



## AN ABSTRACT OF THE THESIS OF

Michael M. Anderson for the degree of Master of Science in Computer Science  
presented on January 1, 2013.

Title: Activity Detection on Free-Living Data Using Change Point Detection

Abstract approved: \_\_\_\_\_

Weng-Keen Wong

(Abstract text)

©Copyright by Michael M. Anderson  
January 1, 2013  
All Rights Reserved

# Activity Detection on Free-Living Data Using Change Point Detection

by

Michael M. Anderson

A THESIS

submitted to

Oregon State University

in partial fulfillment of  
the requirements for the  
degree of

Master of Science

Presented January 1, 2013  
Commencement June 2013

Master of Science thesis of Michael M. Anderson presented on January 1, 2013.

APPROVED:

---

Major Professor, representing Computer Science

---

Director of the School of Electrical Engineering and Computer Science

---

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

---

Michael M. Anderson, Author

## ACKNOWLEDGEMENTS

(Acknowledgement text)

# TABLE OF CONTENTS

|   | <u>Page</u> |
|---|-------------|
| 1 Introduction                                | 1           |
| 1.1 Motivation . . . . .                      | 1           |
| 1.2 Previous Research . . . . .               | 1           |
| 2 Methodology                                 | 2           |
| 2.1 Overview . . . . .                        | 2           |
| 2.2 Datasets . . . . .                        | 2           |
| 2.2.1 OSU Hip . . . . .                       | 2           |
| 2.2.2 UQ . . . . .                            | 3           |
| 2.3 Featurization . . . . .                   | 3           |
| 2.4 Base Classifiers . . . . .                | 6           |
| 2.5 Performance Metrics . . . . .             | 6           |
| 2.6 Top-Down Approach . . . . .               | 7           |
| 2.6.1 Change Point Detection . . . . .        | 7           |
| 2.6.2 Experiments . . . . .                   | 9           |
| 2.7 Bottom-Up Approach . . . . .              | 11          |
| 3 Results                                     | 12          |
| 3.1 Discussion . . . . .                      | 12          |
| 4 Conclusion                                  | 14          |
| 4.1 Summary . . . . .                         | 14          |
| 4.2 Directions for Further Research . . . . . | 14          |
| Bibliography                                  | 14          |

## LIST OF FIGURES

| <u>Figure</u>                                      | <u>Page</u> |
|--|-------------|
| 2.1 OSU Hip Activity Samples . . . . .             | 4           |
| 2.2 UQ Day 2 Activity Samples . . . . .            | 5           |
| 2.3 Change Point Detection . . . . .               | 8           |
| 3.1 CPD-Based Classification Performance . . . . . | 13          |



## Chapter 1: Introduction

### 1.1 Motivation

### 1.2 Previous Research

## Chapter 2: Methodology

### 2.1 Overview

Each dataset that we tested consisted of multiple time series gathered from a number of different subjects, so to perform an experiment on a dataset we began by partitioning the dataset into disjoint subsets of training, validation, and test data. Each individual time series was then partitioned into a set of disjoint windows, and each window was featurized. Once the dataset was featurized, the experiment could be treated as a normal classification problem. We trained a base classifier using the training set, tuned it (when necessary) with the validation set, and obtained results by testing the quality of the resulting tuned model on the testing set.

### 2.2 Datasets

#### 2.2.1 OSU Hip

Our first dataset was collected by the Nutrition and Exercise Sciences department of Oregon State University, and has been used for previous activity detection research with the goal of automatically calculating and monitoring energy expenditure [13] [14]. This dataset consisted of 91 time series collected over a 2-week period in a laboratory environment. The subjects were children between the ages of 5 to 15 (with a mean age of 11 years, and a standard deviation of 2.7 years). Subjects performed 12 different types of activities (as shown in Figure 2.1) over two separate visits, while an ActiGraph GT1M accelerometer worn on their hip collected triaxial acceleration data at a frequency of 30Hz.

