Detecting short time-duration physical activity through statistical modelling of accelerometry data

M. Tadeusiak, J. D. Amor, V. Ahanathapillai and C. J. James
Institute of Digital Healthcare - WMG, University of Warwick, Coventry, CV4 7AL, UK

Abstract

While populations are ageing the development of technology to assist the elderly in maintaining their independence and daily activities becomes important. The majority of in-home accidents are caused by falls of which most occur during postural transitions. This paper presents the study of postural transitions recognition based on recordings of accelerations from a wrist-worn device. The Continuous Profile Model was used in order to obtain the patterns of activities such as: sit-to-stand transition, stand-to-sit transition and walking. Then, a set of training/testing routines were proceeded to assess the accuracy of the algorithm. Finally, the classification of both the regualarised and the naturalised data was successfully performed.

Keywords: *accelerometer, wrist-worn, postural transitions, activity monitoring, classification.*

1 Introduction

According to Eurostat [1], in the EU, the share of the total population aged 65 years or over is projected to increase from 17.1% in 2008 to 23.5% in 2030. In 2050 situation is predicted to be even more dramatic, specifically 16.4% of the world population and 27.6% of the European population are projected to be 65 years and above, and in 14 countries, including nine European ones, more than 10% of the total population will be 80 years or older [2]. Therefore the development of the technology to assist elderly in maintaining their independence and daily activities becomes crucially important.

In England and Wales some 260,000 people aged over 65 attend Accident and Emergency Departments annually due to in-home accidents [3]. The majority of are caused by falls, which makes them the principal threat of health and independence of the elderly. Studies regarding causes of falls and the assessment of the risk of falling are important as they may help to develop appropriate countermeasures.

Nyberg and Gustafson [4] report that the risk of falls is very high among stroke patients. Most falls occur during transfers, when initiating walking or while changing position from sitting to standing or vice versa. Also, Najafi et al. [5] finds the correlation between the length of the posture transition and the risk of fall.

This work is related to the USEFIL project [6] set up in 2011 that aims to develop advanced but affordable in-home unobtrusive activity monitoring solutions. The objective is to use low cost "off-the-shelf" technology and focus on the software meant to run on open-source platforms.

The aim of this study is to determine if the postural transitions can be accurately detected by tri-axial accelerometer-based device worn on wrist. We decided to collect data using a wrist-worn device as it is considered to be the least obtrusive location and generally preferred by the elderly [5]. By analysing the acceleration we are trying to pick up specific moves of the wrist indicating the transition. Our goal is to build a classifier that is able to distinguish between 3 different classes of activities, namely: sit-to-stand transition, stand-to-sit transition and walking. We apply the Continuous Profile Model (CPM) in order to obtain patterns of the activities and to perform classification. We assess its performance by applying several training/testing procedures.

The paper is organised as follows: Section 2 describes the experimental setup and the data collection. Section 3 provides description of training/testing procedures and introduction of CPM. Section 4 presents the results and the last sections focus on discussion and conclusions.

2 Methodology

In order to monitor the activity, especially to be able to detect the sitting-to-standing and the standing-to-sitting transitions, first, a set of specific exercises needs to be recorded. Time series of activity were recorded in a controlled environment and consist of both the regularised and more naturalised behaviour (defined in Section 2.2).

2.1 Device specification

The data were collected with the Z1 Android Watch-Phone (Figure 1) which included a built-in 3 axis accelerometer worn on the wrist. The device runs on an Android 2.2 platform, the specification is: 416 MHz, 256 MB RAM, 8 GB internal memory, which is sufficient for our studies. The recordings were transfered to the computer by USB.

Figure 1: The Z1 Android Watch-Phone.

2.2 Experimental setup and data collection

The data were collected from 15 healthy subjects, 12 males and 3 females. The age ranged between 22 and 37 years, the height varied from 156 to 187 cm, and the weight was between 51 to 100 kg. The device was worn on the left wrist. The participants were asked to perform 3 activities:

1. 10 repetitions of stand-to-sit-to-stand transitions. The time series were split afterwards into two sets consisting of the stand-to-sit (StSi) and sit-to-stand (SiSt) transitions separately;
2. 2 minutes of casual walking;
3. 30 seconds of casual walking followed by StSi and SiSt transitions, all repeated 3 times.

The transitions were evoked by the beep sound emitted every 5 seconds. Although the change of posture takes less, a single recording lasts 5 seconds or 250 samples. The first two sessions were designed to provide the recordings of patterns, being used for training purposes. These recordings are often refer to as 'regularised'. The 3^{rd} set will be used for testing purposes. We will call it the 'naturalised' data. Also, in next sections when classifying we will be assigning activity into classes as follows:

class 1 : sit-to-stand transition (SiSt);

class 2 : stand-to-sit transition (StSi);

class 3 : walking.

3 Data analysis

In this section we present the preprocessing of the data and the training/testing procedures that allow us to assess the performance of different classifiers. Then we describe the Continuous Profile Model and mention the underlying mathematics. Next, we introduce the classifiers we apply to detect activities. Finally, we discuss some measures we employ to asses the accuracy.

3.1 Extraction of samples

The sampling rate of the device was set to 50 Hz. However, as the device wasn't designed for precise measurements and the resulting sampling frequency could slightly vary, the recordings were resampled by software afterwards. We assume 50 Hz is sufficient to record detailed signatures of transitions.

Figure 2 shows examples of series we have recorded. Recordings of transitions are divided into 5 seconds long intervals. StSi and SiSt transitions appear alternately, therefore when extracting, every second sample is taken. Although the length of the original recordings amounts to 5 seconds, the actual transition normally takes about 2 seconds. We have some arbitrariness in choosing only part of the recording for future analysis and in the further studies we are considering series of length: 2.4, 3.2, 4.0 and 5.0 seconds. When obtaining shortened samples we use a window of a chosen length and select a position in order to cover the entire transition. In example depicted in Figure 2, a 2.4 second long window and 0.8 second delay was set. As the reaction time on a beep sound varies for different persons, the offset was set individually. This ensures that the all the transitions are included.

The recordings of walking are continuous, therefore in order to obtain samples a sliding window was used. The length of the window was the same as when extracting samples of transitions. For training purposes no overlap was imposed on the sliding window, while for testing purposes it varied from 50% to 90%.

Due to some artifacts in the recordings (e.g. pulling up a sleeve) some series were excluded in order to provide consistent input for training.

3.2 Training/testing procedures

We split the data into separate training and testing sets in order to evaluate the performance of the algorithm. We apply four different procedures, each of which answers different questions about the performance of the algorithm and deliver valuable information of its limitations. We discuss their significance broadly in Section 5.

3.2.1 Regularised data

Both the training and testing were performed on the data collected in step 1 and 2 (Section 2.2). It can be considered as more regularised data, as all the posture transitions were repeated in regular manner paced by the beep sounds every 5 seconds.

