ESPRESSO: Entropy and ShaPe awaRe timE-Series SegmentatiOn for Processing Heterogeneous Sensor Data

SHOHREH DELDARI, School of Science, RMIT University/ Data61, CSIRO, Australia
DANIEL V. SMITH, Data61, CSIRO, Australia
AMIN SADRI, ANZ, Australia
FLORA SALIM, School of Science, RMIT University, Australia

Extracting informative and meaningful temporal segments from high-dimensional wearable sensor data, smart devices, or IoT data is a vital preprocessing step in applications such as Human Activity Recognition (HAR), trajectory prediction, gesture recognition, and lifelogging. In this paper, we propose *ESPRESSO* (Entropy and ShaPe awaRe timE-Series SegmentatiOn), a hybrid segmentation model for multi-dimensional time-series that is formulated to exploit the entropy and temporal shape properties of time-series. *ESPRESSO* differs from existing methods that focus upon particular statistical or temporal properties of time-series exclusively. As part of model development, a novel temporal representation of time-series *WCAC* was introduced along with a greedy search approach that estimate segments based upon the entropy metric. *ESPRESSO* was shown to offer superior performance to four state-of-the-art methods across seven public datasets of wearable and wear-free sensing. In addition, we undertake a deeper investigation of these datasets to understand how ESPRESSO and its constituent methods perform with respect to different dataset characteristics. Finally, we provide two interesting case-studies to show how applying *ESPRESSO* can assist in inferring daily activity routines and the emotional state of humans.

CCS Concepts: • **Computing methodologies** → *Machine learning approaches*; • **Information systems** → *Data mining*; *Data stream mining*.

Additional Key Words and Phrases: Sensor Data, Time-series, Segmentation, Change Point Detection, Activity Recognition, Wearable, Temporal, Pattern Recognition

ACM Reference Format:
Shohreh Deldari, Daniel V. Smith, Amin Sadri, and Flora Salim. 2020. ESPRESSO: Entropy and ShaPe awaRe timE-Series SegmentatiOn for Processing Heterogeneous Sensor Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 77 (September 2020), 24 pages. https://doi.org/10.1145/3411832

1 INTRODUCTION

Today, there is a growing demand for data mining technologies to transform the complex, unwieldy data collected from a broad diverse range of wearable devices, smartphones, and sensors into compact and actionable information. Whilst supervised methods work well, they require carefully labelled samples. Annotating datasets of wearable sensors can be challenging for a couple of reasons. In addition to the privacy issues associated with collecting human data, the huge volume of data and hierarchical structure of human activities can make the

Authors' addresses: Shohreh Deldari, shohreh.deldari@rmit.edu.au, School of Science, RMIT University/ Data61, CSIRO, Melbourne, VIC, Australia, ; Daniel V. Smith, daniel.v.smith@data61.csiro.au, Data61, CSIRO, Hobart, TAS, Australia; Amin Sadri, amin.sadri@anz.com, ANZ, Melbourne, VIC, Australia, ; Flora Salim, flora.salim@rmit.edu.au, School of Science, RMIT University, Melbourne, VIC, Australia,

2474-9567/2020/9-ART77 $15.00
https://doi.org/10.1145/3411832

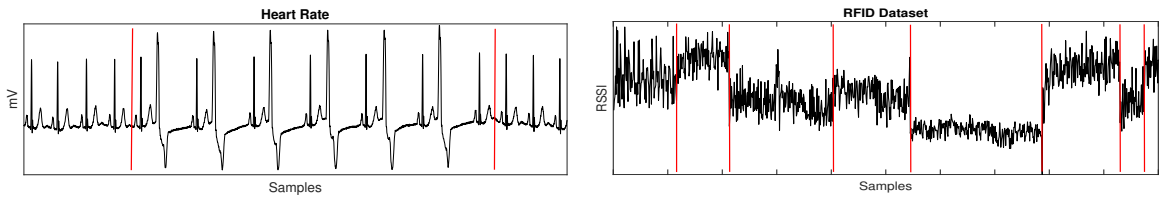
annotation process time-consuming, expensive and sometimes even infeasible. Consequently, unsupervised and self-supervised techniques have gained a lot of attention [37]. Furthermore, automatic knowledge extraction techniques are required to factorise this large volume of sensor data into interpretable pieces of information.

Time-series segmentation is the process of partitioning time-series into a sequence of discrete and homogeneous segments. We propose a new multivariate time-series segmentation technique to be used as a preliminary processing or exploratory data analysis step prior to tasks such as prediction, feature selection, semi-supervised or unsupervised classification. Therefore, the motivation of this paper is to enable a deeper unsupervised exploration of wearable sensor datasets by factorising them into a set of atomic primitives of physical action or emotion. By enabling these primitives to be discovered, the process of feature engineering can be accelerated, whilst in addition, greater insights into the underlying properties of the data can be learned.

Recent studies in Human Activity Recognition (HAR) have demonstrated the effectiveness of using temporal segmentation in combination with classification [3, 6, 25, 42, 46]. In addition to HAR, time-series segmentation have been applied to other modeling tasks with wearable sensors, including trajectory prediction [36], motion-based user authentication [16], life-logging [7], elderly rehabilitation [20], anomaly detection techniques [31], predictions [1, 43, 44], and feature selection [21].

In pervasive computing applications, the time-series being collected will often be heterogenous encompassing a diverse range of characteristics with respect to their dimensionality, continuity, statistical properties and shape. Figure 1 shows two time-series with very different properties. Figure 1(a) shows the repetitive temporal shape patterns of the human heart measured with a wearable electrocardiogram (ECG). Figure 1(b) shows a sequence of human postures that have been measured with a passive RFID tag array; the RSSI of each posture have different statistical properties. This figure clearly shows the semantics of each use case should be extracted by exploiting different time-series properties. Statistical changes can be used to segment the human postures with high precision, however, temporal shape changes will fail in distinguishing these segments. In contrast, exploiting temporal shape changes in the ECG data will be advantageous to segment abnormal heartbeats (the middle segment of the ECG) compared to using statistical changes that are more uniform across the segments in (a) than (b).

While current time-series segmentation methods exploit individual characteristics of the signal, that include the temporal shape, statistics, or probability distribution, we propose Entropy and ShaPe awaRe time-series SEgmentation (*ESPRESSO*), a hybrid model that incorporates multiple signal characteristics. To achieve this, ESPRESSO integrates the search and score based mechanisms of segmentation through a newly proposed shape representation, *WCAC*, and a greedy search that exploits a non-parametric entropy based cost function. The segmentation results are then further enhanced by devising an embedded channel ranking algorithm. *ESPRESSO*



(a) (b)

Fig. 1. a) A time-series of human heart beats with segments that have clear differences in the temporal shape. b) A time-series of RSSI measured by a RFID tag where segments of unique human postures possess clear statistical differences.

has been developed to accurately segment a wider range of time-series by relaxing some of the assumptions imposed by statistical or temporal shape-based methods.

The main challenges of current multi-dimensional time-series segmentation approaches are as follows: 1) Model assumptions: Models make parametric assumptions about the underlying properties of time-series that can limit its application. 2) Model parameterisation: Segmentation models generally utilise a number of parameters and thresholds that need to be carefully tuned based upon domain knowledge. 3) Channel ranking: In multi-dimensional time-series, not all dimensions (channels) are equally important to achieving accurate segmentation. Although there are numerous supervised channel selection techniques for classification tasks, there is little work on ranking the relevance of channels for unsupervised segmentation.

Current work on temporal shape-based segmentation [10] operates under the principle that similar shaped patterns are associated with the same segment class and occur within close temporal proximity. This assumption, however, can lead to degraded segmentation performance under any of the following conditions: a) Several instances of the same class (with the same label) repeat multiple times across the time-series; b) segment classes are not comprised of repeated shape patterns; c) shape patterns drift with time. Each of these conditions are commonly encountered in wearable sensor use-cases. We propose a temporal shape-based segmentation method, *Weighted Chain Arc Curve (WCAC)*, to address these limitations. In addition to temporal shape-based methods, there are a range of statistical based segmentation approaches. Such approaches have commonly employed parametric models in the form of Probability Density Functions *(PDF)* [4, 13, 28] and auto-regressive models [48] but have somewhat limited application given they impose strong assumptions upon the statistical properties of the time-series. Non-parametric kernel based methods have been proposed [3, 18, 26, 47] to offer greater modelling flexibility, but can be difficult to train and provide poor estimates across smaller sample sets.

1.1 Contributions

The main contributions of this paper are as follows:

(1) *ESPRESSO* is a novel time-series segmentation approach which integrates temporal shape and entropy based properties of multidimensional time-series. Unlike most state of the art methods that require several carefully tuned parameters, *ESPRESSO* only depends on one parameter which can be selected with minimal risk, given *ESPRESSO*'s performance is shown to be relatively consistent with respect to this parameter. *ESPRESSO* is shown to outperform four state-of-the-art segmentation methods in terms of their F-score and RMSE measure across seven public datasets of wearable sensors in this experiment.
(2) We propose *WCAC* to address particular limitations of existing shape-based segmentation approaches. In contrast to other temporal shape methods, *WCAC* can accommodate both repeated segments and shifts in temporal shape across the time-series.
(3) An embedded channel ranking has been utilised in *ESPRESSO* to make segmentation more robust to noisy and/or irrelevant channels.
(4) We categorise time-series datasets with regards to their continuity and repetitive patterns. An ablation study is performed to evaluate *ESPRESSO*'s performance with respect to these categories.
(5) Finally, we demonstrate the interpretability of segmentation results obtained by *ESPRESSO* for two real-world use-cases. The first study shows the ability of *ESPRESSO* to discover deviations in the daily routines of people through life-logging data. The second study shows *ESPRESSO* can identify the emotional states of people and provide an interpretation of their emotional transitions.

2 RELATED WORKS

Although the term segmentation is frequently used in the time-series literature, we focus upon approaches that partition time-series from the bottom up by using salient changes in the series to identify individual segment

boundaries, or from the top down by identifying the segment boundaries that optimise a cost function across an entire time-series. First, we review current applications of time-series segmentation in the field of wearable sensors and device-free dataset. We then provide an overview of general time-series segmentation approaches and highlight the current limitations of these approaches.