Data was collected from two separate visits to the lab, where the subjects performed 6 activities per visit. Children were given breaks in between each activities, and each activity lasted 5-10 minutes, however, these unlabelled breaks were removed from the version of the dataset that we used, and additionally only two minutes of data were

available for each subject. Thus, each of the 91 time series contained data from six 120 second long activities, for a total of  $6 * 120 * 30 = 21600$  ticks per time series.

For our experiment, we determined that several of the activities were very similar and that it would be difficult to discriminate between them, so we combined some of them together to create a 7 class version of the data. Our classes were lying down, sitting (hand-writing, computer game), standing (laundry, sweeping, and catch), walking (comfortable, brisk and treadmill walking), dancing, running, and basketball.

### 2.2.2 UQ

This dataset consisted of 23 time series, each containing roughly 10 continuous days worth of data from a single subject. Subjects wore an ActiGraph GT3X+ accelerometer during the entire period, which collected triaxial acceleration data at a frequency of 30Hz, as well as an activPal inclinometer on their thighs. The inclinometer provided what we considered the ground truth of the data by automatically delimitting and classifying intervals using the orientation of the subject at any given moment. It classified a horizontal orientation of the thigh as lying down/sitting, a vertical orientation as standing, and a combination of the two as walking. Figure 2.2 shows samples of accelerometer data from the 3 activities.

This dataset was challenging to work with because of its size, as each individual time series contained roughly 25 million ticks of data. To help alleviate this problem, we split each time series into individual days. We then treated the first day of data that began at midnight from each subject as one whole dataset (UQ Day 1), and the second such day as a separate dataset (UQ Day 2), and did not use any of the data from the remaining days.

## 2.3 Featurization

To formulate our experiments as classification problems, we split each time series into a set of non-overlapping windows and represented each window as a feature vector. How we decided where one window ended (and where the next began) varied between experiments, and is described in sections 2.6 and 2.7. Our feature set was a large collection of statistics that have been shown to be discriminative for activity classification in previous

