Time Series Segmentation Applied to a New Data Set for Mobile Sensing of Human Activities

Arik Ermshaus[1,*], Sunita Singh[1] and Ulf Leser[1]

[1]Humboldt-Universität zu Berlin, Berlin, Germany

Abstract

Human activity recognition (HAR) systems implement workflows that automatically detect activities from motion data, captured e.g. by wearable devices such as smartphones. These devices contain multiple sensors that record human motion as acceleration, rotation and orientation in long time series (TS) data. As a first step, HAR methods typically partition such recordings into smaller subsequences before applying feature extraction and classification. In this study, we evaluate the performance of 6 classical and recently published TS segmentation (TSS) algorithms on a new large HAR benchmark of 126 TS with up to 13 different activities, called MOSAD, recorded with 6 participants using ordinary smartphone sensors. Our results show that the ClaSP algorithm achieves significantly more accurate results compared to the other methods, scoring the best segmentations in 57 out of 126 TS. The FLOSS algorithm also shows promising results, particularly for long TS with many segments. MOSAD is freely available at https://github.com/ermshaua/mobile-sensing-human-activity-data-set.

Keywords

Ubiquitous, Mobile, Sensors, Time Series, Human Activity, Unsupervised, Segmentation, Change Points

1. Introduction

Wearable devices, such as smartphones or smartwatches, contain low-cost sensors that produce large amounts of observational data, commonly known as time series (TS). These recordings capture the chronological order and characteristics of human behaviour as long consecutive segments, which can be extracted to derive health status, fitness, or security information [1]. For instance, elderly people have a higher incidence of falling, which increases their risk of bone fractures and rhabdomyolysis [2]. Mobile sensing data can be used to recognize falls and take appropriate actions [3]. In the related field of medical condition monitoring, activity data can be used to supervise patients suffering from dementia or mental illness [4]. Further industrial and military applications exist that use human motion data to extract knowledge and guide decision-support systems [1].

To do this, first the individual activities within a given TS must be identified. Human activity recognition (HAR) is the research area that aims to accomplish this goal using machine learning (ML) technology [4]. Its literature is vast and contains methods for preprocessing [5], feature extraction [6], classification [1], specialized devices [7], tools [8, 9], and benchmarks [10, 9, 11]. One of the first preprocessing steps of a HAR system is usually the task of activity segmentation. This process partitions a sensor recording into consecutive segments, which can then be further processed. The literature generally refers to this task as TS segmentation (TSS) [12] which technically tries to partition a TS into *homogenous* segments divided by abrupt shifts called change points (CPs). The notion of homogeneity strongly depends on the domain and typically depends on signal characteristics [13] and shapes [14]. We recently evaluated new TSS algorithms that advance the state of the art in TSS on multiple benchmarks [15]. However, these approaches have not been specifically tested for the task of activity segmentation, which is the focus of this study.

We collected and annotated a new mobile sensing human activity data set, called MOSAD, with which we evaluated current TSS technology. The data comprises 14 recordings of 9 triaxial sensor signals from 6 participants who performed up to three different motion sequences that include a total of 20 different annotated activities. We evaluated the TSS methods on each of the 126 TS, as opposed to 14 9-dimensional recordings, because some of the algorithms can only process univariate TS. We specifically focused on a daily setting, in which the subjects conducted typical routines and were recorded with a smartphone, as opposed to more intrusive studies that use specific sensor devices and laboratories [10]. MOSAD enqueues in a list of benchmarks for HAR [4] that capture the heterogeneity of the field concerning sensors and their placement, subjects, places, items, and their combination. Recent segmentation benchmarks also contain motion data sets [14, 15], however, their amount of activities is limited and does not sufficiently cover the complexity of human behaviour. In contrast, MOSAD includes long motion sequences that capture up to 13

Published in the Workshop Proceedings of the EDBT/ICDT 2023 Joint Conference (March 28-March 31, 2023, Ioannina, Greece).