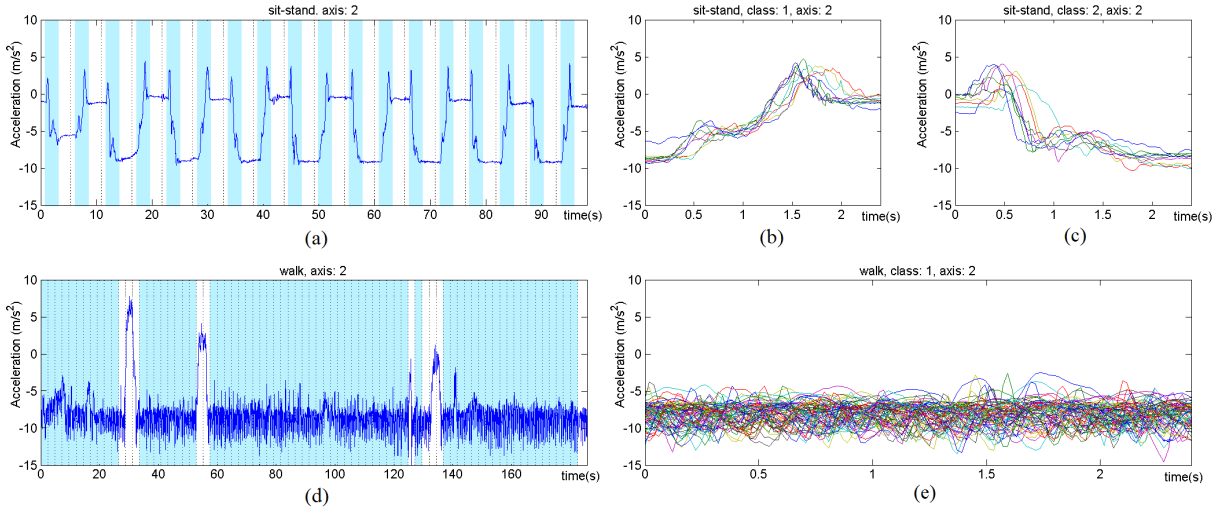


Figure 2: Activity recordings. Plots (a) and (d) show original recordings. Shading indicates which parts of signal where extracted. Plots (b) and (c) show samples of transitions obtained from recording in (a) by extracting a sequence from every second interval (each interval is separated by a beep sound). A window of length 2.4 seconds and an 0.8 seconds long offset was used in order to cover the entire transition. Plot (e) presents samples of walking extracted from (d) by sliding a window of the same length and with no overlap. Some parts where omitted due to artifacts.

• **Procedure 1.** Training on the regularised individual data/testing on the regularised individual data (Leave-one-out).

For each individual all-but-one samples are taken for training, the testing is performed on the remaining sample (Figure 3). The procedure is depicted in Figure 4a. Entire analysis was performed having used all 3 classifiers and for all window lengths: 2.4, 3.2, 4.0 and 5.0 seconds.

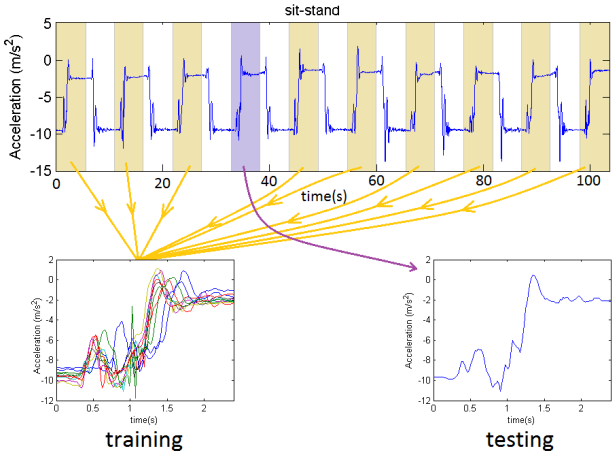


Figure 3: Leave-one-out method. For each person, all but one regularised series were used for training. Testing was performed on the remaining series.

• **Procedure 2.** Training on the regularised combined data/testing on the regularised individual data.

In this approach the regularised recordings over all the subjects were combined and split randomly into training/testing subsets in proportion $^2/_3$ to $^1/_3$. The approach is sketched in Figure 4b.

In order to assess consistency of received results, the procedure was repeated 10 times, each time, due to randomization, different combination of series were constituting training and testing sets. Final scores were averaged over received results.

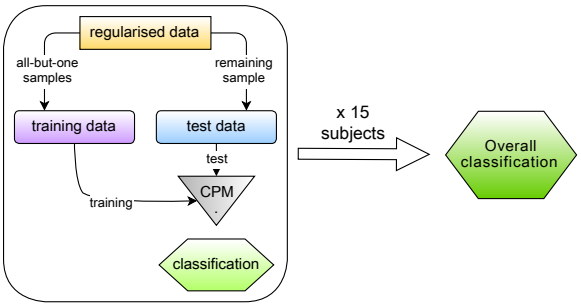
3.2.2 Naturalised data

When collecting naturalised data (Section 2.2), no strict rules about how and when to sit, while walking were imposed on subjects. Subjects were advised to sit down after 6 beep sounds (30 seconds of walking). However, as the location of the subject varied (sometimes they were closer to or further from the closest seat) they might have rushed or slowed down towards a seat. This behaviour causes two outcomes. First, 'naturalisation' of recordings is desirable. The second is more problematic, as there is no certain position we can expect a transition to appear in the recording. Therefore the manual inspection is needed in order to label each sample derived from a naturalised recording.

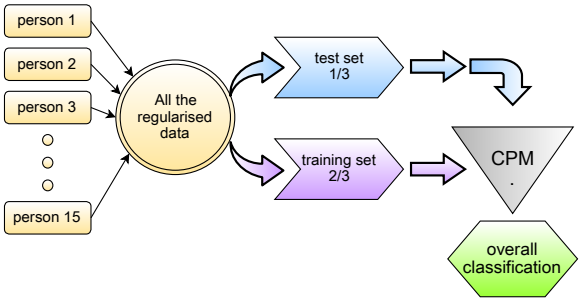
Samples were labelled according to their fitness to a proper class assessed visually. In case when no class could have been assigned clearly, the label '0' was attached and the sample was ignored during testing.

• **Procedure 3.** Training on the regularised individual data/testing on the naturalised individual data.

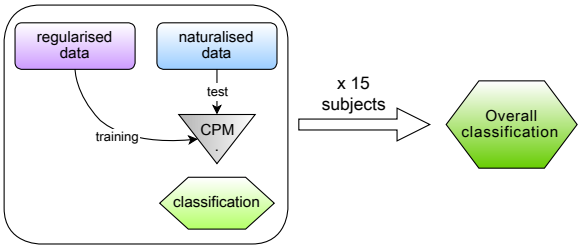
In this case, similarly to procedure 1, training is carried out individually on the regularised sets. Testing, in contrary, is based on naturalised individual recordings. Results are obtained for each person separately and then combined together for the overall assessment, see Figure 4c.



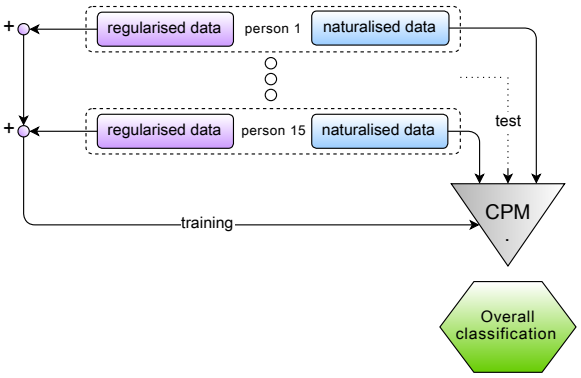
(a) Procedure 1. For each person, all but one regularised series were used for training. Testing was performed on the remaining one. The procedure was repeated for each series. Finally, results over all persons were combined.



(b) Procedure 2. All the regularised recordings were combined and split randomly into training/testing subsets in proportion $^2/_3$ to $^1/_3$.



(c) Procedure 3. Training is performed based on the regularised data and training uses the naturalised data. Results are obtained for each person separately and then combined together for the overall assessment.



(d) Procedure 4. All the regularised recordings are combined for training purposes. Testing is performed on the individual naturalised data. Results over all persons are eventually averaged.

Figure 4: The procedures used for training and testing.

• **Procedure 4.** Training on the regularised combined data/testing on the naturalised individual data.

Similarly to procedure 2, the model was trained on the data combined over all the subjects. The testing stage was performed on the individual naturalised data. After obtaining results for each subject, an average was calculated. Figure 4d illustrates the procedure.

3.3 Continuous Profile Model

Now, we introduce the algorithm we use both for training and testing, the Continuous Profile Model.