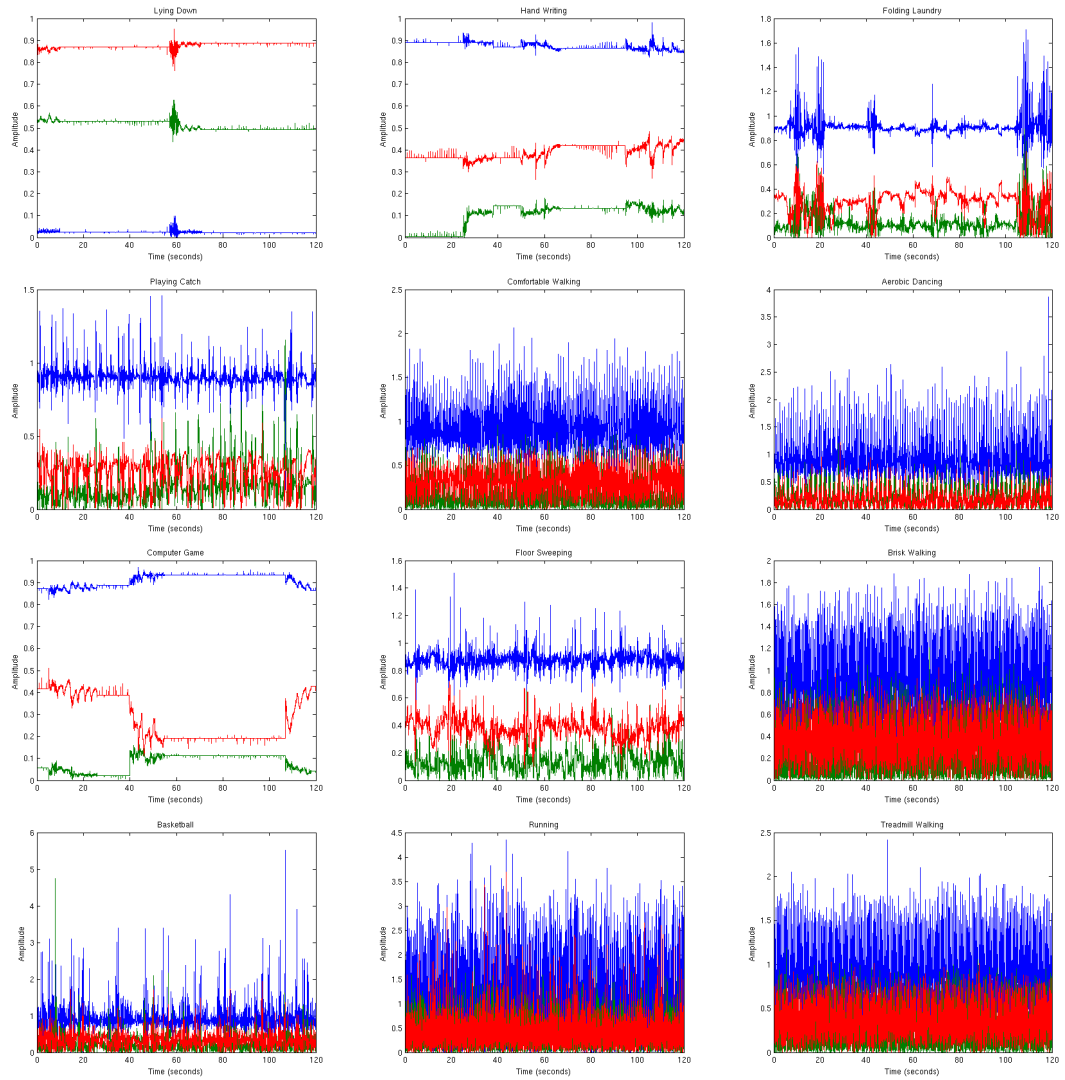


Figure 2.1: OSU Hip Activity Samples

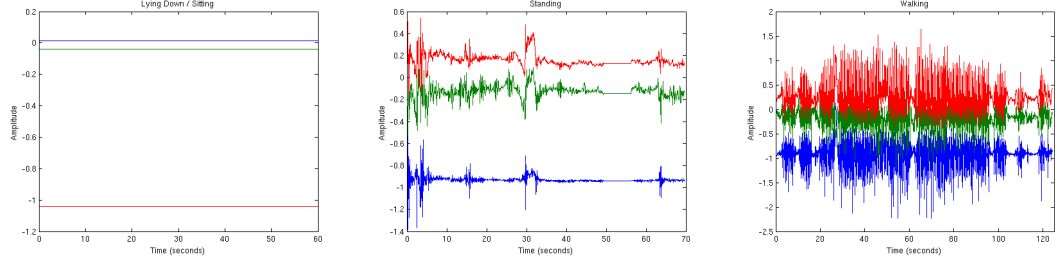


Figure 2.2: UQ Day 2 Activity Samples

research [3] [6] [9] [14]. In all we used 18 statistics that were uniaxial, i.e. were only a function of the data from a single axis of a given window, and 1 biaxial statistic. The uniaxial statistics were applied to data from each axis separately, and the biaxial statistic was applied to data from each of the  $C_2^3 = 3$  possible pairs of axes, for a total of  $18 * 3 + 3 = 57$  features.

One discriminative characteristic of an activity is its overall vigorousness, and the sum and the sample mean both act as simple and obvious ways of measuring this, as more intense activities will tend to involve higher rates of acceleration during movement. We also used the 10th, 25th (quartile 1), 50th (median), 75th (quartile 3), and 90th percentiles of the data, as well as signal power and log energy as supplemental measures of overall intensity.

Another characteristic of an activity is how much it varies in intensity. The sample standard deviation, coefficient of variation, peak-to-peak amplitude (max-min), zero crossings (the number of times the data crosses its median), as well as the interquartile range (75th% - 25th%) were useful for discriminating between activities with a consistent level of intensity (low variance, etc.) and activities that were more rhythmic or staccato in intensity (high variance, etc.).