[*] Corresponding author.

ermshaua@informatik.hu-berlin.de (A. Ermshaus); singhsun@informatik.hu-berlin.de (S. Singh); leser@informatik.hu-berlin.de (U. Leser)

0000-0002-8138-3060 (A. Ermshaus)

different behaviours, which poses a challenging setting for TSS algorithms. Our specific contributions in this paper are:

1. We present MOSAD, a human activity data set recorded with 3 mobile sensors from a smartphone. We recorded 20 different activities in 3 motion sequences with 6 participants, totalling 14 recordings of 9 triaxial sensor readings (126 TS), annotated and preprocessed the data, and make it publicly available on our supporting website [16] for follow-up works.
2. In our experimental evaluation, we used MOSAD to assess 5 state-of-the-art TSS algorithms, namely ClaSP [15], FLOSS [14], ESPRESSO [17], BinSeg [18] and BOCD [19] as well as a simple baseline called Window [13]. ClaSP scores significantly better results than the competitors and achieves the most accurate segmentations in 57 out of 126 cases. FLOSS scores the 2nd best segmentations and is very accurate for TS with many segments.
3. We make a special effort in this work to make our used Python source codes and evaluation framework, Jupyter-Notebooks, as well as all experiment data and visualizations publicly available on our supporting website [16] to foster the reproducibility of our findings and replicability for follow-up works.

The remainder of this paper is organized as follows: In Section 2 we introduce background knowledge and concepts used in this study, Section 3 discusses related works. In Section 4 we present MOSAD, Section 5 evaluates the TSS algorithms with it, and Section 6 summarizes our findings.

2. Background and Definitions

Human activity recognition (HAR) systems use methods such as video analysis [20], environmental sensing [4], or wearable sensors [21] to capture motion sequences. Among these, the use of wearable sensors embedded in smartphones is particularly interesting as they are commonly worn by individuals in similar positions, making the data readily available, comparable, and rich in information about human behaviour.

The most commonly used and insightful wearable sensors in smartphones for HAR are the accelerometer, gyroscope, and magnetometer [22]. The *accelerometer* captures the acceleration forces acting on a mobile device, which can indicate the presence or absence of motion. The *gyroscope* measures the impact of gravity on the device as angular velocity, allowing for estimation of rotation during movement. The *magnetometer* records the effect of the geomagnetic field on the device, providing insight into the orientation of activities. All three sensors produce triaxial measurements in the X, Y, and Z directions, and are typically digitized at a few hundred Hertz (Hz), resulting in long time series (TS) data that capture typical human activities (like standing or going) as homogenous segments. We formalize such recordings with the following concepts:

Definition 1. *A* time series (TS) T *is a sequence of* $n \in \mathbb{N}$ *real values,* $T = (t_1, \ldots, t_n), t_i \in \mathbb{R}$ *that contains the observable output of a sensor over time. The values are also called observations or data points.*

TS are typically sampled at a fixed rate, such that the duration between two consecutive time points is always the same. This simplifies their inspection and is a prerequisite for many advanced analytics. The central property of a TS is that the measurement t_i was recorded *before* t_{i+1} which leads to local patterns that may repeat, drift or suddenly change over time. This allows for the detection of underlying patterns, trends, and anomalies in the data [23].

Definition 2. *Given a TS* T*, a* subsequence $T_{s,e}$ *of* T *with start offset* s *and end offset* e *consists of the contiguous observations of* T *from position* s *to position* e*, i.e.,* $T_{s,e} = (t_s, \ldots, t_e)$ *with* $1 \leq s \leq e \leq n$*. The length of* $T_{s,e}$ *is* $|T_{s,e}| = e - s + 1$.

We use the terms *subsequence* and *window* interchangeably, and also refer to their length as the *width*. Periodic TS repeat similar subsequences of a fixed length, which we call temporal patterns (or periods). However, local parts of TS may still deviate from each other, such as in period length, shape or amplitude, and temporal patterns can drift or change over time. This characteristic of TS makes segmentation challenging, as it requires the identification of patterns that may vary over time.

Definition 3. *Given a TS* T *that captures a motion sequence, a* change point *(CP) is an offset* $i \in [1, \ldots, n]$ *that corresponds to an activity transition. A* segmentation *of* T *is the ordered sequence of CPs in* T*, i.e.,* t_{i_1},...,t_{i_S} *with* $1 < i_1 < \cdots < i_S < n$ *at which the observed routine changed motions.*

In order to induce the segmentation of a TS, an algorithm needs to find the number and locations of all change points (CPs). This task is generally referred to as time series segmentation (TSS) and approached as an unsupervised learning problem [12]. In this work, we study TSS in the context of human activity recognition (HAR), where it is used as a preliminary preprocessing step before feature extraction and classification. The goal of TSS in this context is to partition the sensor recording into consecutive segments, which can then be further processed and used to identify individual activities.