The time series can be aligned in time and scale by using CPM [8][9]. The assumption is that the series come from the same underlying process. In other worlds, each observed time series is considered to be a non-uniformly subsampled, noisy version of a latent trace. The trace can be obtained in a process of data-driven learning. Then the likelihood that a given series diverges from the learned trace can be calculated. The model was successfully applied to align speech signals [9] from multiple speakers and to investigate daily behaviour patterns [10].

CPM is based on the Hidden Markov Model (HMM) [11], see Figure 5. Each hidden state corresponds to a particular location in the latent trace. In CPM in one *time* step, the transition between states can be made only in one direction and the distance to the next state is limited to few *space* steps, although originally it can be arbitrary long.

We enforce the length of the observed time-series to be equally long (this constraint does not come from any mathematical restrictions, but due to implementation), and consists of N samples. Now, the length of the latent trace needs to be $M > N$ (ideally $M \gg N$ so that an observed sample could be mapped more precisely to the appropriate latent value). A local distortion in time scale can be accomplished by jumping to a next hidden state less than M/N steps away (slowing down) or to a one further than this distance (speeding up).

After each transition an observable being a value normally distributed around the sample of the latent trace corresponding to the current hidden state is emitted.

$$\mathcal{L} \equiv \sum_{k=1}^{K} \left[\log\, p(\tau_1) + \sum_{i=1}^{N} \log\, A_{\tau_i}(x_i^k|\overrightarrow{z}) + \sum_{i=2}^{N} \log\, T^k_{\tau_{i-1},\tau_i} \right] \tag{1}$$

$$\mathcal{P} \equiv -\lambda \sum_{j=1}^{M-1} (z_{j+1} - z_j)^2 + \chi \tag{2}$$

This method allows as to build a model capable of representing the set of input timeseries, accounting for local and global time shifts by altering the speed at which the model is traversed. Furthermore, the model can be used to analyse new time series by aligning them to the latent traces and assessing their compatibility.

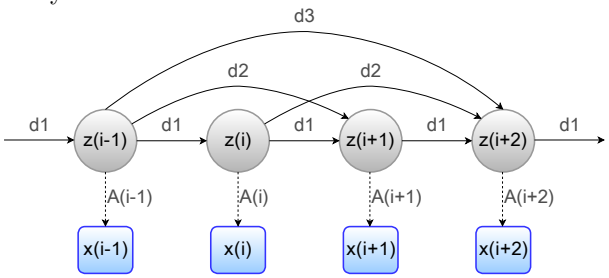


Figure 5: Simplified HMM: gray circles: hidden states corresponding to the latent trace, blue squares: emitted observables, solid arrows: transition probabilities between hidden states, dashed arrows: emission probabilities of observables (note: for sake of the transparency, transitions from/to states outside the diagram were omitted despite they existence).

3.3.1 Mathematical Model

Let $\overrightarrow{x}^k = (x_1^k, x_2^k, ..., x_N^k)$ be the k-th time series, $k = [1..K]$. In the experiments the sampling rate was uniform (50 Hz), although the CPM model introduces no constraints for its regularity.

Further, let $\overrightarrow{z} = (z_1, z_2, ..., z_M)$ be the latent trace, the noiseless, high resolution prototype of observed series $\overrightarrow{x}^k$. In our experiments we have set:

$$M = (2 + \epsilon)N \tag{3}$$

which is double the resolution, plus $\epsilon = 0.05$ for some room on ends.

For a time series to be modelled from the latent trace, a series of hidden states is needed. We call the sequence of hidden states corresponding to time series k, $\overrightarrow{\tau}^k \in [1..M]$ (we assume the activity we record, especially repetitions of the same exercises, to be in the same scale. Therefore, we omit the scale state terms in our further analysis, compare with [9]).

We model the observations x_i^k to be related to the states τ_i^k by the emission probability distribution:

$$A_{\tau_i^k} \equiv p(x_i^k|\tau_i^k, \overrightarrow{z}, \sigma) \equiv \mathcal{N}(x_i^k; z_{\tau_i^k}, \sigma) \tag{4}$$

where σ is the noise level of observed series and $\mathcal{N}(\xi; \mu, \varsigma)$ is a Gaussian probability density for ξ with mean μ and standard deviation ς.

We also need to define the transition probabilities between the states:

$$T^k_{\tau_{i-1},\tau_i} \equiv p^k(\tau_i|\tau_{i-1}) \tag{5}$$

As we allow only forward transitions and we set the jump length limit $J_\tau = 3$ we have:

$$p^k(\tau_i = a|\tau_{i-1} = b) = \begin{cases} d_1^k, & \text{if } a - b = 1 \\ d_2^k, & \text{if } a - b = 2 \\ d_3^k, & \text{if } a - b = 3 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Knowing the latent trace, the transition and emission probability distributions and a noise level, the CPM can be used to model the input time series.

3.3.2 Training

During training, the latent trace, the transition probabilities controlling the Markovian evolution of the time states and the overall noise level are learned. The training is performed with the Expectation-Maximization (EM) algorithm, which includes two steps.

E-step

In E-step the Forward-Backward (FB) algorithm [11] is applied that allows to calculate the probability of the observation sequence $\overrightarrow{x}^k$ given the model, i.e.

$$p(\overrightarrow{x}^k|A_{\tau_i^k}, T^k_{\tau_{i-1},\tau_i}, \sigma) \tag{7}$$

where the likelihood depends on probability distributions of emissions and transitions. For computational reasons, as values of likelihoods might be very small, the usual approach is to compute log-likelihoods. The complete log-likelihood of K observed time series $\overrightarrow{x}^k$, is given by $\mathcal{L}^p \equiv \mathcal{L} + \mathcal{P}$, of the form (1).

$\mathcal{L}$ is the likelihood term that can be obtained using FB algorithm, where $p(\tau_1)$ are priors over the initial states. $\mathcal{P}$ is a penalty term, where the first term is a smoothing penalty punishing too much variation between consecutive latent samples. The bigger the λ is the smoother the latent traces are. In our experiments λ was set to 20. Term χ stands for regulatory terms such as Dirichlet priors to the time transition probabilities so that all non-zero transition probabilities remain non-zero [9].

Based on $\mathcal{L}^p$, summing over all possible states for all input series the expected complete log-likelihood can be computed.

M-step
In M-step the above log-likelihood function (1) is maximised with respect to the parameters that need to be optimised, specifically: $\{z_j\}$, $\{d_v^k\}$, σ. The optimisation provides the best setting for the parameters and allows the HMM to be used to compute the alignment of the input time series.

Expectation-Maximization steps are repeated until convergence of the values being optimised.

3.3.3 Testing

During the testing, input time series are aligned by the CPM to previously learned latent traces. The alignment is based on the Viterbi algorithm [11][12]. The algorithm finds the best state sequence $\overrightarrow{\tau}^k$ for the observed time series $\overrightarrow{x}^k$, linking the observed sequence to the latent trace. The fitness of the aligned series to the latent traces is assessed using different methods described in Section 3.5.1.

3.4 The implementation

In computations we use the CPM toolbox for Matlab[1], developed by Jennifer Listgarten [8]. The toolbox consists of scripts allowing us to apply EM procedure to obtain patterns, align the samples with the Viterbi algorithm and calculate the related log-likelihoods. During the project, whole set of additional functions was implemented to efficiently make use of these tools.