2.1 Wearable Sensors Segmentation

For wearable sensors, segmentation methods use a fixed-length sliding window.. Authors in [42] have compared the effect of window size in detecting different types of activities in HAR applications. To estimate the effect of window size on activity recognition, they divide activities into two main groups: simple activities with periodic actions such as running and waving and more complex, non-periodic actions such as drinking coffee. They found that shorter windows were shown to be effective for simple periodic activities but less reliable to represent the more complex activities. Consequently, recent works, such as [30], have exploited different sized sizes to represent activities of varying complexity. The [25] method proposed an optimization approach to find the optimal window size for activity segmentation, however, this can be challenging task to undertake given the variety of sensors and activities associated with real-world applications.

Temporal segmentation has been shown to be an important pre-processing step in high-dimensional wearable and device-free sensor applications [5, 12, 23]. Segmentation has been mentioned as an open challenge in analyzing life-logging data from wearable sensors in order to have more accurate models [7]. A recently published wearable development toolkit, *WDK*, has incorporated time-series segmentation methods; this demonstrates the impact that these technique have had on wearable sensing applications [12]. Authors in [2, 3] showed that applying classification on top of a segmentation method produced more accurate results than performing classification with a fixed-length sliding window. They proposed a new probability metric, *SEP* to improve current probability density-ratio change point detection techniques in smart home applications. The authors in [46] propose a simple segmentation method that exploit knowledge of the statistical characteristics of low-level human activities (i.e. walking, running) to segment RFID signals. Temporal segmentation of motion data has been studied extensively. The method in [29] proposed using decision trees to find the split points. The authors in [6] proposed a multiple regression model based upon *Expectation Maximization*, *MRHLP*, which identifies activity boundaries as the points where there is a switch in the underlying models.

In addition to classification problems, the authors in [36] have shown that using temporal segmentation in conjunction with a prediction model leads to performance improvements. They have utilized an entropy-based temporal segmentation method to improve the prediction quality of user activity trajectories. For a user identification and authentication application, [16] devised a sequence labelling based segmentation approach to extract physical and behavioural characteristics of individuals from a sequence of their daily activities. Segmentation has also been used for feature selection in datasets of multi-dimensional human activity motion, electroencephalogram (EEG) signals and speech signals [21]. [39, 40] found temporal segmentation could be used to identify a user's behavioral characteristics from their smartphone usage data. Table 1 summarizes some recent wearable sensor applications that benefit from unsupervised segmentation. Other than human-centric applications, time-series segmentation has been applied to a broad range of fields such as sensor data processing, environmental modelling, financial events, music and speech processing, energy consumption predictions and so on. [1] provides a detailed review of time-series change point detection methods.

2.2 Time-series Segmentation Approaches

Time-series segmentation approaches can be divided into supervised and unsupervised techniques. Supervised methods infer the class labels of underlying time-series using binary or multi-class classifiers formed from Hidden Markov Models [38] and Decision Trees [29] in order to identify segment boundaries. Unsupervised techniques

Table 1. List of some of recent wearable sensor analysis applications which benefit from automatic segmentation.

Year	Paper	Applying segmentation for ...
2019	[3]	Activity recognition in smart-home
	[16]	Authentication and Identification
	[21]	Feature selection in HAR and EEG data
2018	[36]	User trajectory prediction
	[46]	HAR using reflection of RFID signals
2017	[39]	User behavioral characteristic base on smartphone usage
	[29]	Physical activity recognition using ACC data
2016	[26]	Complex activity recognition
	[23]	Human motion modelling using ACC data
	[20]	Exercise and physical rehabilitation analysis
2013	[6]	Physical activity recognition using ACC data
	[22]	Exercise and physical rehabilitation for elderly people
	[24]	Exercise and physical rehabilitation analysis

are more commonly utilized than supervised approaches given they do not require training sets of segmented data. Instead unsupervised methods exploit the underlying signal properties to estimate the change points.

2.2.1 Statistical Approaches. The statistical properties of a time-series are most frequently exploited in unsupervised segmentation. These methods can be categorised as top-down optimisation approaches [13, 35] that search for the set of segments that maximise its particular cost function, or bottom up approaches that identify individual segment boundaries from local deviations in the time series [4, 18, 27, 28, 47, 48].

Sadri proposed a top-down temporal segmentation method, *IGTS*, which was based upon the information gain (IG) metric [35]. Segment boundaries were estimated by using a dynamic programming approach to maximise the IG of its constituent segments. Recently, an online version of this unsupervised informatin gain based method is proposed in [51]. A similar top-down approach was used in [13], where a greedy search was used to identify the segment boundaries that maximize the regularized likelihood estimate of a segmented Gaussian model.

The statistical differences between time intervals have commonly been measured with the likelihood ratio formulation [4, 18, 27, 28, 48]. Within this formulation, parametric models have been used to estimate the intervals as Probability Density Functions (PDFs) [4, 28], auto-regressive models [48] or state space models [19]. The parametric assumptions of these models, however, limit the types of statistical changes that can be detected. For instance, by fitting a Gaussian distribution to segments such as in [4, 28], only differences in the mean and/or standard deviation of adjacent intervals can be used in segmentation. Whilst these limitations can be relaxed by considering non-parametric density estimation, this still remains a difficult estimation problem to address.

Flexible non-parametric solutions [18, 27] were proposed to compute the likelihood ratio without the need for density estimation. It was found that estimating the ratio of PDFs directly was a simpler problem to address than density estimation. Hence, a non-parametric Gaussian kernel could be successfully used for this purpose. Kullback-Leibler Importance Estimation Procedure (KLIEP) was used to directly estimate the ratio of PDFs [18]. Liu adopted the Relative unconstrained Least Square Importance Fitting (*RulSIF*) to directly estimate the relative ratio of PDFs [27]. These non-parametric approaches for direct ratio estimation were challenging to train and required a cross-validation procedure for model selection. They also tend to produce poor estimates with small datasets. Yamada utilised the non-parametric *additive Hilbert-Schmidt Independence Criterion(aHSIC)*

for time-series segmentation [47]. Change points were detected by using the *aHSIC* criteria to compute the dependency between time adjacent intervals and the pseudo label of statistical change between the intervals.

Whilst non-parametric approaches offer greater flexibility to modeling statistical change than the earlier parametric methods, they are not universally applicable to HAR applications. They assume statistical homogeneity within each segment and statistical heterogeneity between different segments. Whilst this assumption is appropriate for low level segmentation tasks, it will not always be valid for wearable sensing applications where extracted segments need to characterise complex actions, emotions and behaviours.

2.2.2 Temporal Shape Approaches. The temporal shape is another unique property of time-series that can be exploited in segmentation [10, 17] where changes in the temporal shape patterns of a time-series were used to estimate the segment boundaries. *FLOSS*, Fast Low-Cost Semantic Segmentation [10] works under the principle that patterns of similar shape were each associated with the same segment class and occur within close temporal proximity of each other. The limitations of such assumptions were described in the Introduction section. In contrast to *FLOSS*, which is based on the most similar repeated patterns, the authors in [17] proposed a segmentation model based on rare temporal patterns. Although shape-based methods can be beneficial for time-series composed of repeated shape patterns, performance will degrade when segments are composed of diverse shapes or when the shape patterns of a segment drift over time. Recently, the authors of [45, 53] proposed a new pattern-based primitive, *Chain*, to discover a chain of similarly shaped patterns. To make shape extraction robust against pattern drift, we customize this idea in our proposed shape-based segmentation method.

3 PROBLEM DEFINITION

While we follow the Matrix Profile framework introduced in [50], for completeness, we firstly provide a definition of multi-dimensional time-series.

Definition 1. The high-dimensional time-series X, is an $D \times N$ matrix of N samples and D channels (time-series), such that $X = \{X_i^j | 1 < i < N, 1 < j < D\}$, where X_i^j denotes the ith sample of the jth time-series channel.

Definition 2. The subsequence of a time-series, $X_{i,L}^j$, is a vector of samples in channel j ranging between index i and index $i + L - 1$. L is the length of the subsequence.

Definition 3. *Matrix Profile*, *MP*, is a $D \times N - L + 1$ matrix where MP_i^j denotes the distance between subsequence $X_{i,L}^j$ and its nearest neighbor in X^j. Here we employed Euclidean distance as a similarity metric to compare subsequences.

Definition 4. *Matrix Profile Index* or *MPI* is a $n \times n - L + 1$ matrix where MPI_i^j denotes the index of the nearest neighbor (the most similar subsequence) for the subsequence $X_{i,L}^j$.

According to this definition, the most similar pattern to subsequence $X_{i,L}^j$ is MPI_i^j with the similarity distance of MP_i^j. The authors in [10] defined $ArcCurve$ on top of $MatrixProfile$ in their shape-based segmentation technique.

Definition 5. Arc_i^j is an arc between $X_{i,L}^j$ and MPI_i^j and *ArcCurve*, *AC*, is a vector of the same length of time-series X^j where AC_t^j denotes how many arcs, $Arc_i^j, 1 \leq i \leq N - L + 1$, cross the tth time tick in jth channel.

Problem definition. *Given the multi-dimensional dataset X of d time-series, we attempt to detect the $k - 1$ transition times t_1,t_2,...,t_{k-1} that are indicative of state changes in X. These transition times represent the boundaries needed to extract k segments.* Our proposed method consists of three steps:

(1) Extracting potential segment boundary candidates by analyzing the temporal shape across all dimensions;
(2) Employing a greedy search over boundary candidates in order to identify the set of segments with a minimum average entropy;
(3) Ranking channels and estimating the number of segments.

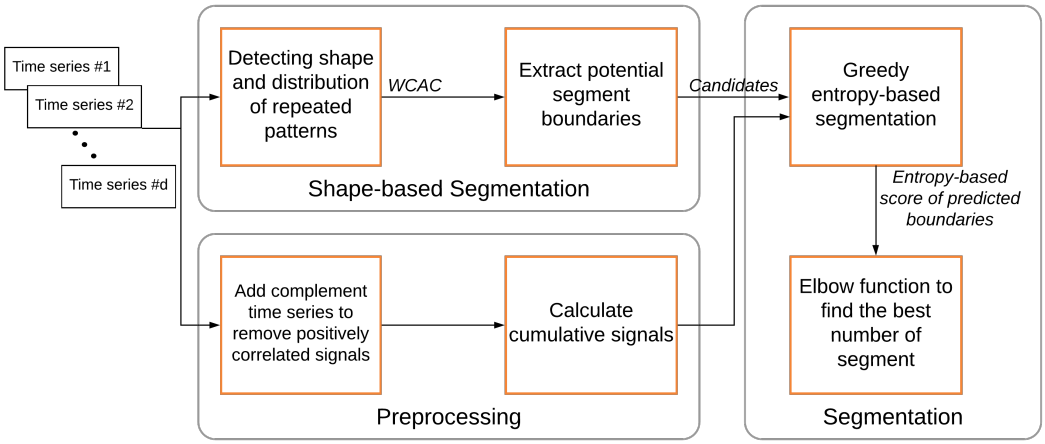


Fig. 2. The workflow of the proposed time-series segmentation method(*ESPRESSO*).