3. Related Work

Human activity recognition (HAR) has been widely studied over the past few decades and has seen many methodical improvements [1, 4] as well as experimental studies [24, 25]. This research field is provided by the availability of benchmark data sets, which are diverse in terms of their overall topics, the devices used and their placement, the number and age of subjects, and the preprocessing of the resulting data [4]. Notable contributions include the HASC corpora [26], OPPORTUNITY [27], PAMAP [10] and mHealth [9].

The literature contains many specific HAR workflows that combine feature engineering and classification techniques [6] to detect activities in these benchmarks. Most of the techniques analyse motion data in sliding windows of fixed sizes, often between 1 and 10 seconds [4]. The advantage of such an approach is its ease of implementation. However, the apparent downside is that it does not differentiate between segments of different length, leading to data heterogeneity in downstream tasks and potential performance losses. A more adaptive method is to capture exactly one activity per window, which is then further processed [28]. To accomplish this, a sensor recording first needs to be segmented, which is the focus of this study.

The TSS literature contains many methods applicable to activity segmentation, as surveyed in [12]. Bayesian approaches split TS into windows and compare their probability distributions to infer CPs. A popular implementation is Bayesian Online Change Point Detection (BOCD) [19], that uses a recursive message-passing algorithm to infer the most recent CP. It has also been extended for short, gradual changes [29]. Another branch of TSS solves optimization problems to induce segmentations. Given a user-selected cost function that defines the notion of segment homogeneity and a search function, the problem can be numerically solved. A catalogue of parametric and non-parametric cost functions have been proposed [13], and ensembled to increase model robustness and accuracy [30]. Accurate exact and approximate search functions include Pruned Exact Linear Time (PELT) [31] and Binary Segmentation (BinSeg) [18].

Recently, TSS methods that impose no assumptions on the observed data points and change types have been published and evaluated on large benchmarks [15]. FLOSS measures the density of similar subsequences in potential segments and greedily extracts the requested amount of CPs [14]. ESPRESSO extends FLOSS with TS chains, positional subsequence information and a more sophisticated entropy-based segmentation procedure [17]. Lastly, ClaSP is one of the most recently published TSS algorithms that formulizes TSS as a collection of hypothetical, self-supervised TS classification (TSC) problems, where the best-performing label configuration induces the segmentation [15]. However, these methods have only been tested on medium-sized TS with only few segments.

Figure 1: Examples of items and places used in MOSAD.

In this study, we evaluate the recent advances in TSS on a new mobile sensing human motion data set, to assess how well these approaches perform on current, large, real-world motion data with new routines and a special focus on large number of segments. We find this especially important for HAR, as human behaviour is complex, evolving and the technology used to capture it frequently changes.

4. MOSAD

We introduce MOSAD (**Mo**bile **S**ensing Human **A**ctivity **D**ata Set), a new multi-modal, annotated TS data set that contains 14 recordings of 9 triaxial smartphone sensor measurements (126 TS) from 6 human subjects performing (in part) 3 motion sequences in different locations. The aim of the data set is to facilitate the study of human behaviour and the design of TS data mining technology to separate individual activities using low-cost sensors in wearable devices. In creating the data set, we focused on capturing pervasive motion sequences to record meaningful human behaviour. For data collection, we utilized the built-in sensors of a smartphone to record the subjects in an unobtrusive and realistic setting, as is commonly done by many people. We annotated the transitions between activities in the recordings and sampled successive data points at a fixed rate of 50 Hz to enable machine learning technology to analyse the measurements and extract knowledge.

In the following subsections, we elaborate in detail on the data set design (Section 4.1), the data collection process (Subsection 4.2), and the annotation and pre-

Subject ID	Gender	Age	Size (in cm)	Weight (in kg)	Motion Sequence
1	M	65	172	75	1
2	M	21	177	80	1,2,3
3	F	24	168	56	1,2,3
4	F	24	169	53	1,2,3
5	F	24	158	48	1,2,3
6	F	53	163	83	1

Table 1
List of participants.

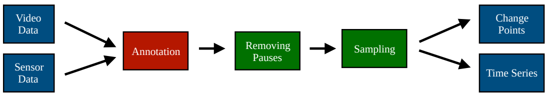

Figure 2: The preprocessing workflow in MOSAD for a single recording. We first annotate the raw sensor data using the ground truth videos. Pauses are then automatically removed and the timestamps resampled. The pipeline outputs a preprocessed multivariate TS with 9 dimensions and a list of associated CPs.

processing steps (Subsection 4.3). We will conclude our technical description of the data set with an overview of its specifications and examples (Subsection 4.4) and information on its availability (Subsection 4.5).