3.5 Experimental Analysis

While training, the set of series corresponding to a single class (the type of activity, see 2.2) was an input, returning the latent trace and related parameters as an output. Figure 6 shows an example of outcomes obtained while training. Among the obtained traces and examples of alignment we can see also the convergence to the optimal latent trace measured by the mean log-likelihood of the fitness of input series to the received trace. These log-likelihoods values vary for different inputs. Nevertheless, as the input series are derived from the same underlaying process we expect them to be similar. Also, the latent trace can be seen as a higher resolution fusion of them, thus it should resemble the inputs. Therefore, we expect the likelihoods of input series fitting to the latent trace to be approximately even. At the same time, as the values of likelihoods are based on transition and emission probabilities and other penalties (1), we cannot expect the values related to different classes to be on the same level. Indeed, the distributions of likelihoods (Figure 7) are concentrated, but the their mean values differ.

[1]the CPM implementation for Matlab `http://www.cs.toronto.edu/~jenn/CPM/`

3.5.1 Classification

When assigning series to the classes based on the latent traces obtained after training, 3 methods were used:

1. **residuals**: first, a sequence is aligned to the latent trace by using the Viterbi algorithm [11]. As the resulting series may not be of the same length (thanks to a non-uniform alignment) as the latent trace (Figure 8), special measure needs to be applied in order to assess the fitness of the sample to a certain class.

   The disparity between a sample $\overrightarrow{x}^k$ and a latent trace $\overrightarrow{z}$ is measured by the average of squared differences between their overlapping parts. The non-overlapping part contributes to the disparity score to an extent depending on its length. The overall error is calcluated according to the formula:

   $$\mathcal{E}_k^C = \sqrt{\frac{1}{|\Omega_O|}\sum_{i\in\Omega_O}(x_i^k - z_i^C)^2} \cdot \left(1 + \frac{|\Omega_{NO}|}{|\Omega_O|}\right) \tag{8}$$

   where $\mathcal{E}_k^C$ is a measure of the difference between series k and the pattern of class C, x_i^k is i^{th} sample of series k, z_i^C is a corresponding sample from the latent trace of class C, Ω_O is set of indices related to the overlap between $\overrightarrow{x}^k$ and $\overrightarrow{z}^C$, Ω_{NO} is set of indices related to the non-overlap between them, and $|\cdot|$ is the number of entries in a set.

   When classifying a sample k, $\mathcal{E}_k^C$ is calculated for each class in order to find the minimal discrepancy;

2. **loglikelihoods**: these values come as a direct output from CPM. Based on the forward-backward algorithm, the log-likelihood that an observed sequence k was produced by the model corresponding to class C is calculated and we denote it as ϕ_k^C. The calculation is run for each class and the sequence is assigned to the class of the maximum score.

3. **evened loglikelihoods**: the loglikelihoods received via method 2 are shifted to bring the means together and make them more comparable. Using the mean values of the normal distributions based on the loglikelihoods computed while training (see Figure 7), the evened log-likelihoods are given by:

   $$\Phi_k^C = \phi_k^C + \Delta\phi_k^C \tag{9}$$

   where the size of the shift is given by:

   $$\Delta\phi_k^C = \eta^C - \overline{\eta} \tag{10}$$

   where η^C is the mean loglikelihood corresponding to class C (the centre of the distributions seen in Figure 7), $\overline{\eta}$ is the average over all η^C.

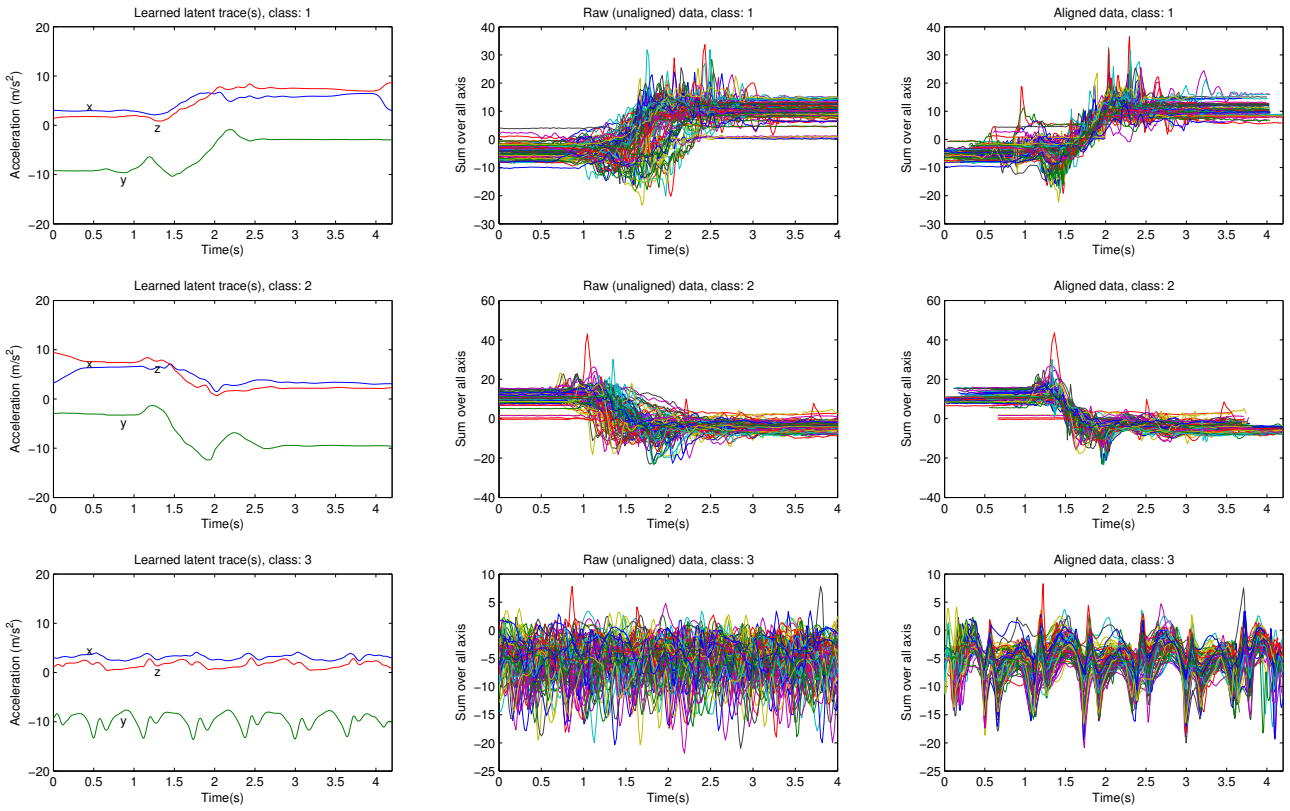


Figure 6: The outcome of training performed by CPM. Each row is related to a different class, from the top: sit-to-stand transition, stand-to-sit transition, walking. In columns from the left: the obtained latent traces; raw input series; the input series aligned with the Viterbi method. Take notice of alignment of 'walking' recordings (bottom right).