Figure 2 provides a visual overview of the work flow associated with our method.

4 ESPRESSO: METHODOLOGY

In this section we describe different parts of our proposed segmentation technique.

4.1 Shape-based Segmentation

The main assumption of our shape-based segmentation methods is that repeated patterns relate to the same class segments, and hence, occur within close temporal proximity. To extract the most similar shapes within each time-series, we utilized the *MP* technique [50] and the *Arc* definition [10]. The FLOSS [10] method works under the principle that the *AC* will have its minimum values at segment boundaries based on the assumption that a large majority of arcs will be confined to individual segments with very few arcs crossing over segments. This assumption is more likely to be violated, however, if the dataset contains repeated segments. Figure 3 shows an example of a sequence of physical activities that starts with jumping, is followed by running and then returns to jumping again. In this case, running subsequences can find their most similar subsequence within the alternative running segment. This leads to a larger number of arcs spanning over the intermediate jumping segment, and hence, higher AC values being produced across this intermediate segment. This degrades the ability of FLOSS to estimate the activity transition times (comparing Figure 3(b) and Figure 3(c)). In order to address this issue, FLOSS [10] defined a temporal constraint parameter to ignore arcs that are longer than a threshold. Setting this threshold, however, requires having detailed knowledge about the particular problem in question. Furthermore, a threshold places an upper limit upon the segment size that can be considered; this is undesirable from an algorithm design perspective. To address this problem, we propose a novel time-series representation primitive named *Weighted Chained Arc Curve* (*WCAC*) to capture the density of pattern repetition with time. The *WCAC* evaluates the arc Arc_i^j according to the temporal distance between each pair of similar subsequences. We show in Section 5 that we can achieve a far more accurate segmentation result by using *WCAC* when compared to *FLOSS*.

To define *WCAC*, we first explain how a chain of similar arcs is calculated. To increase the robustness of the representation to noise and signal drift, we modified the *AC* definition to consider a chain of similar subsequences instead of only the most similar subsequence. The authors in [53] proposed the time-series *Chain* as a new

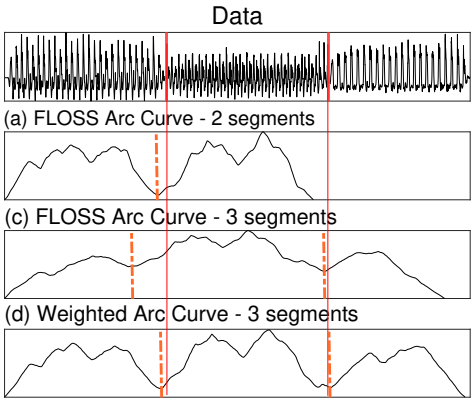


Fig. 3. a)Accelerometer data for a sequence of [jumping,running,jumping]; b)*FLOSS* over the first two segments; c)*FLOSS* over all segments; d)WCAC over all segments.

primitive to sit on top of the MP representation. We modified this chain definition to fit our problem. We believe that considering a chain of patterns is crucial in the context of Human Activity Recognition (HAR), given many activities are associated with motion patterns that can drift with time. We define *ChainedArcCurve* as follows:

Definition 6. *Chained Arc Curve*, $CAC_i^j = \{x_{s1}, x_{s2}, ..., x_{sl}\}, (1 \leq s1..sl \leq n - L + 1)$, is a set of similar subsequences in the jth channel of input. CAC_i^j is ordered in terms of their temporal distance to X_i^j, and for any $1 \leq si \leq l$, we have $x_{si} = MPI^j_{x_{si-1}}$ and $x_{s1} = x_i$, where n is the length of channel and L is the size of the subsequence, l denotes the length of the chain.

Figure 4 illustrates a simplified version of *CAC* with arc chains of second order neighbours. If X_2 is the nearest neighbour (the most similar subsequence) of X_1, and X_3 is the nearest neighbour of X_2, then we add an arc between X_1 and X_3 (if there is no arc yet). The distance of the new arc will be considered as $D(X_1, X_2) + D(X_2, X_3)$. Any higher order arcs will be included in the *CAC* if the distance is less than the specified threshold. To avoid trivial matches, we ignore similar subsequences in an exclusion zone of $L/2$ samples before and after the location of the query. In order to ensure the extracted patterns in the chain are similar, we limit the length of the chain by defining a distance threshold between the first and last subsequence of the chain.

The other modification is to consider the locality of repeated patterns. Each arc in the *CAC* is assigned a weight as an inverse function of its length. Consequently, arcs of a smaller length were provided with greater weight, given they were more likely to belong to the same segment instance than the arcs of a greater length, which

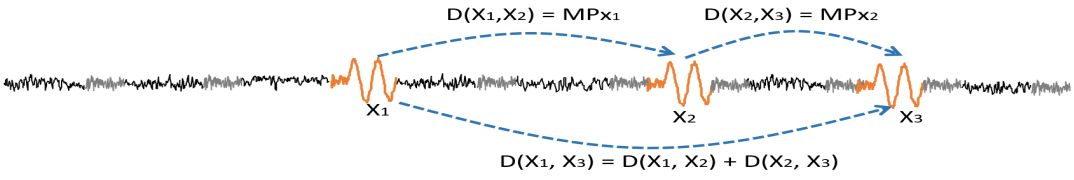


Fig. 4. Chained Arc Curve example.

were more likely to cross over other segments. *WCAC*, is defined as follows:

$$WCAC_i^j = \sum_A \frac{MP_a^j}{|a - MPI_a^j|}, A = \{Arc_a^j \in CAC^j | i \in [a, MPI_a^j]\} \quad (1)$$

Figure 3(d) shows that *WCAC* is a more effective representation to estimate change points, as unlike the original *AC* (Figure 3(c)), there are two local minima in close proximity to the actual segment boundaries.

Algorithm 1: ExtractCAC

```
   Data: MPindex, MP
   Result: Chain of extracted arcs for each
           subsequence(CAC), NewArcsSimilarity
1  CAC= MPindex;
2  newArcs = MPindex(MPindex);
3  while newArcs do
4      NewArcsSim = MP + MP(newArcs);
5      for each arc in newArcs do
6          if newArcsSim(arc) < threshold then
7              add arc to CAC;
8          else
9              remove arc from newArcs ;
10         end
11     end
12     newArcs = MPindex(newArcs);
13 end
14 return CAC, newArcsSim;
```

Algorithm 2: ExtractWCAC

```
   Data: MP, MPindex
   Result: Weighted repesentation of chain of
           arcs(WCAC)
1  CAC, NewArcsSim = ExtractCAC(MP,MPindex);
2  #initialize ;
3  temporalDistance = normalize_distance(CAC);
4  WCAC = zeros(length(MP));
5  #calculate weight for each extracted arc in chains;
6  for each arc in CAC do
7      weights =
         NewArcsSim(arc)/temporalDistance(arc);
8      WCAC = WCAC(arc_start - arc_end) + weights ;
9  end
10 return WCAC;
```

The algorithm to compute the *CAC* and *WCAC* are provided in Algorithm 1 and 2 respectively. The chain (*CAC*) is initialized with the MP index that represents the arcs between each subsequence and its nearest neighbour subsequence, which is the most similar subsequence too (line 1). A set of second-level arcs are then constructed between each subsequence and the nearest neighbor of its own nearest neighbor (line 2). We then look for next level arcs to add to the current chain (lines 3-13). If the new arc meets the distance condition (line 6), it will be added to the chain in line 7. This process is repeated until there are no new arcs to add. To calculate *WCAC* (Algorithm 2), for each arc in *CAC*, the weight will be updated according to the similarity (*NewArcsSim*) and normalized arc length (*temporalDistance*) in lines 6-9.

Change point candidates are estimated based upon the intuition that actual change points should often coincide with either a local or global minimum in the *WCAC*. Figure 5 shows *WCAC*, for six different time-series including a sequence of *sitting, running and sitting* activities. The figure clearly illustrates that in each of the *WCAC* channels, there are local minima in close proximity to the actual change points. These change point candidates are used as the search space for the entropy-based segmentation described in the next section.

4.2 Entropy-based Segmentation.

The shape-based segmentation is unable to detect segment boundaries of time-series that have non-repeating shape patterns. Furthermore, the representation can be biased due to the segment size variation across the time-series. Shorter segments are more likely to possess fewer arcs than longer segments, and hence, *WCAC* can be biased towards estimating change points across the shorter segments. Consequently, we utilise the Information

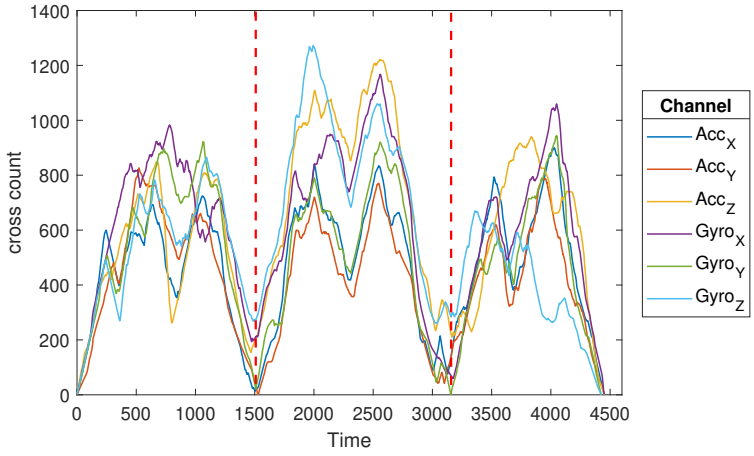


Fig. 5. Example of the Weighted Chain Arc Curve for each of six channels

Gain (IG) metric to evaluate each local minima of *WCAC* as a potential segment boundary. A greedy search procedure is implemented upon the set of boundary candidates in order to identify change points that minimize the entropy-based cost function in (2). Given the first term of the cost function ($H(X, \emptyset)$) is the entropy of the whole time-series as a single segment (no change-point), which have a constant value, maximising IG is equivalent to minimizing the entropy of the constituent segments. The cost function is defined as follows:

$$\mathcal{L} = H(X, \emptyset) - \sum_{i}^{|S' \cup b|+1} \frac{|\mathbf{s}_i|}{|\mathbf{X}|} H(s_i) \tag{2}$$

$$\underset{b \epsilon B}{\text{MAX}}\, \mathcal{L} \tag{3}$$

where X is the time-series of d dimensions, B is the set of segment boundaries that have been selected during previous shape-based segmentation, s_i is the segment between the $i-1$ and ith selected boundaries, and $|.|$ is the length operator. S' is the list of change points that are chosen by our greedy entropy based-method. $H(s_i)$ is the Shannon entropy of the segment i:

$$H(\mathbf{s}_i) = -\sum_{j=1}^{n} p_i^j log p_i^j \tag{4}$$

and p_i^j is the area of segment s_i of series X^j divided by this segment summed across all d time-series.