4.1. Data Set Design

The primary objective of MOSAD is to capture the natural dynamics of human behaviour. To achieve this, we used activities that are widely practised and arranged them in an order based on the places where they were recorded. Human behaviour is not entirely unstructured, as we tend to group similar activities and repeat them periodically. Therefore, recording a common pre-arranged routine is a sensible choice, and it has the advantage of being comparable across subjects and modalities. We designed three motion sequences, comprising 13, 3 and 6 activities, respectively. Two of the sequences were recorded indoors, while the third was recorded outdoors. Figure 1 illustrates pictures of some of the items and places used. The routines consist of a total of 20 different activities, varying in length from a few seconds to a few minutes. The following enumeration lists the ordered activities:

1. **Household** (indoor): descend stairs, climb stairs, vacuum, lie, iron, mop, sit, make bed, stand, slow walk, hang out laundry, walk, fold laundry
2. **Sport** (indoor): spin, sit-ups, modified push-ups
3. **Sport** (outdoor): slow walk, run, walk, rope jump, squat, jumping jack

These routines were performed, in part, by 6 human subjects. See Table 1 for more information. The participants comprise 2 males and 4 females, which cluster into young and mid-aged adults. Such an age gap can be interesting to study, as movement changes with age [1]. The 4 younger participants performed all three routines, while the two older subjects chose to only perform the household activities, resulting in a total of 14 recordings.

All participants were recorded using a Samsung Galaxy M20, which was placed in their front right trouser pocket. This location is common for smartphones and has been shown to be highly accurate for activity recognition [32]. We used the application "Physics Toolbox Sensor Suit" to record the phone's accelerometer and gyroscope data from the IMU component (type "LSM6DSL"), as well as its magnetometer (type "YAS539"). Additionally, we filmed each subject with another smartphone to capture the ground truth movements. The data set includes measurements from all three (X, Y, and Z) sensor axes, and is annotated with activity transitions, resulting in a 9-dimensional multivariate TS and one list of CP positions per subject and routine.

4.2. Data Collection

Before the recordings, we provided instructions on how to perform the motion sequences and set time constraints for each activity. We began the data collection with a preparation phase, in which the smartphone was placed in the subject's pocket, filmed the routine, and ended with a follow-up phase, where the mobile device was eventually removed. During the recording, we gave auditory cues to the subjects to assist with timing and maintaining the order of activities. The subjects were asked to pause briefly between two consecutive activities to prevent transitional movements. These pauses were later removed as part of the preprocessing.

We encountered multiple problems throughout the data collection process, that we want to share to help researchers with follow-up works. In the initial experimentation phase, we used an iPhone 11 Pro, which randomly crashed the application during the recordings, resulting in lost measurements. To overcome this issue, we switched to an Android smartphone, which also increased the sensor sampling rate from 100 to 300 Hz. Additionally, the automatic standby mode of the smartphone resulted in data loss. This can be addressed by using other applications that prevent the smartphone from going to standby while still locking the screen to avoid unwanted interactions.

4.3. Preprocessing

In the preprocessing stage, we carried out three key transformations to create an annotated and consistent data set. Figure 2 illustrates this workflow. Firstly, we annotated the sensor signals with timestamps at which activities change to obtain a ground truth. We use these CPs in the evaluation to assess the quality of the TSS methods. To create the annotations, we carefully examined the videos in conjunction with the sensor measurements to derive precise offsets that capture the exact time points of activity transitions. Secondly, we removed the data points from the preparation phase, pauses between activities and the follow-up phase to clean the signals of unwanted noise. Finally, we downsampled the measurements, using linear interpolation, to a fixed sample rate of 50 Hz. This is a common preprocessing step for sensor signals, as their sampling rate often varies due to physical component inaccuracies. It is necessary to create a multivariate TS that is synchronized throughout time and dimensions, improves interpretability as well as the performance of TS analytics that assume equidistant time gaps between measurements. We selected a sample rate of 50 Hz, as used in the mHealth data set [9], because it has been shown to be appropriate for detecting human behaviour [32].

