Received March 24, 2020, accepted April 9, 2020, date of publication April 21, 2020, date of current version May 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2989165

Visualization of Generic Utility of Sequential Patterns

TOMASZ WIKTORSKI[1], ALEKSANDRA KRÓLAK[2], KAROLINA ROSIŃSKA[2], PAWEL STRUMILLO[2], (Senior Member, IEEE), AND JERRY CHUN-WEI LIN[3], (Senior Member, IEEE)

[1]Department of Electrical Engineering and Computer Science, University of Stavanger, 4036 Stavanger, Norway
[2]Institute of Electronics, Lodz University of Technology, 93-005 Lodz, Poland
[3]Department of Computer Science, Electrical Engineering and Mathematical Science, Western Norway University of Applied Sciences, 5063 Bergen, Norway

Corresponding author: Tomasz Wiktorski (tomasz.wiktorski@uis.no)

This work was supported in part by the University and College Network for Western Norway.

ABSTRACT Most of the literature on utility pattern mining (UPM) assumes that the particular patterns' utility in known in advance. Concurrently, in frequent pattern mining (FPM) it is assumed that all patterns take the same value. In reality, the information about the utility of patterns is not or hardly available in most cases. Moreover, the utility and frequency of the particular pattern are not directly proportional. An algorithm for estimating a generic pattern utility has been recently proposed, but the numeric results might be difficult to interpret. In particular, in datasets with many independent instances or groups. In this paper, we present an approach to generating utility bitmaps that provide visual representation of the numeric data obtained using generic pattern utility algorithm. We demonstrate validity of this approach on two datasets: *PAMAP2 Physical Activity Monitoring Data Set*, an open dataset from the UCI Machine Learning Repository, and an ECG dataset collected using Biopac Student Lab during Ruffier's test. For PAMAP2 dataset, utility bitmaps allow for immediate separation of various physical activities. Variation between participants are present, but do not overshadow differences between the activity types. For the ECG dataset, utility bitmaps immediately indicate age and fitness differences between the participants, even thought this information was not available to the algorithm. In both cases, partial similarity in bitmaps can be traced back to partial similarity in activities or participants generating the data. Based on these tests, the approach seems to be promising for exploratory analysis of large collections of long time series and possibly other sequential patterns such as distance series common in sports data analysis and depth series common in petroleum engineering.

INDEX TERMS Data visualization, DTW, exploratory data analysis, intelligent icons, SAX, sequential pattern, utility, utility visualization.

I. INTRODUCTION

Most of the literature on utility patterns (UPM) assumes a priori knowledge about the utility of given patterns. Research works concentrate on the development of efficient algorithms for detection of patterns with the highest utility [1]. On the other hand not much attention is related to cases where there is no a priori knowledge about the utility of the examined dataset but the higher utility can be assumed for some particular (sequential) patterns. It can be considered as a frequent pattern mining problem (FMP) but in this approach all types of patterns take the same value [2]. In many cases, the patterns of interest may appear not frequently enough in the examined sequence or may be infrequent w.r.t. the set of sequences so that in analysis based on the existing methods they would be omitted.

An algorithm for estimating a generic pattern utility has been proposed in [15], but the numeric results might be difficult to interpret. In particular, in datasets with many independent instances or groups. In this paper, we present an approach to generating utility bitmaps that provide visual representation of the numeric data obtained using generic pattern utility algorithm.

Time-series bitmaps, also called intelligent icons, have been proposed by Kumar *et al.* in [16] and Keogh *et al.* in [17]. Such bitmaps provide an overview of general distribution of

The associate editor coordinating the review of this manuscript and approving it for publication was Philippe Fournier-Viger.

values and their sequence in the dataset in a relatively fix amount of space. The size of a bitmap or an icon does not depend on the length of time series, rather bigger bitmaps provide greater level of detail. These methods are inspired by fractal drawing algorithm Chaos Game [18]. The detailed approach they use depends on reducing value span to only four possible values. They proved to be useful for many sequential datasets, such as DNA, ECG, and other.