Skewness, kurtosis, lag-one-autocorrelation, and peak intensity were useful for discriminating between activities that tend to be similar in their overall intensity and variation in intensity, but that show other types of difference in shape. Skewness indicates whether the data is more concentrated above or below its mean. Kurtosis indicates that the data is concentrated near its mean or conversely that it is fat-tailed. Lag-one-autocorrelation is a measure of the general relationship between data ticks and their immediate neighbors in time. Peak intensity is the number of times that the data reached

its maximum value.

Finally we looked at a single bimodal statistic across each pair of axes, the correlation coefficient, which discriminates between activities where acceleration values in one axis are predictive of acceleration values in another axis, verses those where that is not the case.

## 2.4 Base Classifiers

We tested 3 types of classification models on the featurized versions of our data: decision trees, support vector machines, and neural networks. We used R for our experiments, and used the ‘rpart’, ‘e1071’, and ‘nnet’ libraries to R to build our decision tree, svm, and neural net models, respectively. We treated the decision tree as a simple and quick baseline algorithm, and did not tune it in any way, ignoring the validation set. For all of the neural net experiments, the maximum number of iterations was set to 100000, and the maximum number of weights was set to 1000000.

For the OSU Hip experiments we tuned the cost parameter  $c$  of the svm on the validation set with 6 values:  $\{0.01, 0.1, 1, 10, 100, 1000\}$ . The single-layer feed-forward neural network took two tuning parameters, and we tuned with each element of the set  $H \times W$ , where  $H = \{1, 2, \dots, 30\}$  was the numbers of nodes in the hidden layer, and  $W = \{0, 0.5, 1\}$  was the weight decay parameters.

Since the UQ datasets were an order of magnitude larger, we tuned them slightly differently because of time constraints. Setting the  $c$  parameter to 1000 proved to be prohibitively expensive for the svm model, so we tuned  $c$  from the values  $\{0.01, 0.1, 1, 10, 100\}$ . Running  $30 * 3 = 90$  tuning experiments for the neural networks was also prohibitively expensive, so we drew from  $H \times W = \{5, 10, 15\} \times \{0, 0.5, 1\}$ .

## 2.5 Performance Metrics

To measure the performance of our classification algorithms we used two metrics. Accuracy is defined as the number of ticks that an algorithm correctly classifies in a time series, over the total number of ticks in the time series. Since we were also interested in our algorithms’ feasibility for activity classification in realtime, we used detection time as a second metric. Detection time is defined as the average amount of time required for

an algorithm to begin correctly classifying data after a true activity change has occurred.

## 2.6 Top-Down Approach

### 2.6.1 Change Point Detection

For this approach, the data was split into non-overlapping segments for featurization using techniques from the statistical field of change point detection. Change point detection has found application in many problem domains that require analysis of time series data from dynamic systems, including failure detection [1], quick detection of attacks on computer networks [11], and monitoring of heartbeat fluctuations during sleep [8]. Change point detection assumes that each tick of a time series is a draw from some process, but that the process may suddenly change as time passes. The goal is to predict when these changes have occurred. A *score* is generated for each time tick, and if the score is above a given threshold a change is predicted to have occurred between that tick and its immediate predecessor. To generate a score at a time tick, a window of data that immediately preceeds it (the *reference data*) is compared to it along with a window of data that immediately follows it (the *test data*).

Model-based approaches to change point detection assume that each tick in a time series is a draw from some underlying probability distribution. Scores are generated by estimating the distribution of the reference data and the test data, and then by calculating the likelihood that the two distributions are different. Where it is reasonable to assume that the data belongs to a particular family of distributions then parametric estimation methods have been employed [12]. If no such modeling assumptions are reasonable then non-parametric methods have also been found to be viable [4]. Finally, distance-based approaches such as Singular Spectrum Analysis generate scores through other metrics of dissimilarity or difference between the reference data and the test data [5]. Notationally, we say that for each tick  $i$  in a time series:

$$s_i = D(x_{r,i}, x_{t,i})$$

Where  $s_i$  is the score of the  $i$ th tick,  $x_{r,i}$  is the reference data associated with the  $i$ th tick,  $x_{t,i}$  is the test data associated with the  $i$ th tick, and  $D(A, B)$  is a function that computes the dissimilarity between a matrix of data  $A$  and matrix of data  $B$ , and