Algorithm 3 describes the greedy search used to estimate the segment boundaries (*GreedyEntropySearch*), whilst Algorithm 4 explains the complete *ESPRESSO* approach. During each iteration of the *GreedyEntropySeg*, the remaining segment boundary candidates b (not currently in S') were used to split an existing segment of the time-series into two segments. The entropy of the new segments were then computed (line 6). The candidate b that produced the two segments of lowest entropy were then selected and added to the set S'. This greedy search was repeated for each dimension of the time-series in *ESPRESSO* (Algoirthm 4).

Algorithm 3: GreedyEntropySeg

```
Data: TS, candidtes
Result: Selected boundaries (S) and
        corresponding entropy (H)
TT, H, S = empty;
while (!meetKneePoint(H)) do
    minH = Inf;
    for b in candidates do
        TT = sort(TT.append(b));
        h = Entropy(TS, TT);
        if h ≤ minH then
            bestB = b; minH = h;
        end
    end
    S.append(bestB); H.append.(minH);
    candidate.remove(b);
end
return S, H;
```

Algorithm 4: ESPRESSO

```
Data: TS
Result: TT, score
MP, MPindex = calculateMatrixProfile(TS);
WCAC = ExtractWCAC(MP, MPindex);
candidates = findLocalMinima(WCAC);
for each dimenssion do
    TT[dim], entropy[dim] =
      GreedyEntropySeg(TS,candidates[dim]);
end
best_index = IndexOfMin(entropy);
return TT[best_index], entropy[best_index];
```

4.3 Channel Ranking

In this study, we showed that the segmentation accuracy was positively correlated with the entropy of the estimated segments in (2). Consequently, we attempt to exploit this discovery to rank the channels of our multi-dimensional time-series X and select the channel that was most likely to provide a higher segmentation accuracy than X. For each channel j in X, a set of boundary candidates (B^j) were estimated from its $WCAC^j$ representation. A greedy search of B^j (as outlined in section 4.2) was then used to segment X based upon maximising the IG metric of (2).

4.4 Estimating the Number of Segments

The number of segments were estimated by analyzing the extent to which a new segment contributes to decreasing the entropy of the segmented time-series. The relationship between entropy and the number of segments have been proven to be monotonic increasing [35], given the entropy of constituent segments will always decrease as k is increased. The following knee point detection equation in (5), which was proposed by [35], was used to estimate k, where $\mathcal{L}_k$ denotes the reduced entropy for k segments.

$$\underset{k}{\text{MAX}}\left(\frac{\mathcal{L}_k - \mathcal{L}_{k-1}}{\mathcal{L}_{k+1} - \mathcal{L}_k}\right) \tag{5}$$

5 EXPERIMENTS AND EVALUATION

We introduce the seven datasets and four benchmark segmentation methods used in our experiment. The metrics used to evaluate the performance of ESPRESSO and the benchmark segmentation methods are then defined. Finally, the results of the experiment and some use-case studies are presented. The source code is available at the GitHub page: https://github.com/cruiseresearchgroup/ESPRESSO.

Table 2. Properties of the datasets

Dataset	Context	Sensors	Length[1]	Dimension[2]	Segments[3]
HandGesture[5]	10 hand gestures	Accelerometer,Gyrometer	133.3K	18	600
PAMAP[32, 33]	14 physical activities	Accelerometer,Temperature	42.4k	10	21
USC-HAD [52]	12 physical activities	Accelerometer,Gyrometer	93.6k	6	36
EYE state[34]	close/open eye	EEG	2k	8	5
WESAD[41]	Stress, amusement, and meditation state	ECG, EMG, EEG, Respiration	~300k	4	8
RFID[49]	12 physical activities	RFID Readers	15.3K	12	84
Emotion[15]	Neutral, negative and positive pleasure	EDA, EMG, BVP, ECG, Temp,Respiration,Heartrate	~37k	10	4

* [1] Length of each time-series, [2] Number of dimensions, and [3] Number of segments.

Table 3. Categories of datasets. Datasets that contain both features are marked with ✓;✗

Feature \ Dataset	PAMAP	RFID	Hand Gesture	USC-HAD	WESAD	EYE	Emotion
Continuous (✓) / Non-Continuous (✗) segments	✗	✗	✓	✗	✗	✓	✓
Repetitive (✓) / Non-Repetitive (✗) patterns	✓;✗	✓;✗	✓;✗	✓;✗	✗	✗	✗

5.1 Datasets

Seven public datasets comprised of smartphones, RFID tags and different wearable sensors including motion sensors, physiological sensors and eye wear computing sensors have been used to test the segmentation performance of *ESPRESSO* and benchmark methods. Table 1 provides a detailed summary of the seven public datasets used in this experiment. We evaluated our method on two different types of wearable sensor datasets; continuous (C) and non-continuous (NC). The continuous datasets were comprised of sensor data collected across an uninterrupted sequence of different human activities. In these datasets, the activity transition times were manually recorded by human observers. The second type of datasets were comprised of individual recordings of human activity that were manually stitched together to form an activity sequence. Consequently, the transition between adjacent activities were far more discontinuous in the second type of datasets than the first type. In addition, the datasets were categorised based on whether they contained repetitive temporal patterns (R) or were exclusively non-repetitive patterns (NR). Table 3 presents the datasets and categories they are each associated with.

USC Human Activity Dataset (USC-HAD). [1] [52]: The USC-HAD dataset includes twelve human activities that were each recorded separately across fourteen subjects. Each human subject was fitted with a 3-axis accelerometer and a 3-axis gyrometer that was attached to the front of the right hip and sampled at 100Hz. Activities were repeated five times for each subject and consisted of: walking forward, walking left, walking right, walking upstairs, walking downstairs, running forward, jumping up, sitting, standing, sleeping, elevator up, and elevator down. To perform experiments, the different set of activities were manually stitched randomly, therefore, USC-HAD was considered a NC dataset.

[1] http://sipi.usc.edu/had/

Physical Activity Monitoring for Aging People (PAMAP). [2] [32, 33]: This dataset includes fourteen low-level (such as walking and sitting) and high-level (such as ironing which consists of two or more low-level) human activities undertaken by eight different subjects. Each subject was fitted with an IMU (inertial measurement units) sensor on their wrist, chest and ankle. Each participant was given both an indoor and outdoor activity schedule to perform the following activities sequentially: lying, sitting, standing, walking very slow, normal walking, Nordic walking, running, ascending stairs, descending stairs, cycling, ironing, vacuum cleaning, jumping rope and playing soccer. Each IMU collected observations of temperature, 3-axis acceleration, 3-axis angular velocity (gyroscope), and the 3-axis magnetic field (magnetometer) at a sampling rate of 100Hz. As a result of missing readings in some of the sensors, only a subset of this IMU dataset was used in the experiment; the data from all three accelerometers and the hand fitted thermometer.

Hand Gesture Dataset. [3] [5]: Our experiment used the Hand Gesture dataset, a collection of twelve hand movement activities performed by two subjects. Activities were captured by three IMUs that were attached to the subject's hand, upper arm and lower arm, respectively. The activities that were recorded within the experiments included: opening the window, closing the window, drinking, watering plants, cutting, chopping, stirring, reading a book, a tennis forehand, a tennis backhand and a tennis smash.

Device-free posture recognition by RFID (RFID). [49]: In this experiment nine passive RFID tags were placed on a wall. The experimental dataset consist of six subjects that each performed twelve predefined postures between the wall and an RFID antenna. Each posture was performed for 60 seconds. RFID was a NC dataset given it was formed by concatenating the twelve postures for each of the six subjects.

Emotion Dataset. [4] [14, 15]: This dataset has been collated to study the physiological response to different emotional states and to identify the effect of physical activity on these emotion state. Five hours of physiological data from E4-wristband, Biosignalsplux device, and smart-phone were collected from 18 subjects with respect to three different emotion categories, High Positive Pleasure High Arousal (HPHA), High Negative Pleasure High Arousal (HNHA), and Neutral.

WESAD Dataset. [5] [41]: WESAD is a well-known dataset for stress that has been acquired from multi-modal wearable sensors. This dataset is comprised of physiological and motion data from chest and wrist-worn sensors of 15 subjects. In this study, only chest-worn sensors with a down-sampling rate of 10 were used to detect stress, amusement and meditation segments.

EYE State Dataset. [6] [34]: This dataset consists of 14980 samples of 15 EEG sensors collecting eye state data for 117 seconds. The labels (close/open) are manually annotated using video collected during the measurements.

5.2 Benchmark Methods

The performance of the proposed *ESPRESSO* method was compared to four state-of-the-art algorithms: Fast Low-Cost Semantic Segmentation (*FLOSS*) [10], Information Gain-based Time-series Segmentation (*IGTS*) [35], additive Hilbert-Schmidt Independence Criterion (*aHSIC*) [47], and Relative unconstrained Least Square Importance Fitting (*RuLSIF*) [27]. To avoid inconsistencies and implementation errors, we evaluated our method against benchmark algorithms with publicly available source code.

