrects acceleration values close to zero.[5]
- When a movement ends (e.g. a step), the velocity should be zero. However, errors in the integration of the acceleration can lead to a final non-zero velocity. This can be corrected by reducing it to zero when there is a long period of zero acceleration[5] or when detecting a step[6]. The latter will be most useful when running, since periods of rest are not expected.

The step-frequency can be measured by detecting the steps in the data and using the time-span of one step. The ankle-data, which will have approximately the same acceleration as the foot will show a peak each time the foot lands.

The steps can be detected by searching the data for these peaks or for zero-crossings.[6]

Most methods of calculating the step length make use of some user-specific parameters[6]. Since this is not available, it has to be calculated differently. One way is to derive it from the speed and step frequency, but this may not be sufficiently accurate.

4 Machine Learning Techniques

Once features are selected and derived, these can be used to apply a machine learning algorithm on.

The paper of Albert et al.[2] describes five techniques: SVM, Sparse Multinominal Logistic Regression, Naive Bayes, k-nearest neighbors, decision trees. The best results were achieved with SVM and SMLR. K-nearest neighbors and decision trees also performed quite well. Naive Bayes however, returned poor results. This motivates us to try SVM first and compare it with decision trees and k-nearest neighbors. Since SMLR is not part of the course, using it would be optional.

5 Evaluation of Classifier

Cross validation is used in both [2] and [3] for evaluating the classifier and features. This is the technique we will apply as well.

6 Time Spent

This is the time spent on the first part of the project:

Jessa:

What	**time**
Literature Study [1][2][3][4]	4:00
Writing report 1	3:45

Koen:

What	**time**
Literature Study [4][5][6]	5:00
Writing report 1	3:45

References

[1] Mohamed R. Al-Mulla, Francisco Sepulveda, and Martin Colley. A review of non-invasive techniques to detect and predict localised muscle fatigue. *Sensors*, 11(4):3545–3594, 2011.

[2] Mark V. Albert, Konrad Kording, Megan Herrmann, and Arun Jayaraman. Fall classification by machine learning using mobile phones. *PLoS ONE*, 7(5):e36556, 05 2012.

[3] Stephen J. Preece, John Yannis Goulermas, Laurence P. J. Kenney, and David Howard. A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. *IEEE Trans. Biomed. Engineering*, 56(3):871–879, 2009.

[4] David Sachs. URL: `http://www.youtube.com/watch?v=C7JQ7Rpwn2k`, 2010.

[5] Kurt Seifert and Oscar Camacho. Implementing positioning algorithms using accelerometers. *Freescale Semiconductor*, 2007.

[6] Oliver Woodman and Robert Harle. Pedestrian localisation for indoor environments. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 114–123. ACM, 2008.