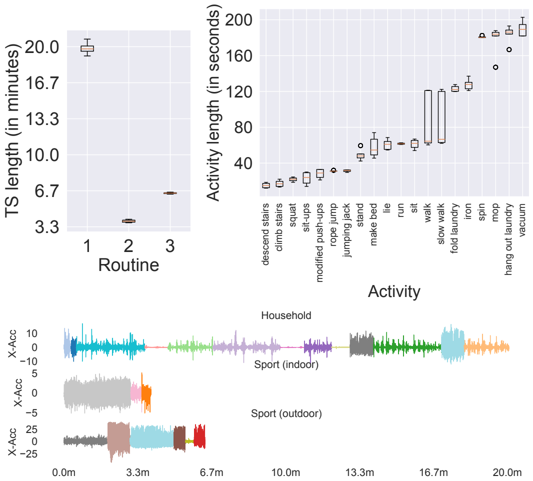

Figure 3: Top: Overview of data set characteristics. Bottom: Example X-axis acceleration recordings for participant 2. Each activity segment is coloured.

4.4. Data Set Overview

In total, MOSAD comprises 14 9-dimensional sensor recordings, totalling 126 TS, from the 6 participants performing up to 3 motion sequences. On average, the routines 1-3 contain 19.9, 3.9 and 6.5 minutes worth of activity data, with only small deviations per participant (Figure 3 top left). The single activities last between 12 seconds for descending the stairs and 203 seconds of vacuuming, also with small variances except for (slow) walking (Figure 3 top right), the only two activities performed in two routines (household and outdoor sport), yet with different time constraints.

In Figure 3 bottom, we illustrate, as an example, the X-axis acceleration recordings of participant 2 performing the 3 motion sequences. The single activities are coloured and show subtle but also more pronounced differences between each other.

4.5. Data Availability

We make MOSAD publicly available under the CC BY-SA licence. Users are permitted to share and adapt the data with author credit, but must also use this licence. The data set can be downloaded on our supporting website [16], including a data loader for Python and supplementary materials.

5. Experimental Evaluation

In this section, we present the experimental evaluation of different TSS methods on MOSAD. We first describe our setup in Subsection 5.1, then compare the accuracy of 6 TSS competitors in Subsection 5.2. To further demonstrate the characteristics of our data set, we discuss the results of the two best-performing methods using a selected example in Subsection 5.3. We make all used source codes, Jupyter-Notebooks, and the raw measurement sheets available on our website [16].

5.1. Benchmark Setup

We conducted all experiments on an Intel Xeon E7-4830 with a clock speed of 2.20 GHz, 440 GB of RAM, and 80 cores, using Python 3.8.

Competitors We compare 6 established and recently published TSS algorithms on MOSAD. We include a simple baseline called Window [13] that measures the discrepancy between successive subsequences using a Mahalanobis cost function. As the width of the subsequences, we chose 5 seconds of motion data (250 data points), which is sufficient to capture the transitions of human behaviour. We also include two classical methods, BinSeg [18] with auto-regressive cost (optimization-based) and BOCD [19] (Bayesian), that are very popular and have performed well in a recent evaluation [33]. Additionally, we include the density-based algorithm FLOSS [14] and its variant ESPRESSO [17]. Both require a window size, which should roughly capture one instance of a temporal pattern, that we set to 1 second (50 data points). FLOSS also uses a sliding window to achieve more accurate results in long TS data that we set to 20 seconds of sensor data (1k data points). ESPRESSO needs a TS chain length that we set to 3 as in [15]. Lastly, we report results for our own method ClaSP [15]. Except BOCD and ESPRESSO, all the aforementioned algorithms use a minimum segment size parameter that we set to 5 seconds. This increases their accuracy, but it exhibits domain knowledge. We also set the number of segments as a hyper-parameter for all algorithms, as only ESPRESSO and ClaSP can automatically infer this information without adaptation.

Evaluation Metric The quantitative evaluation of TSS algorithms is challenging as signals and ground truth annotations can be ambiguous. Therefore, the literature contains a range of evaluation measures that quantify different notions of quality. These measures can be divided into classification-based and clustering-based approaches.