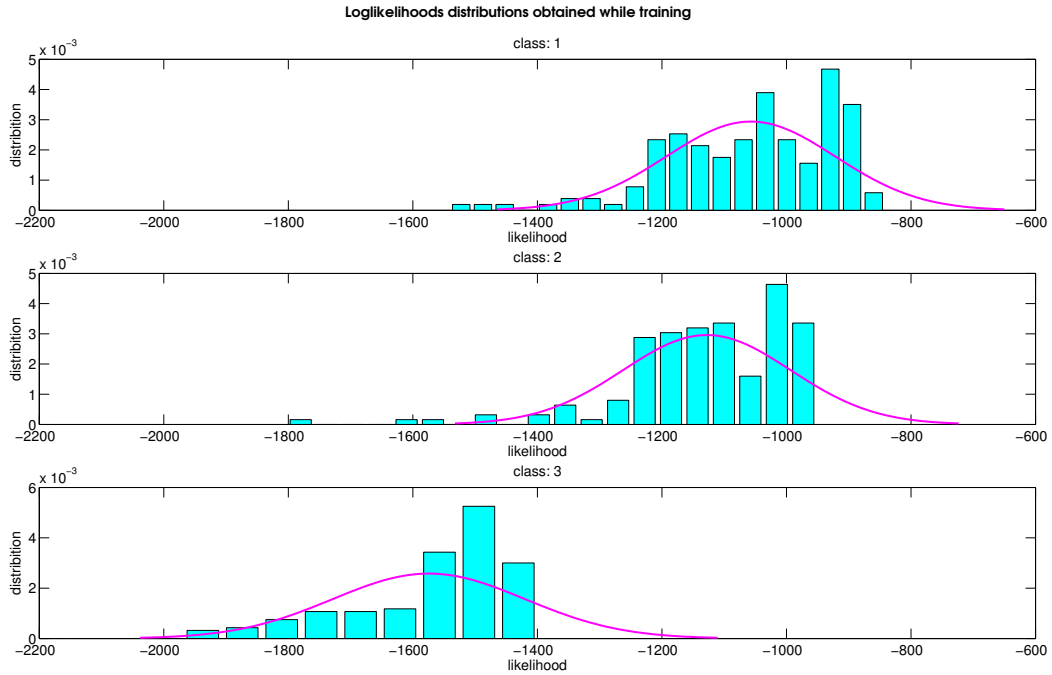
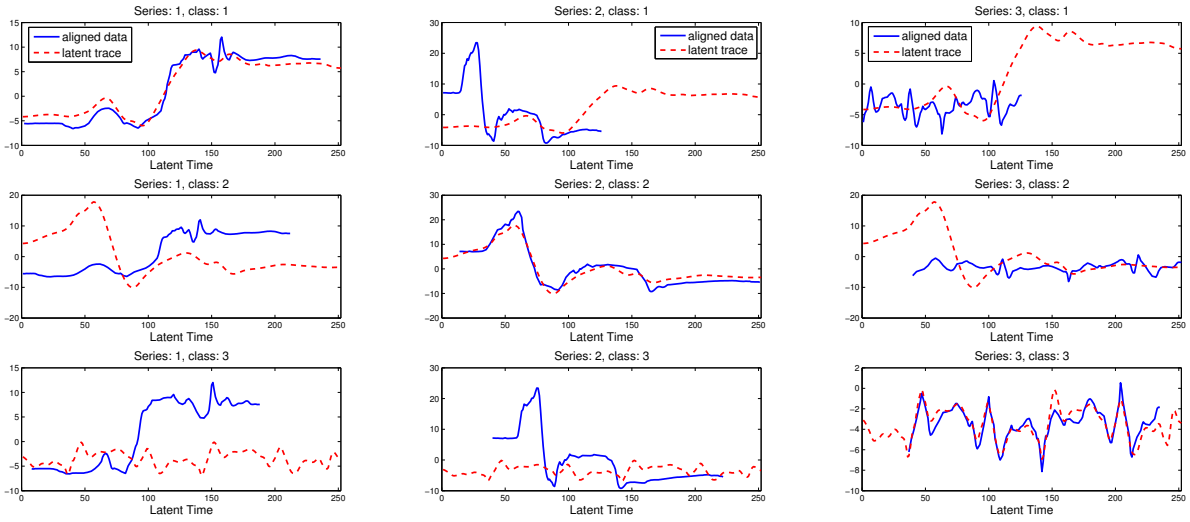
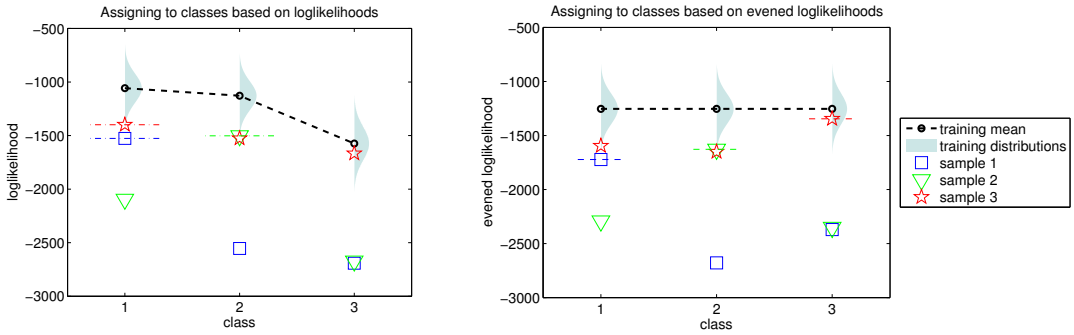


Figure 7: Distributions of likelihoods corresponding to 3 different classes obtained while training. The pink curve is an adjusted normal distribution. The distribution for class 3 is clearly shifted comparing to others.



(a) Sample from class 1. (b) Sample from class 2. (c) Sample from class 3.

Figure 8: Method of residuals. Solid blue line: aligned series, red dashed line: latent traces. Three samples corresponding to each class: StSi transition, SiSt transition and walking (columns, respectively) are aligned to every latent trace (rows). According to expectations we can observe the best alignment on diagonal. Particularly, in (b) we can inspect visually the best match of a recording of SiSt transition to class 2. In this case, the values of errors $\mathcal{E}_2^C$ returned by the method of 'residuals' are: {14.3, 2.3, 10.3}. The lowest value correctly corresponds to class 2.



(a) Loglikelihoods (b) Evened loglikelihoods

Figure 9: Illustration of classification to the proper class based on log-likelihoods. Figures show log-likelihoods that a given sample fits to one of the classes. In this example, i^{th} sample corresponds to i^{th} class. A sample is classified to the class of the score. In case (a), sample 3 is not properly assigned due to differences in scales. In case (b), all the values are shifted to bring the means of distributions together, which implies proper assignment.

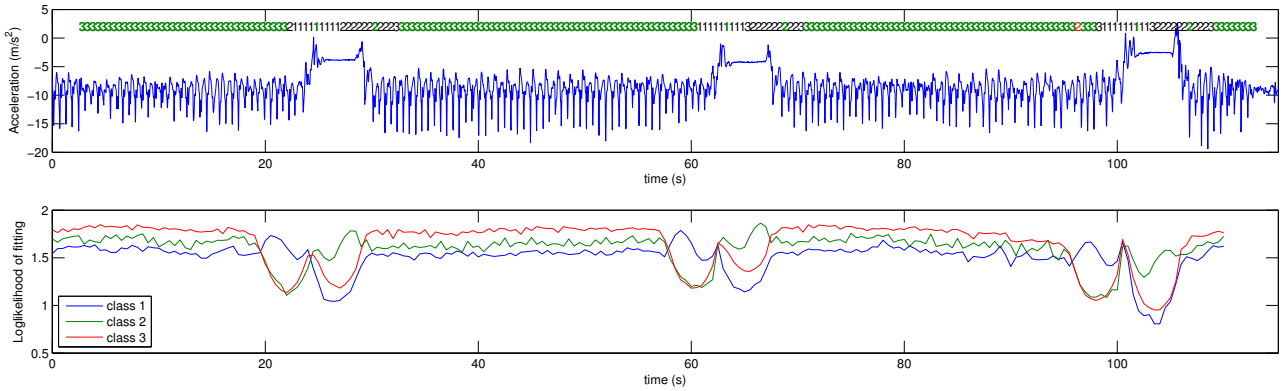


Figure 10: The example of 'naturalised' data with the classification performed by the 'evened log-likelihoods' classifier. The upper plot shows the original recording (we can recognize transitions) and values indicating the class each sample was assigned to. Values in green are assigned properly; values in red signalise mistake; values in black are ignored due to their ambiguity (see comment in Section 3.2.2). The lower plot shows the values returned by the classifier for each class. A sample is assigned to the class of the maximal score.

Figure 9 illustrates the assignment of three samples corresponding to three different classes respectively. For each sample the probability of fitting to a particular class is computed by using methods 2 and 3. A sample is classified to the most probable class.