[2] http://www.pamap.org
[3] https://github.com/andreas-bulling/ActRecTut
[4] https://www.comtec.eecs.uni-kassel.de/emotiondata/
[5] https://ubicomp.eti.uni-siegen.de/home/datasets/icmi18
[6] https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State

FLOSS is a shape-based segmentation method that sits on top of the Matrix Profile time-series representation, and *IGTS*, is an Information gain based segmentation method, which have both been described in previous sections. *IGTS* requires no input parameters, whilst *FLOSS* requires the subsequence length as its input parameter. *RuLSIF* is based on estimating the relative probability density ratio of subsequences. The number of subsequences in each round and regularization constant were fixed at 10 and 0.01, respectively, as suggested by the authors. To enable a fair comparison with this method, we evaluate its performance across different subsequence lengths. The final method, *aHSIC*, is a multi-dimensional time-series segmentation combined with channel selection. Firstly, this method selects important channels using a supervised learning method and then scores each time step according to the proposed dependency measure regarding a pseudo-binary-label. The regularization constant and the kernel parameter was set as 0.01 and 1, respectively, as suggested in their paper. Segmentation performance was compared across a range of subsequence sizes that were unique to each dataset based on its sample rate and the minimum segment duration of 0.5 seconds. The range of subsequence sizes varied from the narrowest set of 10 to 40 samples for the RFID dataset to the widest set of 100 to 900 samples for the PAMAP dataset.

5.3 Evaluation Metrics

The performance of the segmentation algorithms were evaluated with respect to the following metrics:

(1) F-score: the F-score is defined as the harmonic mean of the Precision ($\frac{TP}{TP+FP}$) and Recall ($\frac{TP}{TP+FN}$). Each estimated segment was defined as a True Positive (TP) when it was located within a specified time window of the ground truth segment boundaries and a False Negative (FN) when it fell outside the time window of all the ground truth segment boundaries. When multiple segment estimates fell within a specified time window of the ground truth segment boundary, only the closest estimate was considered to be TP and the remaining estimates were considered to be False Positives (FP). As a consequence of the sampling rate of sensors in the dataset, the time window (i.e. segmentation threshold) was set to 0.5 seconds for the EYE dataset and 2 seconds for each of the remaining datasets.
(2) RMSE: The Root Mean Square Error (RMSE) was computed between the ground truth segment boundary time and its nearest estimated segment boundary time. The RMSE was then normalized into the range of [0, 1] by dividing it by the time-series duration.
(3) MAE: To compare the performance of the proposed shape-based segmentation method, *WCAC*, and the other shape-based segmentation method, *FLOSS*, we employed the Mean Absolute Error (MAE) as used in [9]. For this particular study, segmentation performance was evaluated as the MAE between the estimated segment boundaries and ground truth segment boundaries.

The F-score metric depends upon the selection of a threshold value. For example, consider the actual segment transition time is at 250 seconds and a threshold of five seconds is set. Suppose two segmentation methods estimate segments boundaries, A and B, at 256 seconds and 300 seconds, respectively. The F-score metric will evaluate the A and B estimates as False Negatives despite A being a far superior estimate to B. The RMSE and MAE, will address this problem given their continuous metric space ensures A will be represented as a superior estimate to B. Typically all metrics might incorporate the error of several transition estimates that are in closest proximity to a single ground truth boundary. Consequently, we ensure each ground truth boundary is exclusively mapped to only one segment boundary estimate to ensure that metrics are not biased by change point estimates being clustered around a subset of segment boundaries. Existing studies often evaluate performance using one of these metrics, however, we believe including the F-score as well as RMSE (or MAE) provides a more comprehensive evaluation.

5.4 Evaluation

In this section, we first compare the effectiveness of the proposed shape-based segmentation technique, *WCAC* against *FLOSS*. Then, we investigate the performance of *ESPRESSO* against four state-of-the-art segmentation techniques. Finally, we undertake an ablation study to compare the performance of the shape and entropy based components of *ESPRESSO*.

5.4.1 WCAC. In this section, we compare the performance of our proposed shape-based segmentation method, *WCAC*, and state of the art shape based segmentation method *FLOSS*. To compare the effectiveness of these methods, the Hand Gesture, USC-HAD and RFID datasets were selected given they contained a diverse set of sensors and contained repetitive temporal patterns. The experiments were repeated over a set of subsequence lengths ranging from between 10 and 40 samples for the Hand Gesture dataset 6(a), 50 and 550 samples for the USC-HAD dataset 6(b) and 20 and 100 samples for the RFID dataset 6(c). The subsequence lengths were set based upon the sampling rate of each dataset. The minimum subsequence length was set to 0.5 seconds, whilst the maximum subsequence length was set to half of the minimum segment size in the dataset.

Figure 6 shows the segmentation performance with respect to the Mean Absolute Error (MAE). These figures show that the proposed *WCAC* method does not only have consistently superior segmentation performance to *FLOSS*, but was far more insensitive to the subsequence length that was used.

5.4.2 ESPRESSO. The performance of *ESPRESSO* was compared to four competing segmentation techniques: *FLOSS*, *IGTS*, *aHSIC*, and *RuLSIF*. We performed extensive experiments across seven datasets and over a range of different subsequence lengths. Figure 7 compares the F-score of the *ESPRESSO*, *aHSIC*, *RulSIF* and *FLOSS* methods with respect to the subsequence length across three datasets; Hand Gesture, PAMAP and RFID. The *IGTS* method was not included in this comparison given it does not utilise subsequences in order to perform segmentation.

For the PAMAP and RFID datasets, *ESPRESSO* was shown to offer a very high level of segmentation performance that was superior to three benchmark algorithms across all subsequence lengths, apart from a small set of subsequence lengths (with 33 to 37 samples) in the RFID dataset where aHSIC's performance was equivalent. In addition, ESPRESSO's performance was consistently high across all subsequence lengths, a desirable property of the algorithm.

For the Hand Gesture dataset, *ESPRESSO* achieved superior segmentation performance to the benchmark methods for a large majority of the subsequence lengths. The exception was a narrow range of subsequence lengths (with 35 to 55 samples) where *FLOSS* outperformed *ESPRESSO*. *FLOSS*'s performance was found to be far more sensitive to subsequence length than *ESPRESSO*, given it exhibited a significant performance decline outside of this optimal subsequence range.

Table 4 shows the average F-score and RMSE of *ESPRESSO* and four benchmark methods across each of the seven datasets. For each of the methods apart from *IGTS*, the F-score and RMSE were averaged across all of the subsequence lengths and subjects of each dataset. For *IGTS*, the F-score was only averaged across the dataset subjects given its a top-down method that does not utilise subsequences. In the PAMAP and WESAD datasets, given the computational cost of *aHSIC* became prohibitively high over longer window sizes, the window size was fixed at 50 samples as suggested in their paper.

The segmentation results in Table 4 indicate that *ESPRESSO* was superior to the benchmark methods, on average, across all datasets with an F-Score improvement of 45.6%, 7%, 44.4%, and 45.2% over the *FLOSS, IGTS, aHSIC*, and *RuLSIF* methods, respectively. In addition, *ESPRESSO* had an average RMSE improvement of 140%, 21%, 92% and 224% over the *FLOSS, IGTS, aHSIC*, and *RuLSIF* methods, respectively. *ESPRESSO* had a clear advantage over the *FLOSS, RulSIF, aHSIC* methods across each of the datasets. The RFID dataset was the only one where *IGTS* was shown to be superior to *ESPRESSO* across both performance metrics. *ESPRESSO* was advantageous over *IGTS*

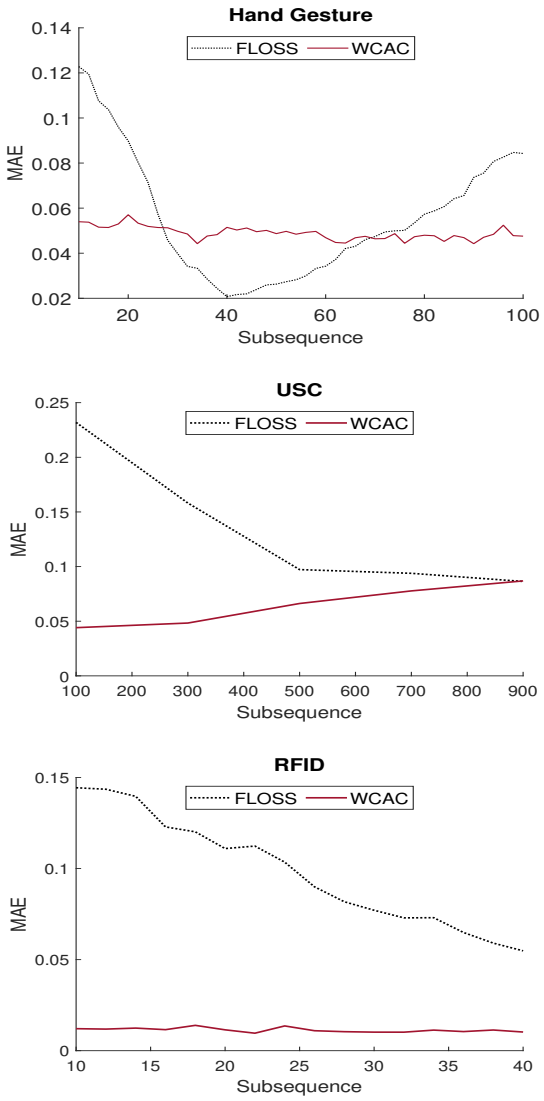


Fig. 6. Comparing the performance of *FLOSS* and the proposed weighted chained arc curve (*WCAC*) for the (a) Hand gesture, (b) USC-HAD, and (c) RFID datasets.