However, the limited value span and focus on representing the whole dataset might be a limitation in many applications. Sometimes small changes are of critical importance to show difference between instances, but might be too small to change value distribution statistics for the whole datasets. One example is a difference in heart rate between activities such as ascending stairs and Nordic walking in PAMAP2 dataset. In earlier work, we demonstrated that it is possible to identify a set of short heart rate subsequences that separate such very similar (in terms of heart rate) activities.

While these subsequences show high utility they are simply a set of numbers and might be difficult to interpret at first. Especially, in presence of large number of instances or groups in the dataset. In this paper, we present an approach to converting these numeric results to easy to interpret bitmaps, that we call utility bitmaps. In contrast to earlier time-series bitmaps and intelligent icons, utility bitmaps are based only on representing the most useful subsequences and can use any value span.

We provide a proof of the effectiveness of our approach using *PAMAP2 Physical Activity Monitoring Data Set* [3], [4], an open dataset from the UCI Machine Learning Repository [13] and an ECG dataset collected using Biopac Student Lab during Ruffier's test.

The reminder of the paper is organized as follows. In Section II, we provide a short overview of related work. In Section III, we explain the core problem in details and follow with specification of a solution in Section IV. We validate the solution with PAMAP2 dataset in Section V and with ECG dataset in Section VI. We conclude and outline future work in Section VII.

II. RELATED WORK

A. TIME SERIES DATA REPRESENTATION

Wang *et al.* in [19] present an experimental comparison of representation methods for time series data. Representation methods can be divided into data adaptive and non-data adaptive. Data adaptive methods include: piecewise polynomials, Adaptive Piecewise Constant Approximation (APCA), Singular Value Decomposition (SVD), symbolic, and trees. Symbolic methods include commonly used SAX [11] and its derivatives iSAX [21], 1D-SAX [12].

Non-data adaptive representation methods include: wavelets, random mappings, spectral, and Piecewise Aggregate Approximation (PAA). Spectral methods include commonly used Discrete Fourier Transformation (DFT).

Authors report that methods such as SVD and PCA are not feasible for large datasets. On the other hand, there are multiple projects that use DFT, DWT, PAA, and SAX for large datasets.

All representation methods are compared using Tightness of Lower Band (TLB), which is a ratio of lower bound distance to true Euclidean distance. TLB has a value between 0 and 1. The lower the value the bigger reduction in number of disk accesses can be expected, while preserving quality of the representation.

Based on the experiments performed by authors of [19] on 80 different datasets from UCR Time Series Repository [20], the differences on average were rather small. However, periodic datasets favor spectral representations, such as DFT, while bursty data favor APCA.

Due to small differences in performance, representation is often selected based on specific needs of a particular use-case. Sometimes, certain methods tend to be more popular in some applications areas. DFT and wavelets are common in signal processing, while SAX and other symbolic methods are common in generic and long time series analysis. To fit a particular use-case authors often create custom extension to one of the main methods. A good example is ESAX [22], which adapts SAX for financial applications. It allows for better representation of minima and maxima that would otherwise be omitted in the regular SAX.

B. UTILITY OF SEQUENTIAL PATTERNS

In recent years the topic of mining for patterns of high importance has been widely explored. Patterns can be defined as frequent subsequences (sequential patterns), ordered set of elements or events, or can be characterized by more or less complicated set of association rules in case of more complex patterns. One of the properties of interest is a utility. Utility can be considered as a subjective measure based on expert knowledge or user preferences. It may be defined as a performance metric or measure how much given pattern contributes to a predefined objective function. Existing utility pattern mining algorithms can be divided into four categories: (1) apriori-based, (2) tree-based, (3) projection-based, and (4) vertical-/horizontal-data-based. Overview and in-depth analysis of existing algorithms can be found in [1].

Frequent pattern mining algorithms, such as Eclat, TreeProjections, and FP-growth were discussed in a survey by Aggarwal *et al.* in [2]. These algorithms have one main limitation, namely they use only one measure of pattern's relevance, that is frequency. The occurrence frequency of particular pattern may not be enough for decision making and some other features or more information may be required. Chan *et al.* in [23] presented high-utility itemset mining (HUIM), where more factors important in mining the high-utility patterns were considered, such as the unit profit of the item or the quantity of items. The concept of HUIM based on external information was proposed by Yao *et al.* in [36] but this approach suffers the problem of ‘‘combinational explosion’’. Transaction-weighted utilization (TWU) model was defined by Liu *et al.* in [25] to hold the downward closure property for mining the required information. A number of extensions based on TWU model was developed, i.e. HUI-Miner [26], HUP-tree [27], skyline HUIM [28], high-utility occupancy pattern mining (HUOPM) [29], IHUP [33], or UP-growth+ [34].