3.5.2 Assessment of performance

In order to assess which classifier performs better we need to have a set of labelled samples to use as the test data. Due to shortage of time dedicated for the project we use a simple prototype where labels are assigned manually. As the samples were extracted from the original recording by a sliding window with an overlap of 90%, many of them contain some parts of transitions. The main idea is to label only those samples for which there is no ambiguity to which activity they correspond. This is assessed visually for each sample. For each transition only one, the most obvious sequence is labelled according to its class, all the neighbouring series are labelled as '0' which excludes them from further analysis.

The upper plot in Figure 10 shows the original recording and labels assigned to each of extracted sequences of length 5 seconds (the boundaries are not shown for better clarity). Labels in green indicate correct assignments, in red - mistakes, the black ones, manually labelled as '0', are ignored. In this case, classifier 3 was used. In the lower plot of Figure 10 we can observe the evolution of evened log-likelihoods for each class. We can clearly see that values related to classes 1 and 2 increase close to transitions, while for class 3 they suddenly drop.

Although, this method of assessment of classifiers is not perfect - we can imagine a case when the transition is performed in a peculiar way so it is partly assigned correct and partly incorrectly. It may happen that the sample chosen manually would be labelled wrongly, while the neighbours correctly. In this case the assignment would be counted as an error, while it shouldn't be.

The problem of assessing the classifier based on multiple assignment of overlapping series can be a topic for a long discussion and it is important problem to be solved if this algorithm was to be applied in real life.

3.5.3 Accuracy Measures

While assessing the performance of algorithm multiple measures were used. Accuracy is measured by:

$$P_k = \frac{TP_k}{TP_k + FP_k} \tag{11}$$

$$R_k = \frac{TP_k}{TP_k + FN_k} \tag{12}$$

where: P is *precision*, R is *recall*, TP, FP, TN, FN are the number of *true positive*, *false positive*, *true negative*, *false negative* classifications respectively. Index k indicates the affiliation to a certain class.

We can interpret *precision* as the fraction of correct classifications relative to all the classifications to a particular class. *Recall* would be the fraction of correct classifications compared to all the real samples of a particular class.

We also introduce the F_1 score:

$$F_{1k} = 2\frac{P_k \cdot R_k}{P_k + R_k} \tag{13}$$

F_1 is a harmonic mean of *precision* and *recall*, which means it is biased towards the smaller of them. We will often use this useful measure.

It is worth to notice that when there is no sample assigned to a class, both TP and FP are zero and precision is undefined $\frac{0}{0}$, thus we receive 'NaN'. As during testing, there is always some number of samples from each class, it is clearly a mistake. In this case, while calculating averages, we consider these scores as zeros.

Unfortunately, the measures defined as above can be misleading when the number of samples in classes is not equal. The common case in our analysis is that number of 'walking' samples significantly exceeds the number of 'StSi' and 'SiSt' samples. Often, we observe the situation when many samples of class 3 (walking) are assigned as class 1 (StSi) which considerably influences precision related to class 1, but only slightly affects recall of class 3.

In order to eliminate this bias, weights related to the number of samples need to be introduced:[2]

First, we define the *weighted confusion matrix*, $C = (c_{ij})$, where $i, j \in \{1, 2, 3\}$ correspond to classes: StSi, SiSt, walking, respectively. c_{ij} is the weighted number of assignments of samples from class i (reference) to class j (prediction). The example for class 1 is shown in Figure 11.

Each entry of matrix C is weighted according to the number of samples from the respective class, such as:

$$c_{ij} = a_{ij}\omega_i \tag{14}$$

where a_{ij} is actual number of assignments of samples from class i to class j and a weight:

$$\omega_i = 1 - \frac{N_i}{N} \tag{15}$$

where N_i is the number of samples from class i and N is sum of them. Now we can define weighted measures:

$$TP_k = c_{kk} = a_{kk}\omega_k \tag{16}$$

$$FP_k = \sum_{i \neq k} c_{ik} = \sum_i c_{ik} - TP_k \tag{17}$$

$$FN_k = \sum_{j \neq k} c_{kj} = \sum_j c_{kj} - TP_k \tag{18}$$

$$TN_k = \sum_{\substack{i \neq k \\ j \neq k}} c_{ij} = \sum_{i,j} c_{ij} - TP_k - FP_k - FN_k \tag{19}$$

[2]Other, easier solution would be to even the number of samples of each class, regrettably because of the classification method we use this solution cannot be applied.

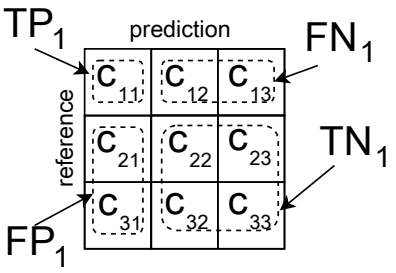


Figure 11: Confusion matrix.

We also define the *normalised confusion matrix*, $D = (d_{ij})$. Let a_{ij}^l be the number of assignments of samples from class i to class j for person l. Also, let N_i^l be the number of samples from class i for person l. d_{ij} is defined as the sum of assignments a_{ij}^l over all the individuals, normalized with respect to the overall number of samples in each class:

$$d_{ij} = \frac{\sum_{l=1}^{L} a_{ij}^l}{\sum_{l=1}^{L} N_i^l} \tag{20}$$

The difference between these two confusion matrices is that, the *weighted CM* is modified for each individual separately, while the *normalised CM* gives a global measure of confusion between classes.

4 Results

All 4 procedures were run for 4 different window lengths. Each time 3 classifiers were applied. For every setting we produce a table with the individual and averaged precision, recall and F_1 scores for each class, confusion matrix and also a box plot presenting the distribution of individual results. It gives altogether 48 collections of tables, matrices and plots. Therefore it is reasonable to focus only on the most relevant cases. Also, to give a general measure of the accuracy of classification in certain cases, we will use the doubly averaged F_1 score: first we average over all subjects receiving means related to classes, then we average once more over the classes (Table 3 is a good illustration, we receive the value in the bottom right corner as the result). From now on we will refer to this value as 'averaged F_1 score'.

4.1 Procedure 1

In this procedure training and testing were run on regularised data in 'leave-one-out' manner. As we observe nearly zero misclassifications, we present only the averaged F_1 scores for all methods versus the window length (Figure 12(a)). Method of residuals return 100% accuracy for every window length for every class. Method of log-likelihoods perform slightly worse, but still close to 100%. Method of evened log-likelihoods returns the best averaged F_1 score for the shortest window, namely: 99.09% and declines to 93.93% when the window lengthen.

4.2 Procedure 2

Training was performed on $^2/_3$ of all the regularised data and testing on the remaining $^1/_3$. The summary of the results is presented in Figure 12(b). Methods of residuals and evened log-likelihoods perform satisfactory, close to 97-98% regardless the window length. The method of 'log-likelihoods' stays below 70%, only to jump to 85% for the longest window. The best results were obtained for 4 seconds-long window by classifier 3 (evened log-likelihoods) with F_1 scores reaching: $\{99.52\%, 96.89\%, 97.85\%\}$ for each class respectively.

4.3 Procedure 3

The CPM was trained on the individual regularised data and tested on the individual naturalised data. The averaged F_1 scores (Figure 12(c)) are low, majority under 80%. The classification improves for the shortest window. The best performer - method of 'residuals' reaches the highest result around 83%, however, the individual results differ considerably. Among different subjects, they range from 'NaN' (we consider these scores as zeros) to 100%, see Figure 13.