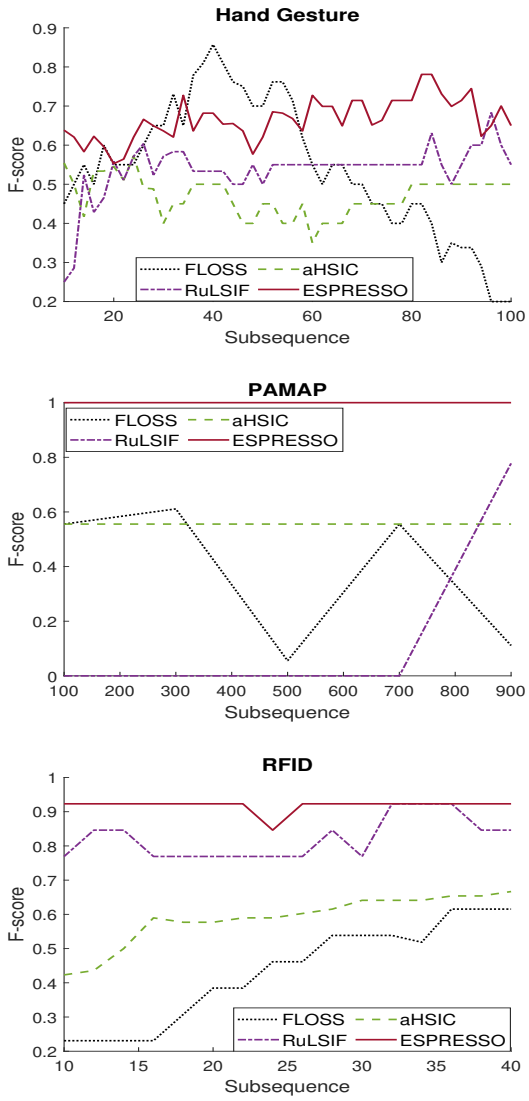


Fig. 7. The segmentation performance (F-score) of the *ESPRESSO* method and three benchmark methods across a range of subsequence lengths for three datasets (a) Hand Gesture (b) PAMAP and (c) RFID

across four of the seven datasets (RFID, Hand Gesture, USC-HAD and WESAD) given both of its performance metrics were superior. We hypothesize that RFID was an optimal dataset for a top-down entropy based approach such as *IGTS*, given the segments had salient statistical differences (as shown in Figure 1(b)).

The effectiveness of *ESPRESSO* was then examined with respect to particular dataset characteristics in Figure 8:

Table 4. Evaluation results

	Feature \ Dataset	PAMAP	RFID	Hand Gesture	USC-HAD	WESAD	EYE	Emotion
F-score	IGTS	1	**0.9554**	0.3825	0.7333	0.6154	0.5116	0.5556
	FLOSS	0.3778	0.4106	0.5379	0.3733	0.1795	0.4252	0.4722
	aHSIC	0.5556	0.7787	0.2188	0.4	-	0.5312	0.00
	RuLSIF	0.1556	0.8560	0.2529	0.4133	0.3667	0.5336	0.2222
	ESPRESSO	**1**	0.9378	**0.6209**	**0.7467**	**0.6410**	**0.5821**	**0.5833**
RMSE)	IGTS	**0.0024**	**0.0401**	0.4270	0.1939	0.2195	0.0997	**0.0607**
	FLOSS	0.2779	0.3969	0.3166	0.3267	0.5140	0.1114	0.1219
	aHSIC	0.1659	0.1411	0.4069	0.3147	-	0.1070	0.2359
	RuLSIF	0.8375	0.1013	0.3792	0.3338	0.2873	0.2727	0.5746
	ESPRESSO	0.0030	0.0692	**0.2764**	**0.1933**	**0.1936**	**0.05**	0.0719

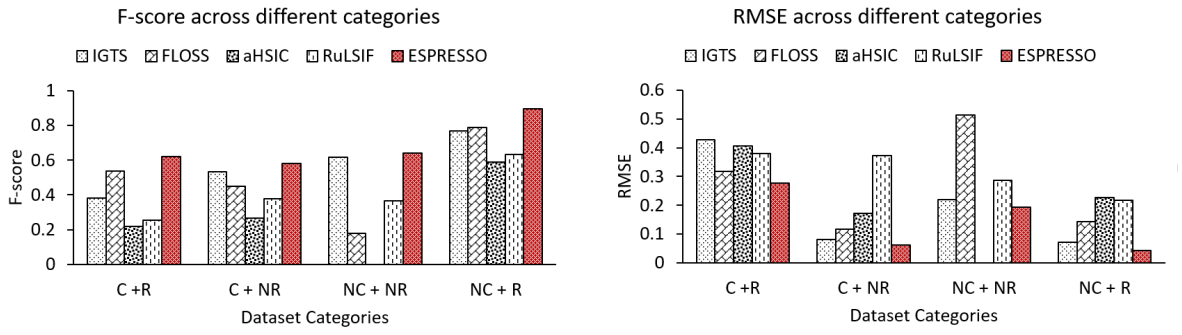


Fig. 8. A comparison of *ESPRESSO* and four benchmark methods across different dataset categories; continuous datasets with repetitive patterns (C+R), continuous datasets with non-repetitive patterns (C+NR), non-continuous datasets with non-repetitive patterns (NC+NR) and non-continuous datasets with repetitive patterns (NC+R) (F-score(left) and RMSE(right)).

- **Datasets with continuous (C) and non-continuous (NC) segments:** the datasets associated with the C and NC categories are shown in Table 3. *ESPRESSO* was shown to be superior to each of the benchmark methods for the C and NC categories with an average F-score of 0.59 and 0.67, respectively. Figure 8 shows *ESPRESSO's* performance advantage over the statistical based methods (*IGTS*, *RulSIF*) was more significant for the continuously recorded datasets than the non-continuous datasets. This suggests *ESPRESSO* was better equipped to detect segment transitions with higher correlation than statistical methods. In contrast, *ESPRESSO's* performance advantage over the shape-based *FLOSS* method was greater for the NC datasets, in particular, the NC datasets with non-repeating temporal patterns.
- **Datasets with repeating temporal patterns (R) and non-repeating temporal patterns (NR):** HAR datasets consist of physical activities that contain both repeating (such as walking or stirring) and non-repeating (such as sitting or opening the window) actions. The association between the seven datasets used in the experiment and action categories are shown in Table 3. In Figure 8, *ESPRESSO*'s performance was shown to be far superior across datasets that possess at least some repeated patterns (R) (average F score of 0.83) when compared to the datasets composed of non-repeating patterns (NR) exclusively (average F

score of 0.6). This can be attributed to *ESPRESSO* using *WCAC* to identify potential segment boundary candidates; *WCAC* is far more effective at detecting segment transitions across time-series with repeating shape patterns. *ESPRESSO* was shown to be superior to each of the four benchmark methods across both categories of temporal patterns. Figure 8 shows *ESPRESSO's* performance advantage was greater across R datasets than NR datasets. This could be attributed to ESPRESSO's unique ability to exploit shape and statistical properties of time-series in the R datasets where both properties are useful to segment the combination of repetitive and non repetitive patterns.

5.4.3 Ablation Study. An additional study is performed to investigate the temporal shape and statistical components of *ESPRESSO* independently. Table 5 compares the segmentation performance of *ESPRESSO* with its constituent shape-based method (*WCAC*) and entropy-based method (*GreedyEntropySeg*). Furthermore, Figure 9 compares the segmentation performance of the *WCAC* and *GreedyEntropySeg* methods across the four dataset categories introduced in section 5.4.2.

The segmentation performance of *ESPRESSO* was largely attributed to the strong and consistent segmentation performance of the *GreedyEntropySeg* method with respect to the NC and NR dataset categories. In contrast, the *WCAC* method was shown to offer a more significant contribution to *ESPRESSO*'s segmentation performance for the R dataset categories (with repeating patterns). This can be attributed to *WCAC* needing to exploit repeated temporal shape patterns in order to perform accurate segmentation of the time-series.

Table 5. Ablation study

	Feature \ Dataset	PAMAP	RFID	Hand Gesture	USC-HAD	WESAD	EYE	Emotion
F-score	WCAC	0.4218	0.5237	0.5381	0.3881	0.1795	0.4404	0.3889
	GreedyEntropySeg.	1	**0.9554**	0.3825	0.7333	0.6154	0.5116	0.5556
	*ESPRESSO**	**1**	0.9378	**0.6609**	**0.7467**	**0.6410**	**0.5325**	**0.5833**
RMSE	WCAC	0.1977	0.2528	0.3221	0.4109	0.5140	0.1	0.1222
	GreedyEntropySeg.	**0.0024**	**0.0401**	0.4270	0.1939	0.2195	0.0997	**0.0607**
	*ESPRESSO**	0.0030	0.0692	**0.2764**	**0.1933**	**0.1936**	**0.05**	0.0719

* Combination of WCAC + GreedyEntropy Segmentation.

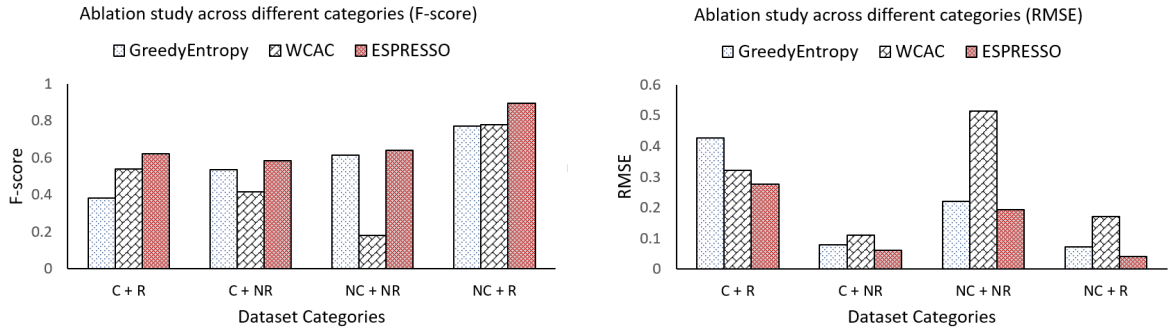


Fig. 9. Investigating the contribution of shape and entropy based methods of *ESPRESSO* across: continuous datasets with repetitive patterns (C+R), continuous datasets with non-repetitive patterns (C+NR), non-continuous datasets with non-repetitive patterns (NC+NR) and non-continuous datasets with repetitive patterns (NC+R) (F-score(left) and RMSE(right)).

6 REAL WORLD CASE STUDIES

6.1 Case Study1: Unsupervised Inference of Work Routines and Behavior Deviations

In this section we show how performing segmentation with the data of wearable sensors can help to extract a user's daily life patterns and to identify any deviations in their daily routines. In this study, we used an existing life-logging dataset from NTCIR-13 Life-logging track [11]. The data from three biometric sensors (calories burnt, heart rate, and skin temperature) and a step counter were used in this study. The aim of this study was to model the physical activity intensity of a user on a daily basis and then compare these in order to detect any activity deviations [8]. To extract different levels of activity intensity, we changed the granularity of the estimated segments from 4 segments to 11 segments per day.