The use of utility concept was extended to sequential pattern mining (SPM) [30], [31]. In the survey by Fournier-Viger *et al.* [5] developments in SPM are discussed. They consider high-utility sequential pattern mining as a popular extension to SPM when frequency of occurrence is not sufficient and utility metric is required. They also assume that there exists prior knowledge about the utility that is associated with particular elements of the dataset. Application of utility concepts and uncertain sequence data in mining the average-utility pattern was discussed by Lin *et al.* in [32]. They proposed considering the size of the itemset as a measure of utility.

Outlier detection is one of the real applications of the utility concept, what is presented in the survey of algorithms for detecting outliers in temporal data [7]. The authors include the discussion on outliers subseries, however, the measures used assume a comparison only with the remaining parts of the particular time series. A similar approach is taken by Keogh *et al.* [6] in their work on finding surprising patterns. However, in the existing research the situation, where the high-utility itemset does not differ significantly from its surrounding, has not been given much attention yet.

C. VISUALIZATION OF TIME SERIES AND SEQUENTIAL PATTERNS

Visualization of data makes people understand meaning of data much quicker than a textual description. Visualization techniques in [35] are classified into 7 categories: 3D/volumetric charts (3D brain maps, interactive geo-spatial maps), icons, maps, multidimensional charts (area charts, bar graphs, bipartite graphs, box plots, bubble charts, causal network visualizations and heatmaps, key performance indicators, line graphs, pie charts and scatter plots), tables, temporal/timeline graphs (simple time series graphs with or without color coding) and textual descriptions. It has been found in [36] that positional and colour visual encodings are recommended for detection tasks and time series visualisations. On the other hand, more effective for the task of comparison are techniques using area visual encodings. Another important aspect is coordinate system. In general Cartesian coordinate system for time series visualizations is more effective than Polar.

Visualization techniques for sequential patterns can be divided into five types: individual representations, flow diagrams, aggregated pattern visualizations, placement strategies, and episode visualizations [37]. Each of these types has its advantages and disadvantages. Individual representations are characterized by poor scalability. On the other hand using this technique the user is able to completely identify all the pattern elements. Better scalability is achieved using flow diagrams, this technique also provides relatively good support for comparison of the patterns. Most types of flow diagrams include also some kind of interestingness measures. The greatest advantage of the aggregated pattern visualization technique is the possibility to present many missing events that are not included in the pattern itself. This technique can be applied for visualization of maximal, closed or generator patterns as it shows the amount of information that is lost in the process of compression. The pattern placement strategies are presenting the abstract visualizations of the patterns, what provides very good scalability both in terms of alphabet size and the number of patterns. Finally, the episode visualizations enable the identification of periodic occurrence-patterns and show the occurrences of patterns present in an event sequence.

III. BACKGROUND

This work focuses on visualization of sequential patterns. Time series are the most common example of such patterns. They can represent virtual variables (e.g. CPU load, network traffic) or real variables (temperature changes, HR variations). Typical sequential patterns are a variation of a single variable with time. However, more complex sequential patterns also exist, e.g. purchase patterns, while more convoluted, have an underlying time component.

In order to present our visualization method, we repeat a basic definition of a time series, following the original work on generic utility of sequential patterns [15].