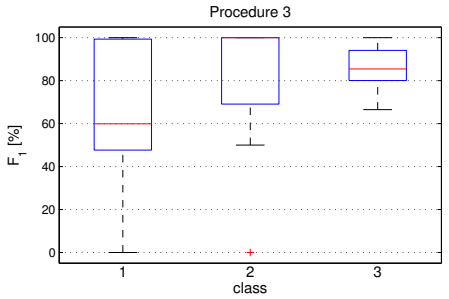


Figure 13: Distribution of individual F_1 scores obtained by the method of 'residuals' for a window length 3.2 seconds in procedure 3. 'NaN' values were considered as '0'.

4.4 Procedure 4

Training was performed on combined regularised data, testing on individual naturalised recordings. Averaged F_1 scores are shown in Figure 12(d). The best results are returned by method of 'evened log-likelihoods' for the window length 3.2 seconds - the averaged F_1 score reached 92.7%. On page 12, we present full set of results related to this case. Table 1 presents the *precision*, *recall* and F_1 averaged over all persons. The values range between 84.77% and 98.60% among all classes.

However, the mean is not a robust estimator as it can be considerably biased by outliers. In this case, the median is more appropriate. In Figure 14 we can see box plots presenting the distributions of scores among subjects. The median of all results is 100% and general performance is quite impressive. Nevertheless, there exist some outliers - individuals for whom the classification comes out very poorly, in worst case

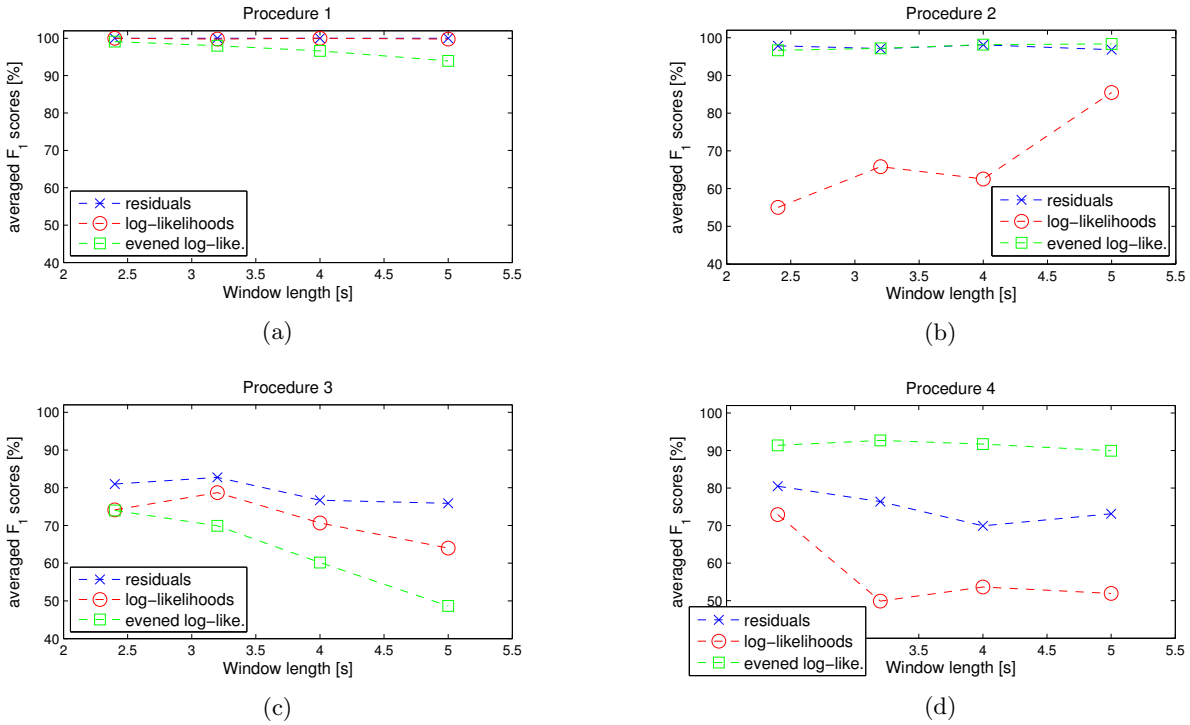


Figure 12: The averaged F_1 scores for all methods in relation to the window length. Methods and markers: 'residuals' - blue crosses, 'log-likelihoods' - red circles, 'evened log-likelihoods' - green squares.

with the F_1 score about 16.84% (person number 14 in Table 3), although in contrast to procedure 3, this time no 'NaN' values appear.

Moreover, the confusion matrix (Table 2) presents the misclassification between classes. The results obtained are not far from the optimum as all the diagonal values are above 90%. As our goal is to detect transitions, we concern ourselves most with the correct classification of class 1 and 2, thus it is particularly admirable that there is no confusion between them.

5 Discussion

Procedure 1 provides us with very good results (Figure 12(a)). Almost all the assignments are correct. CPM seems to be performing very well when dealing with regularised data, provided that it is trained on recordings specific to a particular person.

In contrast to training and testing on each individual separately, in procedure 2 the training is performed on all the data combined together. It should help to asses how accurate the classification is when based on the general patterns.

Although slightly worse than in procedure 1, the scores are still excellent. It is good news considering possible future implementations. It means that algorithm could be suitable for real world use with no user training period. It is important as a device ready to measure the activity straight after being put on the wrist, without additional calibration, would be most desirable.

Procedure 1 and 2 provide results obtained in arguably artificial way as the testing was performed on the regularised data, therefore the variability of samples were significantly suppressed. The next approaches focus on more relevant case, when the testing is performed on the naturalised data.

In procedure 3, although the highest average F_1 score amounts to 83%, the performance of the model is not reliable in a sense that results for different individuals can differ extremely. Comparing to procedure 1, when the training set was almost equal (reduced by one sample being left out, altered each time) the performance is mediocre. We suppose that it is because of over-fitting to the regularised data, in which case another regularised sample should be recognised properly, while a naturalised sample, due to its irregularity, might be assigned wrong.

As the over-fitting might be the case when testing on naturalised data, in procedure 4 we train on combined data. The idea is that inter-person variability may, to some extent, play a role of intra-person variability and therefore diminish the over-fitness.

In procedure 4, we observe a significant improvement comparing to procedure 3, especially taking into account that F_1 scores are stable along different window lengths. The accuracy for majority of persons reaches 100%, on average scores are close to 90% and there is no confusion between transitions.

The results based on naturalised recordings, which we consider the most relevant, are optimistic. The CPM seems to work satisfactory. However, we are aware that there exist factors that could have affected the performance of the model and that there are limitations unrevealed in this project.

class	precision	recall	F1
1	84.77%	95.65%	**89.88%**
2	98.60%	93.33%	**95.89%**
3	94.21%	90.55%	**92.34%**

Table 1: Accuracy table returned by classifier 3 for window length 3.2 s in procedure 4. Precision, recall and F_1 scores are averaged over all subjects.

class		assignment		
		1	2	3
reference	1	**95.7%**	0.0%	4.4%
	2	0.0%	**93.3%**	6.7%
	3	9.0%	0.7%	**90.3%**
number of samples		46	45	3871

Table 2: Normalised confusion matrix returned by classifier 3 for window length 3.2 s in procedure 4. Shows the number of classifications divided by the real number of samples from a certain class (presented in the bottom row). The diagonal corresponds to correct assignments.