Figure 10 shows the average extracted transition times over 19 weekdays (orange) and 8 holiday/weekend (blue). As we increase the number of segments in a day, the granularity of the task routine increased. There are several distinct differences between weekdays and weekends in terms of transition times, segment lengths at different times of the day, and the timing of the first and last estimated segments (morning and evening). For example, during the weekdays, the user usually starts the day at around 4:00 am, whilst at weekends, this segment starts at approximately 5:00 am. It should be noted that this start time is associated with changes in the biometric parameters and may be related to the subject's biological clock. It also shows that users are likely to wake up later on the weekends due to having no work obligations. There was similar shifts in segment transitions in the middle of the day (noon) and later in the day (evening) which are highlighted using grey arrows in the Figure. To detect unusual daily patterns, we devised a threshold-based algorithm to find any deviation from normal daily routines. We analyzed the deviation in daily patterns from the reference routine and consider it as an *unusual day* if the dissimilarity metric was greater than a predefined threshold. Using the corresponding images as ground truth, we attempt to explain the reasons behind the strong deviation in activity levels. Table 6 shows the *unusual days* that had been identified and provides an interpretation of each day.

6.2 Case Study2: Detecting Emotion Changes

The well-known stress and emotional affect dataset WESAD [41] were analysed with a new perspective. This dataset includes wearable physiological and motion sensors commonly used in medical applications such as electrocardiogram (ECG), electromyogram (EMG), Blood Volume Pulse (BVP), temperature, electrodermal activity (EDA), respiration and the accelerometer. Data is labelled in 5 different categories including: Baseline, Stress, Amusement, Medication and Not Defined (ND).

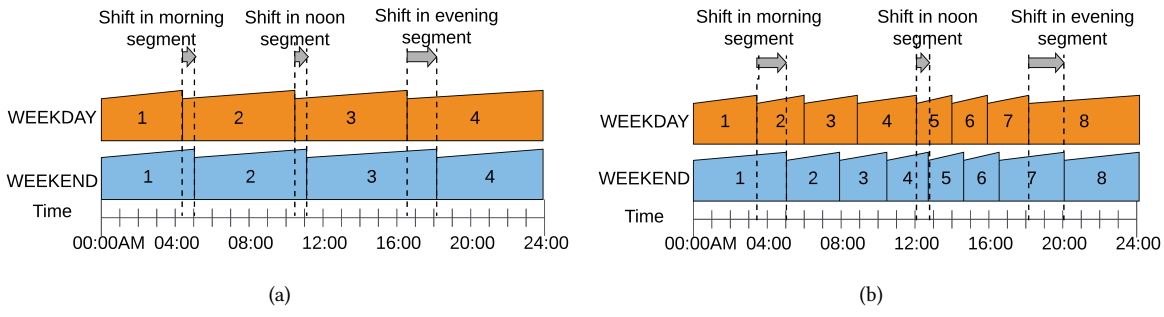


Fig. 10. Inferring daily routine from weekday(orange) and weekends(blue) dividing each day into: a) 4, b) 8 segments.

Table 6. List of identified unusual days

Date	Reason of deviation
17-8-2016	The user left work earlier to shop and have lunch
24-8-2016	The user did not go to work.
29-8-2016	The user caught a bus instead of driving
30-8-2016 and 8-9-2016	The user caught a flight and then went back to work.

Through this experiment we aim to answer the following questions: 1) What sort of emotion transitions (for example transition from"Stress" to "Meditation") are more detectable? 2) How long does it take for a physiological response to occur for each category of emotion?

We applied *ESPRESSO* to the data of 15 subjects in two sets of experiments. The first involved estimating the emotion transition times between "Stress", "Amusement" and "Meditation" segments. The second involved estimating emotional transition times across the entire set of data containing Not Defined segments. For each experiment, we evaluated the estimated emotion transitions in terms of the True Positive Rate (TPR). Section 5.3 explains how we consider an extracted boundary as the True Positive. TPR is calculated by dividing the number of true positives by the total number of segments. The detection threshold was changed from 15 seconds to 275 seconds which accounts for the delay in the physiological response. Figure 11 shows the TPR across different threshold values for the emotion transitions. These figures show that transitions into and out of the "Stress" state can be accurately detected in less than 100 seconds. This is due to the strength of the physiological response to "Stress" and hints at the negative impact that stress can have on human health. In contrast, we hypothesise the "Amusement" to "Meditation" transitions are detected with lower accuracy for several reasons. Firstly, the duration of "Amusement" segments is much smaller than other emotion states. Secondly, the "Amusement" emotion is unlikely to have as strong a physiological effect as "Stress" may have. Thirdly, subjectivity is often introduced into such experiments. For the "Amusement" segments, subjects are provided with 11 "funny" clips,

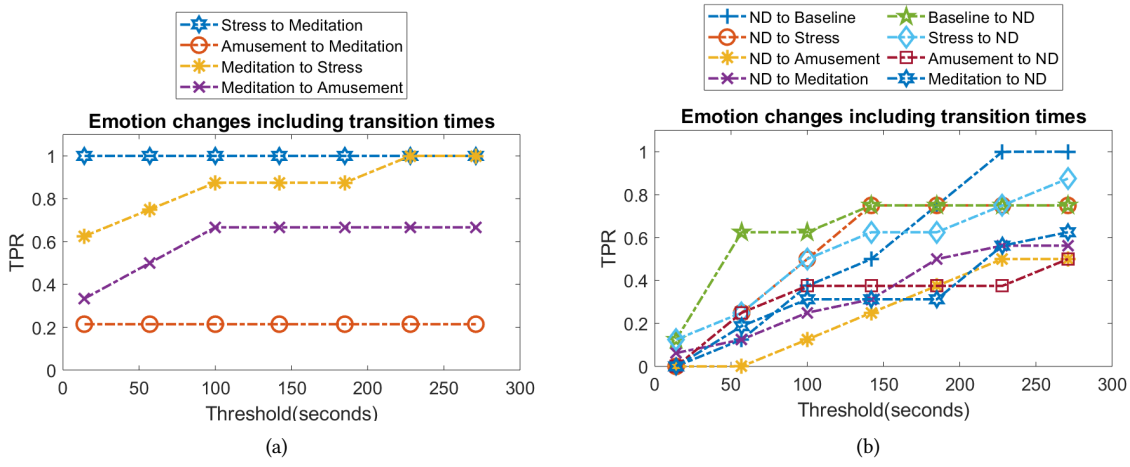


Fig. 11. The detection of emotion transitions for (a) three emotion states and the (b) whole WESAD dataset

however, these clips may not be found amusing by all subjects. We believe this study opens a new avenue towards improving personality-inference based applications by considering the subject's reactions and response time in different situations dynamically.

7 CONCLUSION

We propose a novel unsupervised method for multivariate time-series segmentation, *ESPRESSO*, and test it on a range of wearable sensing and device-free applications. *ESPRESSO* has a hybrid formulation that enables time series to be segmented on the basis of its temporal shape and statistical properties (i.e. entropy). The proposed temporal shape representation, the Weighted Chained of Arc Curve (*WCAC*), was used to detect potential candidates of segment boundaries. Segments were then estimated with the *GreedyEntropySeg* method that performed a greedy search upon this limited space of boundary candidates using an entropy-based metric.

Our proposed shape segmentation method, Weighted Chained of Arc Curve, *WCAC*, was shown to consistently outperform a state of the art temporal shape based method across three public datasets. This improvement was attributed to our novel primitive, *WCAC*, addressing the limitations of current shape based methods such as segmenting repeated segments or temporal shape patterns that drift over time.

Experiments were run across a diverse set of seven public datasets of wearable and device free sensors and showed that *ESPRESSO* achieved an average segmentation performance improvement (in terms of F-score and RMSE) over four state of the art methods *FLOSS*, *aHSIC*, *RulSIF* and *IGTS*. Furthermore, it was demonstrated that *ESPRESSO* outperformed the four benchmark methods across different categories of the data related to the repetition of patterns and the continuity of segments. An ablation study of *ESPRESSO* demonstrated the *WCAC* method offered a more significant contribution to segmenting time series with repetitive patterns, whilst the *GreedyEntropySeg* method offered a greater contribution to segmenting time series with non-repetitive patterns and time series composed of non-continuous segments.

We demonstrated the value of using ESPRESSO in two real-world use-cases of inferring daily activity routines and emotional state transitions in an unsupervised context.

Future work will involve extending the current method to an online version of the channel ranking algorithm where the set of channels accounting for system change can be dynamically selected. Currently, we use the highest ranked channel for segmentation. We also aim to consider subset of channels and their correlation for future segmentation based channel selection.

ACKNOWLEDGMENTS

We acknowledge the support of Australian Research Council Discovery DP190101485, and CSIRO Data61 Scholarship program (Grant number 500588). The authors would like to acknowledge Dr Eamonn Keogh's invaluable comments during the early discussions of this project.

REFERENCES

[1] Samaneh Aminikhanghahi and Diane J. Cook. 2017. A Survey of Methods for Time Series Change Point Detection. *Knowledge and information systems* 51, 2 (01 May 2017), 339–367.

[2] Samaneh Aminikhanghahi and Diane J Cook. 2017. Using change point detection to automate daily activity segmentation. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. 262–267.

[3] Samaneh Aminikhanghahi and Diane J Cook. 2019. Enhancing Activity Recognition Using CPD-based Activity Segmentation. *Pervasive and Mobile Computing* 53 (2019).

[4] Michèle Basseville, Igor V Nikiforov, et al. 1993. *Detection of Abrupt Changes: Theory and Application*. Vol. 104. prentice Hall Englewood Cliffs.

[5] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A Tutorial on Human Activity Recognition Using Body-worn Inertial Sensors. *ACM Computing Surveys (CSUR)* 46, 3 (2014), 1–33.

[6] Faicel Chamroukhi, Samer Mohammed, Dorra Trabelsi, Latifa Oukhellou, and Yacine Amirat. 2013. Joint Segmentation of Multivariate Time Series with Hidden Process Regression for Human Activity Recognition. *Neurocomputing* 120 (2013), 633–644.

[7] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R Millán, and Daniel Roggen. 2013. The Opportunity challenge: A benchmark Database for On-Body Sensor-based Activity Recognition. *Pattern Recognition Letters* 34, 15 (2013), 2033–2042.

[8] Shohreh Deldari, Jonathan Liono, Flora D Salim, and Daniel V Smith. 2019. Inferring Work Routines and Behavior Deviations with Life-logging Sensor Data. In *Proceedings of ACM International Conference on Web Search and Data Mining (WSDM) workshop on Task Intelligence (TI@WSDM) (2019)*. ACM.