Time series S of a length T can be defined as a list of values ordered by time, Eq. 1. Usually time distance between consecutive values is uniform for the particular time series, in such a case we say that the time series has a constant sampling rate.

$$S = (s_1, \ldots, s_t, \ldots, s_T), \quad t \in \mathbb{N}, \ s_t \in \mathbb{R} \tag{1}$$

Given a time series S, a subsequence S_{sub} of S of length L, Eq. 2, is a series consisting of L contiguous elements from S.

$$S_{sub} = (s_l, \ldots, s_{l+L-1}), \quad l \in \mathbb{N} \, and \\ 1 \leqslant l \leqslant T - L + 1, \ s_l \in \mathbb{R} \tag{2}$$

A distance measure D, presented in Eq. 3, between two time series S_1 and S_2 is a function that calculates the similarity between the given time series. Depending on the particular distance measure given, time series might have to be of the same length.

$$D(S_1, S_2) \rightarrow [0, \infty) \tag{3}$$

A subsequence distance measure between time series S and a shorter time series Q can be defined as the minimum distance between the Q and any subsequence of S, Eq. 4.

$$D_{sub}(S, Q) = min(D(S_{sub}, Q)) \tag{4}$$

All data used further in the paper were z-normalized and SAX [14] representation was obtained. Following parameters were used, length reduction was 1 to 50, meaning 50 time points of original time series were corresponding to each

SAX symbol. Various alphabet sizes were tested, but no major difference in results was observed. Results presented further in the paper use alphabet of size 4. Mostly, due to the same alphabet size being used in the main related paper.

IV. METHOD SPECIFICATION

In this section, we first provide an overview of an approximate method to calculating generic utility of sequential patterns, including some improvements we introduced with this work. This method is described in detail and evaluated in [15]. Next, we present an approach to generating a simple, yet informative, visualization of such patterns in form of utility bitmaps.

A. GENERIC UTILITY OF SEQUENTIAL PATTERNS

The generic utility of sequential patterns was proposed in [15] as a utility measure for datasets that lack user-specified utility values. The approach is on a surface similar to term frequency, inverse document frequency method. However, the frequency is replaced with a separation calculated using subsequence Dynamic Time Warping (DTW) similarity. DTW is a widely used similarity measure for time series and sequence data in general.

Since longer sequences might be naturally preferred as more unique, the method compensates for that using a length adjustment factor integrated in the utility formula. The factor integrates both the typical length of a sequence in the dataset and the length of the particular tested subsequence.

We define the utility of the subsequence U_{sseq} in Eq. 5, as a normalized difference of subsequence DTW distance in the sample s and out in the population P, specified in Eq. 6. In the original work the difference was calculated based on the distance in the group and out of the group. Such approach assumes that group information is known for at least part of the dataset. In this work this limitation was lifted and further results show that the method can produce expected results without group information, at least for the purpose of the visualization. In fact, as demonstrated later, it can reveal existing groups. The difference in DTW distance is further adjusted by length penalizing factor. Important to note is that normalization is with a range of $[-1, 1]$.

A high positive value is an indication that the tested subsequence is a good discriminator between a particular sequence and the remaining sequences. It can be a good candidate to use in learning algorithms or for visualization. Values around zero imply that the tested subsequence is equally common for the particular sequence and the remaining sequences. A high negative values, on the other hand, means that the tested subsequence is better represented in the remaining sequences. It can be used as a form of negative feedback to learning algorithms.

$$U_{sseq} = \widehat{(D_P - D_s)} * L_{adj} \tag{5}$$

$$D_P = S{-}DTW(sseq, P) \tag{6a}$$

$$D_s = S{-}DTW(sseq, s) \tag{6b}$$

To adjust for natural preference for longer subsequences adjustment term L_{adj} is introduced in Eq. 7. It is particularly important for scenarios when each sequence is considered against the remaining corpus of sequences, rather than group of sequences against another group. In the base the adjustment terms we have length of the subsequence L, in the exponent we have penalizing factor $-pf$ is the exponent.

Such a function ensures that for a subsequence of length equal 1 the value is 1 and the longer the subsequence gets the closer the L_{adj} gets to 0. Factor pf can adjust how fast that happens. In the earlier work that parameter was defined to be a logarithm of a length of an average sequence in the corpus, we follow the earlier definition here.

$$L_{adj} = L^{-pf} \tag{7a}$$

$$pf = log(L_{avg}) \tag{7b}$$

B. UTILITY BITMAPS DEFINITION

Execution of the generic utility algorithm provides a list of subsequences with highest utility. In a standard version all subsequences are of the same length X, the amount of the subsequences Y can also be defined by the user. Each element x of each subsequence y has a discrete value in a range Z. Such a list of subsequences $m_{y,x}$ can be represented as a row-major matrix $M_{y,x}$, as in Eq. 8.