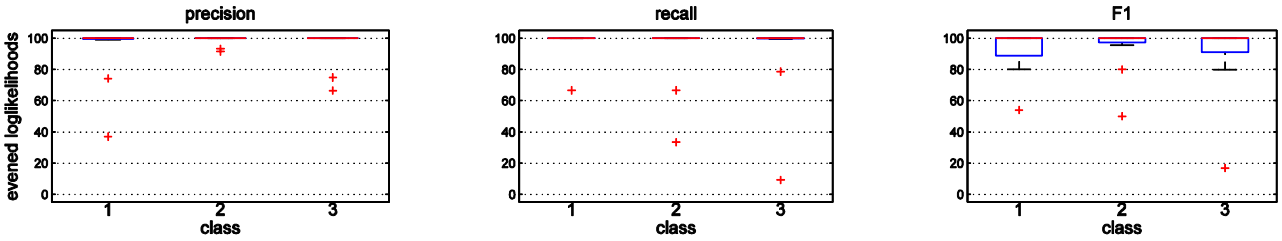


Figure 14: Boxplots of individual results returned by classifier 3 for window length 3.2 s in procedure 4. On each box, the central red mark is the median, the edges of the box are the 25^{th} and 75^{th} percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually as '+'.

subject	class 1			class 2			class 3			average
	precision	recall	F1	precision	recall	F1	precision	recall	F1	F1
1	100.00	66.67	80.00	100.00	66.67	80.00	74.80	100.00	85.58	81.86
2	100.00	66.67	80.00	100.00	33.33	50.00	66.43	100.00	79.83	69.94
3	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
4	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
5	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
6	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
7	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
8	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
9	99.25	100.00	99.62	100.00	100.00	100.00	100.00	99.62	99.81	99.81
10	74.15	100.00	85.16	93.16	100.00	96.46	100.00	78.60	88.02	89.88
11	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
12	99.18	100.00	99.59	100.00	100.00	100.00	100.00	99.58	99.79	99.79
13	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
14	37.00	100.00	54.02	91.44	100.00	95.53	100.00	9.19	16.84	55.46
15	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
average	84.77	95.65	89.88	98.60	93.33	95.89	94.21	90.55	92.34	**92.70**

Table 3: Individual results returned by classifier 3 for window length 3.2 s in procedure 4. The table shows individual results for each subject and the averaged values in the last row. For each class precision, recall and F_1 scores are shown. The last column is individually averaged F_1 score over all classes (NaN's were considered as zero's). The value in the bottom right corner is averaged F_1 score over all subjects and classes and it corresponds to the value being called the 'averaged F_1 score'.

First, the number of 100% scores should enhance our vigilance. We recognize two reasons for such numerous fraction of the highest possible scores. The first reason, evident in procedure one, is related to the exceptional regularity of recordings making the training and testing sets too similar to leave the chance for a mistake (unless one of the sequences vary significantly from the others). Therefore, as we already have signalised, we don't consider the first procedure to be very contributive to the accuracy assessment. The second reason, apparent in procedure 4, is connected with insufficient number of series while testing. There are only 3 recordings of each transition for each individual (15) which sum to 45 samples altogether. Further, the data were collected in controlled way, even so called naturalised recordings must be much clearer than it would be if we recorded truly natural activity, as the persons were trying not to make unnecessary moves.

Another problem is the sensitivity to the position of the watch on the wrist. We record accelerations in three dimensional Cartesian coordinate system relative to the watch. If it is misplaced, new signatures of transitions might not fit to the learned patterns, however we haven't specifically checked the robustness of the model to this issue.

Some data preprocessing could help with tackling this problem, e.g. calculating the root mean square of all accelerations. The RMS shouldn't be affected by the position of the watch. Unfortunately, the best averaged F_1 score we get is only 47%, but what is more worrying, the confusion between two transitions was in the best case 60%.

There are also some methods we have tried and which are worth to mention as they add some instructive information about CPM. While looking at the raw data (Figure 10) the difference between two transitions is noticeable. It is just change of level with some transient states between, while walking, if averaged over many realisations, would be a flat line due to its oscillations. In order to investigate this, we applied an 0.3 s long sliding window to calculate the local trend, and then run CPM on both, the trend and residuals. As expected, the results for residuals are hopeless, mainly because the confusion between the transitions - at least 24% (the averaged F_1 score at best: 62%). For trends, the scores are much better with no confusion and averaged F_1 score at 88% for the shortest window, although still less than scores obtained on original recordings. It means that trends play crucial role, but are not alone responsible for the classification.

There is still further work to be done to employ the CPM in activity monitoring. In order to eliminate very poor performance in some individual cases, it is necessary to investigate in detail what differ the outliers from the others. It could be also worth to try to introduce thresholds while classifying. So far, a sample is assigned to a class to which it fits best, regardless of actual fitting score. Some cutoffs imposed on values returned by classifier (see bottom plot in Figure 10) would bring about fraction of samples remaining unknown, but this might be better than misleading classification.

Also, in order to seriously employ the algorithm, it is necessary to develop a meta-classifier able to fairly precisely point the location of a transition. So far, due to the overlap of series, there are multiple correct classifications allocated around the transition, see Figure 10, but this is a positive outcome. Our prototypical idea of picking up only one, to most obvious sample (see Section 3.5.2), doesn't take advantage of this multiplicity. Nevertheless, we can imagine a meta-classifier which will take the number of consecutive assignments to the same class into account and based on this will be able to exclude false instances and attach some measure of uncertainty to the detected activities.

6 Conclusions

The obtained results suggest that CPM can perform satisfactory in aligning the raw data. A particularly interesting example is an alignment revealing a walking rhythm (compare sequences before and after aligning in Figure 6). This ability allows to obtain the underlying pattern and to classify the raw series to different classes

Even for naturalised data the model is able to satisfactory detect the transitions, commonly with no mistakes. The accuracy of the activity detection, measured by the mean F_1 score, reaches: 89.9%, 95.9% and 92.3% for stand-to-sit transition, sit-to-stand transition and waking respectively. Although, we are aware of the limitations, this preliminary studies deliver strong arguments to consider the applications based on the Continuous Profile Model to be prospective tools in activity monitoring.

Acknowledgment

I would like to thank my supervisor James Amor for inspiring discussions, insightful remarks and for keeping me positive.

References

[1] Eurostat: *Regional population projections*, Statistics Explained (2013/6/4), `http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Regional_population_projections#`

[2] E. Diczfalusy: *The demographic revolution and our common future*, Maturitas. 2001 Feb 28;38(1):5-14.

[3] J. Askham, E. Glucksman, P. Owens, C. Swift, A. Tinker, G. Yu: *Home and leisure accident research: A review of research on falls among elderly people, Age Concern Institute of Gerontology*, Kings College, London, U.K., 1990.

[4] L. Nyberg, Y. Gustafson: *Patient falls in stroke rehabilitation. A challenge to rehabilitation strategies*, Stroke. 1995 May;26(5):838-42.

[5] B. Najafi, K. Aminian, F. Loew, Y. Blanc, P.A. Robert: *Measurement of StandSit and SitStand Transitions Using a Miniature Gyroscope and Its Application in Fall Risk Evaluation in the Elderly*, IEEE Trans Biomed Eng, 2002 Aug; 49(8): 843-51.

[6] The USEFIL project: `http://www.usefil.eu/`.

[7] V. Ahanathapillai, J.D. Amor, M. Tadeusiak, C.J. James: *Wrist-worn accelerometer to detect postural transitions and walking patterns*, MEDICON2013, 2013.

[8] J. Listgarten: *Analysis of sibling time series data: alignment and difference detection*, PhD Thesis, Department of Computer Science, University of Toronto, 2006.

[9] J. Listgarten, R.M. Neal, S.T. Roweis, A. Emili: *Multiple Alignment of Continuous Time Series*, Advances in Neural Information Processing Systems 17, MIT Press, 2005.

[10] J.D. Amor: *Detecting and Monitoring Behavioural Change Through Personalised Ambien Monitoring*, University of Southampton, Faculty of Engineering and the Environment Institute of Sound and Vibration Research, PhD Thesis, 2011.

[11] L.R. Rabiner: *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceeding of the IEEE, Vol. 77, No. 2, Feb 1989;

[12] A.J. Viterbi: *Error bounds for convolutional codes and an asymptotically optimal decoding algorithm*, IEEE Trans. Informat. Theory, vol. IT-13, pp. 260-269, Apr 1967.