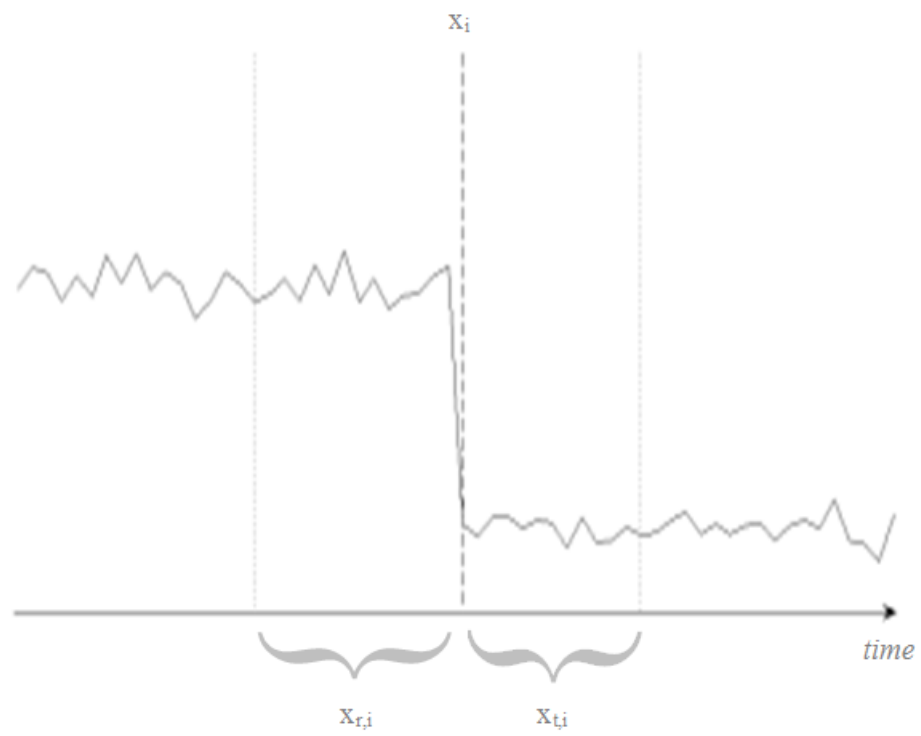


Figure 2.3: Change Point Detection



is particular to the given change point algorithm. Note that for a given algorithm it may not be possible to generate scores right at the very beginning of the time series (insufficient reference data) or at the very end of a time series (insufficient test data).

### 2.6.2 Experiments

There are many different modeling assumptions and associated algorithms for generating change point detection scores, and one simple baseline approach that we wanted to test was the Shewhart Control Chart. This approach assumes that the reference data is drawn from a multivariate normal distribution, and that scores are calculated by the Mahalanobis distance of the target time tick from the estimated multivariate normal:

$$s_i = \sqrt{(\bar{x}_{r,i} - x_i)^T S_{r,i}^{-1} (\bar{x}_{r,i} - x_i)}$$

where  $\bar{x}_{r,i}$  is the sample mean of the reference data,  $S_{r,i}$  is the sample covariance matrix of the reference data, and  $x_i = x_{t,i}$  is the  $i$ th data point [7].

We were also interested in testing the performance of a newer and more sophisticated change point detection algorithm: the Kullback-Leibler Importance Estimation Procedure (KLIEP), introduced by Kawahara and Sugiyama [2] [10]. This approach generates scores using the Kullback-Leibler (KL) divergence between the reference data and the test data. One method of doing this is to estimate the density of the reference distribution and test distribution separately, and then compare them using a likelihood ratio (known in the change point detection literature as *importance*). Instead, KLIEP estimates the importance directly using a non-parametric model.