While classification-based approaches check if a detected CP matches the annotation (with some slack), clustering-based approaches report the exact degree of deviation. In this study, we have chosen to use the latter and have selected the Covering measure [33]. This measure quantifies the overlap between predicted and annotated segments. It is defined as follows:

Let the interval of two successive CPs $[t_{i_k}, \ldots, t_{i_{k+1}}]$ denote a segment in T and let $segs_{pred}$ as well as $segs_T$ be the sets of predicted or ground truth segmentations, respectively. For notational convenience, we always consider $t_{i_0} = 0$ as the first and $t_{i_C} = n + 1$ as the last CP to include the first (last) segment. The Covering score reports the best-scoring weighted overlap between a ground truth and a predicted segmentation (using the Jaccard index) as a normed value in the interval $[0, \ldots, 1]$ with higher being better (equation 1).

$$Covering = \frac{1}{\|T\|} \sum_{s \in segs_T} \|s\| \cdot \max_{s' \in segs_{pred}} \frac{\|s \cap s'\|}{\|s \cup s'\|} \tag{1}$$

It is defined for sets with varying sizes (including being empty). In order to perform a performance comparison, we run the TSS algorithms with each of the 126 TS and aggregate the resulting Covering scores into a single ranking. First, we compute the rank of the score of each method per TS, where the best method is assigned rank 1, the 2nd best method is assigned rank 2, and so on. Then, we average the ranks of a method on all TS to obtain its overall rank. To visualize the final ranking, we use critical difference (CD) diagrams, as introduced in [34]. The best-ranking approaches with the lowest (average) ranks are shown to the right of the diagram; see, for instance, Figure 4 (left). Groups of approaches that are not significantly different in their ranks are connected by a bar, based on a Nemenyi two-tailed significance test with $\alpha = 0.05$.

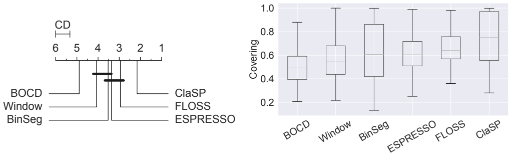

Figure 4: Covering segmentation performances on the 126 TS in MOSAD for the 6 state-of-the-art competitors. The ClaSP algorithm ranks first and shows significantly better results.

	mean	median	std
ClaSP	**74.0%**	**75.0%**	21.1%
FLOSS	66.3%	64.0%	14.4%
ESPRESSO	61.8%	60.3%	15.8%
BinSeg	62.9%	60.6%	23.5%
Window	56.1%	54.4%	18.6%
BOCD	50.2%	49.2%	16.0%

Table 2
Summary Covering performances for the 6 competitors on the 126 TS in MOSAD. ClaSP shows by far the best results.

5.2. Segmentation Performance

We compare the average Covering ranks of the 6 competitors on the entire MOSAD data set. The CD diagram in Figure 4 (left) shows that ClaSP (2.2) ranks first, followed by FLOSS (2.9), ESPRESSO (3.4), BinSeg (3.5), Window (4.1), and BOCD (4.9). FLOSS, ESPRESSO and BinSeg form a group of insignificantly different approaches, while Window and BOCD perform much worse. Considering the wins or ties (first position in ranking), ClaSP scores the best performances for 57 TS, followed by FLOSS (36), ESPRESSO (17), Window (10), BinSeg (8), and BOCD without any wins (counts do not sum up to 126 due to ties). We analysed the properties of the 69 TS where ClaSP ranks 2nd position or worse, but did not find any obvious commonalities, such as certain routines, sensors, or subjects.

Considering these dimensions in isolation, we observe similar results compared to the global ranking. For the single routines, FLOSS scores the best results for the household sequence, while ClaSP wins both the indoor and outdoor sport motions. This indicates that the temporal constraint in FLOSS, if properly set, enables it to score good results in long activity recordings. We also analysed the 42 TS per sensor type individually and find that ClaSP scores best-ranking results for all three of them, namely acceleration, gyroscope and magnetometer. A similar finding holds if we aggregate our results by subject; ClaSP scores the best segmentation performances for participant 1-5 while FLOSS wins for subject 6 (who only performed the household routine). The rankings show that ClaSP scores state-of-the-art performances across TS and their characteristics. FLOSS also scores

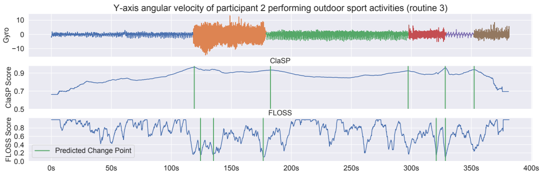

Figure 5: The outdoor sport routine 3 performed by participant 2. The TS (top) visualizes the Y-axis gyroscope that captures the motion. The score profiles for ClaSP and FLOSS (centre and bottom) are illustrated with predicted CPs (green).

good results and outperforms ClaSP in certain specific settings, though without a significant difference.