[9] Shaghayegh Gharghabi, Yifei Ding, Chin-Chia Michael Yeh, Kaveh Kamgar, Liudmila Ulanova, and Eamonn Keogh. 2017. Matrix Profile VIII: Domain Agnostic Online Semantic Segmentation at Superhuman Performance Levels. In *IEEE International Conference on Data Mining (ICDM) (2017)*. IEEE, 117–126.

[10] Shaghayegh Gharghabi, Chin-Chia Michael Yeh, Yifei Ding, Wei Ding, Paul Hibbing, Samuel LaMunion, Andrew Kaplan, Scott E Crouter, and Eamonn Keogh. 2019. Domain Agnostic Online Semantic Segmentation for Multi-dimensional Time Series. *Data Mining and Knowledge Discovery* 33, 1 (2019), 96–130.

[11] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Bernd Munzer, Rami Albatal, Frank Hopfgartner, Liting Zhou, and Duc-Tien Dang-Nguyen. 2019. A Test Collection for Interactive Lifelog Retrieval. In *International Conference on Multimedia Modeling*. Springer, 312–324.

[12] Juan Haladjian. 2019. The Wearables Development Toolkit: An Integrated Development Environment for Activity Recognition Applications. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 3, 4 (2019), 1–26.

[13] David Hallac, Peter Nystrup, and Stephen Boyd. 2019. Greedy Gaussian Segmentation of Multivariate Time Series. *Advances in Data Analysis and Classification* 13, 3 (2019), 727–751.

[14] Judith S Heinisch, Christoph Anderson, and Klaus David. 2019. Angry or Climbing Stairs? Towards Physiological Emotion Recognition in the Wild. In *IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 486–491.

[15] Judith S. Heinisch, Isabel Fernanda Hübener, and Klaus David. 2018. The Impact of Physical Activities on the Physiological Response to Emotions. In *IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 824–829.

[16] Anna Huang, Dong Wang, Run Zhao, and Qian Zhang. 2019. Au-Id: Automatic User Identification and Authentication Through the Motions Captured from Sequential Human Activities Using RFID. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 3, 2 (2019), 1–26.

[17] David Tse Jung Huang, Yun Sing Koh, Gillian Dobbie, and Russel Pears. 2014. Detecting Changes in Rare Patterns from Data Streams. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. Springer, 437–448.

[18] Masashi Kawahara and Yoshinobu Sugiyama. 2012. Sequential Change-Point Detection Based on Direct Density-Ratio Estimation. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 5, 2, 114–127.

[19] Y. Kawahara, T. Yairi, and K. Machida. 2007. Change-Point Detection in Time-Series Data Based on Subspace Identification. In *Seventh IEEE International Conference on Data Mining (ICDM)*. IEEE, 559–564.

[20] Agnes WK Lam, Danniel Varona-Marin, Yeti Li, Mitchell Fergenbaum, and Dana Kulić. 2016. Automated Rehabilitation System: Movement Measurement and Feedback for Patients and Physiotherapists in the Rehabilitation Clinic. *Human–Computer Interaction* 31, 3-4 (2016), 294–334.

[21] Wei-Han Lee, Jorge Ortiz, Bongjun Ko, and Ruby Lee. 2018. Time Series Segmentation through Automatic Feature Learning. *CoRR* abs/1801.05394 (2018). arXiv:1801.05394 http://arxiv.org/abs/1801.05394

[22] Zhen Li, Zhiqiang Wei, Wenyan Jia, and Mingui Sun. 2013. Daily Life Event Segmentation for Lifestyle Evaluation based on Multi-sensor Data Recorded by a Wearable Device. In *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society)*. 2858–2861.

[23] Jonathan Feng-Shun Lin, Michelle Karg, and Dana Kulić. 2016. Movement Primitive Segmentation for Human Motion Modeling: A Framework for Analysis. *IEEE Transactions on Human-Machine Systems* (2016), 325–339.

[24] Jonathan Feng-Shun Lin and Dana Kulić. 2013. Online Segmentation of Human Motion for Automated Rehabilitation Exercise Analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 22, 1 (2013), 168–180.

[25] Jonathan Liono, A Kai Qin, and Flora D Salim. 2016. Optimal Time Window for Temporal Segmentation of Sensor Streams in multi-activity recognition. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. 10–19.

[26] Li Liu, Yuxin Peng, Shu Wang, Ming Liu, and Zigang Huang. 2016. Complex Activity Recognition Using Time Series Pattern Dictionary Learned from Ubiquitous Sensors. *Information Sciences* 340 (2016), 41–57.

[27] Song Liu, Makoto Yamada, Nigel Collier, and Masashi Sugiyama. 2013. Change-point Detection in Time-series Data by Relative Density-Ratio Estimation. *Neural Networks* 43 (2013), 72–83.

[28] Qin Ni, Tim Patterson, Ian Cleland, and Chris Nugent. 2016. Dynamic Detection of Window Starting Positions and its Implementation within an activity recognition framework. *Journal of Biomedical Informatics* (2016).

[29] Mohd Halim Mohd Noor, Zoran Salcic, I Kevin, and Kai Wang. 2017. Adaptive Sliding Window Segmentation for Physical Activity Recognition Using a Single Tri-axial Accelerometer. *Pervasive and Mobile Computing* 38 (2017), 41–59.

[30] Liangying Peng, Ling Chen, Zhenan Ye, and Yi Zhang. 2018. AROMA: A Deep Multi-Task Learning Based Simple and Complex Human Activity Recognition Method Using Wearable Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 2, 2 (2018), 1–16.

[31] Venkatesh Rajagopalan and Asok Ray. 2006. Symbolic Time series Analysis via Wavelet-based Partitioning. *Signal processing* 86, 11 (2006), 3309–3320.

[32] Attila Reiss and Didier Stricker. 2012. Introducing a New Benchmarked Dataset for Activity Monitoring. In *IEEE 16th International Symposium on Wearable Computers (ISWC)*. IEEE, 108–109.

[33] Attila Reiss, Markus Weber, and Didier Stricker. 2011. Exploring and Extending the Boundaries of Physical Activity Recognition. In *2011 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 46–50.

[34] Oliver Rösler and David Suendermann. 2013. A First Step Towards Eye State Prediction Using EEG. *Proc. of the AIHLS, International Conference on Applied Informatics for Health and Life Sciences* (2013).

[35] Amin Sadri, Yongli Ren, and Flora D Salim. 2017. Information Gain-based Metric for Recognizing Transitions in Human Activities. *Pervasive and Mobile Computing* 38 (2017), 92–109.

[36] Amin Sadri, Flora D Salim, Yongli Ren, Wei Shao, John C Krumm, and Cecilia Mascolo. 2018. What Will You Do for the Rest of the Day? an approach to continuous trajectory prediction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 2, 4 (2018), 1–26.

[37] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2019. Multi-task Self-Supervised Learning for Human Activity Detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 3, 2 (2019), 1–30.

[38] Rubén San-Segundo, Jaime Lorenzo-Trueba, Beatriz Martínez-González, and José M Pardo. 2016. Segmenting Human Activities Based on HMMs Using Smartphone Inertial Sensors. *Pervasive and Mobile Computing* 30 (2016), 84–96.

[39] Iqbal H Sarker, Alan Colman, Muhammad Ashad Kabir, and Jun Han. 2017. Individualized Time-series Segmentation for Mining Mobile Phone User Behavior. *Comput. J.* 61, 3 (2017), 349–368.

[40] Iqbal H Sarker and Flora D Salim. 2018. Mining User Behavioral Rules from Smartphone Data through Association Analysis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 450–461.

[41] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 400–408.

[42] Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul JM Havinga. 2016. Complex Human Activity Recognition Using Smartphone and Wrist-Worn Motion Sensors. *Sensors* 16, 4 (2016), 426.

[43] Hui Song, AK Qin, and Flora D Salim. 2018. Evolutionary Multi-objective Ensemble Learning for Multivariate Electricity Consumption Prediction. In *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–18.

[44] Hui Song, A Kai Qin, and Flora D Salim. 2016. Multivariate Electricity Consumption Prediction with Extreme Learning Machine. In *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2313–2320.

[45] Shaopeng Wang, Ye Yuan, and Hua Li. 2019. Discovering All-Chain Set in Streaming Time Series. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD*. 306–318. https://doi.org/10.1007/978-3-030-16148-4_24

[46] Yanwen Wang and Yuanqing Zheng. 2018. Modeling RFID Signal Reflection for Contact-free Activity Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 2, 4, Article 193 (2018), 22 pages. https://doi.org/10.1145/3287071

[47] Makoto Yamada, Akisato Kimura, Futoshi Naya, and Hiroshi Sawada. 2013. Change-point Detection with Feature Selection in High-dimensional Time-series Data. In *Proc. of 23th International Joint Conference on Artificial Intelligence (IJCAI)*.

[48] Kenji Yamanishi and Jun-ichi Takeuchi. 2002. A unifying framework for detecting outliers and change points from non-stationary time series data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 676–681. https://doi.org/10.1145/775047.775148

[49] Lina Yao, Quan Z Sheng, Wenjie Ruan, Xue Li, Sen Wang, and Zhi Yang. 2015. Unobtrusive Posture Recognition via Online Learning of Multi-dimensional RFID Received Signal Strength. In *IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS)*. 116–123.

[50] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. 2016. Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View that Includes Motifs, discords and shapelets. In *IEEE 16th international conference on data mining (ICDM)*. IEEE, 1317–1322.

[51] Masoomeh Zameni, Amin Sadri, Zahra Ghafoori, Masud Moshtaghi, Flora D. Salim, Christopher Leckie, and Kotagiri Ramamohanarao. 2019. Unsupervised Online Change Point Detection in High-Dimensional Time Series. *Knowledge and Information Systems (KAIS)* (2019), 719–750.

[52] Mi Zhang and Alexander A Sawchuk. 2012. USC-HAD: a Daily Activity Dataset for Ubiquitous Activity Recognition Using Wearable Sensors. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*. 1036–1043.

[53] Yan Zhu, Makoto Imamura, Daniel Nikovski, and Eamonn Keogh. 2017. Matrix Profile VII: Time Series Chains: A New Primitive for Time Series Data Mining (best student paper award). In *IEEE the 17th International Conference on Data Mining (ICDM)*. IEEE, 695–704.