Let the estimate of the importance  $\hat{R}$  be represented by this model:

$$\hat{R} = \frac{p_t}{\hat{p}_r} = \sum_{j=1}^{n_t} \alpha_j K_G(x, x_{t,j})$$

Where  $p_r$  and  $p_t$  are the probability densities of the reference data and the test data,  $n_t$  is the number of ticks in the test window,  $\alpha$  is a vector of model parameters to solve for,  $x$  is the concatenation of the reference and the test data,  $x_{t,j}$  is the  $j$ th element of the test data, and  $K_G(A, B)$  is the Gaussian kernel with width  $\sigma$ :

$$K_G(A, B) = \exp\left(-\frac{\|A - B\|^2}{2\sigma^2}\right)$$

Now solve for  $\alpha$  so that the empirical KL divergence between  $\hat{p}_t$  and  $p_t = p_r \hat{R}$  is minimized, which is equivalent to the following convex optimization problem:

$$\begin{cases} \max_{\alpha} & \sum_{j=1}^{n_t} \log\left(\sum_{k=1}^{n_t} \alpha_k K_G(x_{t,j}, x_{t,k})\right) \\ \text{s.t.} & \frac{1}{n_r} \sum_{j=1}^{n_r} \sum_{k=1}^{n_t} \alpha_k K_G(x_{r,j}, x_{t,k}) = 1 \\ & \text{and } \alpha_1 \dots \alpha_{n_t} \geq 1 \end{cases}$$

Finally, the scores that we wish to generate are just the estimate of the importance given by the solution to the complex optimization problem, i.e.  $s_i = \hat{R}_i$ .

Since this approach uses a Gaussian kernel, it requires the selection of a kernel width  $\sigma$  for each time tick. We used an implementation of KLIEP that is available at Sugiyama's website, which included a cross-validation procedure for the value of  $\sigma$ . The CV procedure chooses a number of disjoint splits of the test data along with a number of different candidate  $\sigma$ 's, and runs KLIEP with each combination of split and candidate  $\sigma$ . Then it chooses the candidate  $\sigma$  that, on the average across all of the splits, maximizes the KL divergence (the  $\max_{\alpha}$  equation above) the most.

For the OSU Hip dataset, we used this CV to choose the the kernel width at each individual time tick. This computationally- intensive approach was impractical for the UQ dataset because it is orders of magnitude larger, so instead of running it on every tick of that data, we ran the CV procedure on a number of random ticks drawn from the data. From this we were able to empirically identify 0.01 as a plausible  $\sigma$ , and so fixed  $\sigma$  at that value for our experiments on this dataset.

Previous research [14] found that on the average a window size of 10 seconds contained just enough information to discriminate well between OSU Hip activities. Since this window size worked well in previous experiments, and since the activities of the UQ dataset were comparable in average length, we decided to fix our reference window size at 10 seconds for both datasets. Because we were interested in minimizing detection time, and because 1 second was the smallest window that we felt could provide some

information about an activity, we fixed our test window size at 1 second for both datasets.

Once scores were generated, we chose a number of threshold values to determine which scores were considered high enough to predict a changepoint. Threshold values were chosen by considering the false positive rates of change prediction for the change point detection algorithms. A smaller false positive rate corresponded to a higher and more conservative threshold, which split the time series into fewer segments for featurization. A larger false positive rate corresponded to a lower threshold, which split the time series into more segments.

## 2.7 Bottom-Up Approach

## Chapter 3: Results

Figure 3.1, shows accuracy and detection time as a function of the false positive rate per second, for the OSU Hip dataset, for each of the SVM, Decision Tree, and Neural Net base classifiers.