The summary statistics in Table 2 and Figure 4 (right) validate our findings. ClaSP scores the highest mean Covering performance of 74.0% with a rather large standard deviation of 21.1%. This can be explained by a large performance difference of circa 30 percentage points (pp) between the household and sport routines. On average, ClaSP improves compared to the 2nd best competitor FLOSS by 7.7 pp. In a pairwise comparison of ClaSP against the other algorithms, it achieves between 79 wins against FLOSS and 110 wins vs BOCD. However, its average performance still is only 74.0%, which needs improvement.

5.3. Outdoor Sport Motion Segmentation

Besides the quantitative analysis, we also assess the quality of computation for the best-ranking methods, ClaSP and FLOSS, by human inspection. Figure 5 shows a selected example of participant 2 performing the outdoor sport routine 3 (top) and the score profiles from both algorithms, including CP predictions (green vertical lines). ClaSP assigns high scores to TS offsets that are likely CPs, while FLOSS annotates low scores. Local maxima (minima) are then extracted with different peak detection algorithms to locate the CP predictions. Both algorithms produce good segmentations in this example. ClaSP detects all 5 activity transitions, while FLOSS locates 3. The profiles, however, are quite different. ClaSP shows smooth scores with only small deflections, while FLOSS has many smaller peaks and is noisy. This complicates the CP detection and leads to worse results.

6. Conclusion

In this paper, we have introduced a new mobile sensing human activity data set, called MOSAD, which includes 126 TS from 6 participants performing three motion sequences. The data set utilizes current sensor technology and captures two age groups. It is freely available on our supporting website [16] and can be used in follow-up research.

We have also evaluated 6 state-of-the-art TSS algorithms using MOSAD, and find that the ClaSP algorithm significantly outperforms its competitors. These findings are consistent with recently published benchmark results [15]. It achieves the highest accuracy in 57 out of the 126 TSS in MOSAD, increasing the average accuracy by 7.7 pp compared to the 2nd best method.

Based on our results, we conclude that the ClaSP algorithm can be a suitable method for activity segmentation. However, we acknowledge that there is still room for improvement in performance. Future work will investigate the use of multivariate TSS methods to exploit temporal dependencies between sensors.

References

[1] O. D. Lara, M. A. Labrador, A survey on human activity recognition using wearable sensors, IEEE Communications Surveys & Tutorials 15 (2013) 1192–1209.

[2] A. Ungar, M. Rafanelli, I. Iacomelli, M. A. Brunetti, A. Ceccofiglio, F. Tesi, N. Marchionni, Fall prevention in the elderly., Clinical cases in mineral and bone metabolism : the official journal of the Italian Society of Osteoporosis, Mineral Metabolism, and Skeletal Diseases 10 2 (2013) 91–5.

[3] J. Yin, Q. Yang, J. J. Pan, Sensor-based abnormal human-activity detection, IEEE Transactions on Knowledge and Data Engineering 20 (2008) 1082–1090.

[4] M. A. R. Ahad, A. D. Antar, M. Ahmed, Iot sensor-based activity recognition - human activity recognition, in: Intelligent Systems Reference Library, 2021.

[5] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, P. J. M. Havinga, Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey, in: ARCS, 2010.

[6] S. Pirttikangas, K. Fujinami, T. Nakajima, Feature selection and activity recognition from wearable sensors, in: Ubiquitous Computing Systems, 2006.

[7] C. Park, J. Liu, P. H. Chou, Eco: an ultra-compact low-power wireless sensor node for real-time motion monitoring, IPSN 2005. Fourth International Symposium on Information Processing in Sensor Networks, 2005. (2005) 398–403.

[8] N. Kawaguchi, N. Ogawa, Y. Iwasaki, K. Kaji, Distributed human activity data processing using hasc tool, in: UbiComp '11, 2011.

[9] O. Baños, C. Villalonga, R. García, A. Saez, M. Damas, J. A. Holgado-Terriza, S. Lee, H. Pomares, I. Rojas, Design, implementation and validation of a

novel open framework for agile development of mobile health applications, BioMedical Engineering OnLine 14 (2015) S6 – S6.

[10] A. Reiss, D. Stricker, Introducing a new benchmarked dataset for activity monitoring, 2012 16th International Symposium on Wearable Computers (2012) 108–109.

[11] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, J. del R. Millán, Collecting complex activity datasets in highly rich networked sensor environments, 2010 Seventh International Conference on Networked Sensing Systems (INSS) (2010) 233–240.