### 3.1 Discussion

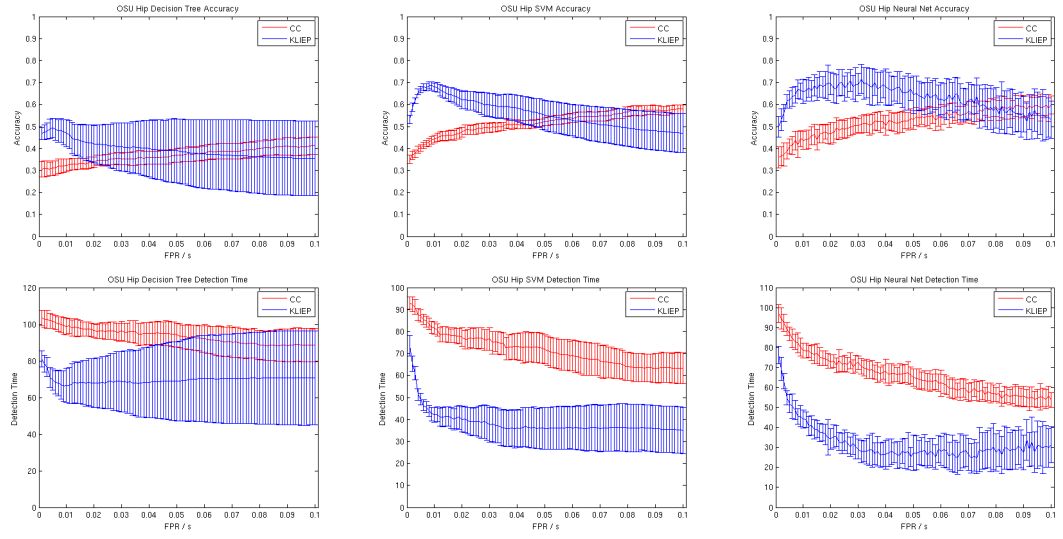


Figure 3.1: CPD-Based Classification Performance

## Chapter 4: Conclusion

### 4.1 Summary

### 4.2 Directions for Further Research

## Bibliography

- [1] Suk Joo Bae, Byeong Min Mun, and Kyung Yong Kim. Change-point detection in failure intensity: A case study with repairable artillery systems. *Computers and Industrial Engineering*, 64:11–18, January 2013.
- [2] Yoshinobu Kawahara and Masashi Sugiyama. Change-point detection in time-series data by direct density-ratio estimation. *Proceedings of the SIAM International Conference on Data Mining*, pages 389–300, 2009.
- [3] Z. Li. Exercises intensity estimation based on the physical activities healthcare system. In *Communications and Mobile Computing, 2009. CMC’09. WRI International Conference on*, volume 3, pages 132–136. IEEE, 2009.
- [4] David S. Matteson and Nicholas A. James. A nonparametric approach for multiple change point analysis of multivariate data. 2012.
- [5] V. Moskvina and A. A. Zhigjavsky. An algorithm based on singular-spectrum analysis for change-point detection. *Communication in Statistics. Statistics and Simulations*, 32:319–352, 2003.
- [6] M.P. Rothney, M. Neumann, A. Béziat, and K.Y. Chen. An artificial neural network model of energy expenditure using nonintegrated acceleration signals. *Journal of Applied Physiology*, 103(4):1419–1427, 2007.
- [7] W. Shewhart. Quality control charts. *Bell System Technical Journal*, pages 593–603, 1926.
- [8] M. Staudacher and all. A new method for change-point detection developed for on-line analysis of the heart beat variability during sleep. *Statistical Mechanics and its Applications*, 349:582–596, April 2005.
- [9] J. Staudenmeyer, D. Pober, S. Crouter, D. Bassett, and P. Freedson. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *Journal of Applied Physiology*, pages 1300–1307, 2009.
- [10] Masashi Sugiyama and all. Direct importance estimation with model selection and its application to covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 2008.

- [11] A.G Tartakovsky and all. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE Transactions on Signal Processing*, 54:3372–3382, September 2006.
- [12] G. Thatte and all. Parametric methods for anomaly detection in aggregate traffic. *IEEE/ACM Transactions on Networking*, 19:512–519, April 2011.
- [13] S.G. Trost, W.K. Wong, K.A. Pfeiffer, and Y. Zheng. Artificial neural networks to predict activity type and energy expenditure in youth. *Medicine and Science in Sports and Exercise*, pages 1801–1809, September 2012.
- [14] Yonglei Zheng. Predicting activity type from accelerometer data. Master’s thesis, Oregon State University, August 2012.