[12] S. Aminikhanghahi, D. J. Cook, A survey of methods for time series change point detection, Knowledge and Information Systems 51 (2017) 339–367.

[13] C. Truong, L. Oudre, N. Vayatis, Selective review of offline change point detection methods, Signal Processing 167 (2020) 107299.

[14] S. Gharghabi, C.-C. M. Yeh, Y. Ding, W. Ding, P. R. Hibbing, S. R. LaMunion, A. Kaplan, S. E. Crouter, E. J. Keogh, Domain agnostic online semantic segmentation for multi-dimensional time series, Data Mining and Knowledge Discovery 33 (2018) 96 – 130.

[15] A. Ermshaus, P. Schäfer, U. Leser, Clasp: parameter-free time series segmentation, Data Mining and Knowledge Discovery (2023).

[16] MOSAD, Code and Raw Results, https://github.com/ermshaua/mobile-sensing-human-activity-data-set, 2023.

[17] S. Deldari, D. V. Smith, A. Sadri, F. D. Salim, Espresso: Entropy and shape aware time-series segmentation for processing heterogeneous sensor data, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 4 (2020) 77:1–77:24.

[18] A. K. Sen, M. S. Srivastava, On tests for detecting change in mean, Annals of Statistics 3 (1975) 98–108.

[19] R. P. Adams, D. J. MacKay, Bayesian online changepoint detection, arXiv preprint arXiv:0710.3742 (2007).

[20] M. A. R. Ahad, Vision and sensor-based human activity recognition: challenges ahead, in: Advancements in Instrumentation and Control in Applied System Applications, IGI Global, 2020, pp. 17–35.

[21] S. A. Elkader, M. Barlow, E. Lakshika, Wearable sensors for recognizing individuals undertaking daily activities, Proceedings of the 2018 ACM International Symposium on Wearable Computers (2018).

[22] F. Demrozi, G. Pravadelli, A. Bihorac, P. Rashidi, Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey, IEEE Access 8 (2020) 210816–210836.

[23] A. Ermshaus, P. Schäfer, U. Leser, Window size selection in unsupervised time series analytics: A review and benchmark, 7th Workshop on Advanced Analytics and Learning on Temporal Data (2022).

[24] C. R. Wren, E. M. Tapia, Toward scalable activity recognition for sensor networks, in: Location- and Context-Awareness, 2006.

[25] A. M. Swartz, S. Strath, D. Bassett, W. L. O'Brien, G. A. King, B. E. Ainsworth, Estimation of energy expenditure using csa accelerometers at hip and wrist sites., Medicine and science in sports and exercise 32 9 Suppl (2000) S450–6.

[26] N. Kawaguchi, Y. G. Yang, T. Yang, N. Ogawa, Y. Iwasaki, K. Kaji, T. Terada, K. Murao, S. Inoue, Y. Kawahara, Y. Sumi, N. Nishio, Hasc2011corpus: towards the common ground of human activity recognition, in: Ubiquitous Computing, 2011.

[27] R. Chavarriaga, H. Sagha, A. Calatroni, S. T. Digumarti, G. Tröster, J. del R. Millán, D. Roggen, The opportunity challenge: A benchmark database for on-body sensor-based activity recognition, Pattern Recognit. Lett. 34 (2013) 2033–2042.

[28] A. Khan, Y.-K. Lee, S. Lee, T.-S. Kim, A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer, IEEE Transactions on Information Technology in Biomedicine 14 (2010) 1166–1172.

[29] E. Draayer, H. Cao, Y. Hao, Reevaluating the change point detection problem with segment-based bayesian online detection, Proceedings of the 30th ACM International Conference on Information & Knowledge Management (2021).

[30] I. D. Katser, V. Kozitsin, V. Lobachev, I. Maksimov, Unsupervised offline changepoint detection ensembles, Applied Sciences 11 (2021) 4280.

[31] R. Killick, P. Fearnhead, I. A. Eckley, Optimal detection of changepoints with a linear computational cost, Journal of the American Statistical Association 107 (2012) 1590 – 1598.

[32] G. Bieber, J. Voskamp, B. Urban, Activity recognition for everyday life on mobile phones, in: Interacción, 2009.

[33] G. J. J. van den Burg, C. K. I. Williams, An evaluation of change point detection algorithms, ArXiv abs/2003.06222 (2020).

[34] J. Demšar, Statistical Comparisons of Classifiers over Multiple Data Sets, The Journal of Machine Learning Research 7 (2006) 1–30.