$$M_{y,x} = \begin{bmatrix} m_{1,1} & m_{1,\dots} & m_{1,X} \\ m_{\dots,1} & m_{\dots,\dots} & m_{\dots,X} \\ m_{Y,1} & m_{y,\dots} & m_{Y,X} \end{bmatrix} \tag{8}$$

Eq. 9 presents an example utility list EUL of length $Y = 3$ seen as the amount of rows in the matrix, and subsequences of length $X = 4$ seen as the amount of elements in each row in the matrix, and value range $Z = 5$.

$$EUL = \begin{bmatrix} 5 & 3 & 1 & 1 \\ 1 & 3 & 3 & 1 \\ 5 & 3 & 3 & 1 \end{bmatrix} \tag{9}$$

Both the generic utility numbers and the relatively highly processed values of the subsequences can be difficult to interpret. In particular, in datasets with many independent instances or groups. However, converting such numeric data to colors or shades of gray can have positive effects on interpretability.

We define a color mapping function cm in Eq. 10. The function maps from a domain of values V to a domain of colors C, defined as standard additive color model consisting of red R, green G, blue B components. Such color mapping functions are available in most of data analytic tools.

$$cm : V \mapsto C \tag{10a}$$

$$\text{where } C = (R, G, B) \tag{10b}$$

We apply a typical grayscale color mapping function to our example utility list, it results in a new matrix presented in Eq. 11. Value of *245* is the selected maximum and value of *0* is the minimum. Since the color model is additive, the potential maximum values of *255* for each *RGB* component would result in color white. The minimum values of *0* for each *RGB* component result in color black.

In our experiments we noticed that some users misinterpreted color white as a lack of information instead of low or high value. Therefore, we recommend limiting color spectrum in a way that pure white is avoided. This is why selected maximum value is *245* and medium value becomes *123*. Sometimes it might also be helpful to reverse the colors, that is for dark shades to represent high values and respectively light shade to represent low values. It can be achieved by a simple reversed scaling. Such choice is made in Sections V and VI, since it was preferred by users during testing. At the same time, we do not recommend such reversed scaling if full color mapping is used. This is the reason why all the formulas and the algorithm use regular scaling in (11), as shown at the bottom of this page.

Finally, we use list dimensions as *x,y* coordinates and convert *RGB* values to actual shades of gray. Results are visible in Figure 1. Pure white color was adjusted to a slight shade of gray, as recommended.

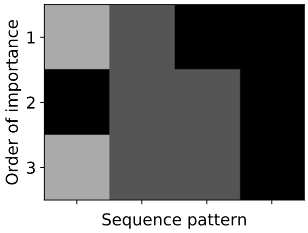

FIGURE 1. Example utility bitmap.

Along the *y* axis we see 3 different subseries, where importance decreases from top to bottom, as marked by the ticks. Along the *x* axis we see pattern progression along each of the subseries. In this case, darker tones represent lower values and lighter tones higher values. This can be adjusted with a different color map.

Another element that can be adjusted is how scaling is performed. It is possible to perform scaling for multiple instances or groups. Data can be scaled for the whole dataset or per individual or group. We found all strategies useful when testing the approach for various datasets.

C. UTILITY BITMAPS ALGORITHM

After presenting a definition of utility bitmaps in the previous subsection, we present a basic Algorithm 1 that transforms a list of subsequences with highest utility to utility bitmaps.

The input list of sequences *TS* is identical to the output list of the algorithms from [15]. The length of sequences *X* can be either set by the user or optimized by the algorithm. The amount of sequences *Y* is in principle selected manually, but a simple threshold rule can also be established. The output of the algorithm is a matrix *BM* with *Y* rows and *X* columns.

In lines 1-3 we define a simple grayscale color mapping function *gray_cm*. It takes 5 arguments. First argument *v* is the value to be scaled. Next two arguments *min_v* and *max_v* are extrema of values present in the dataset to be scaled. In our case it will be *TS* and these extrema are obtained in lines 5-6. The last two arguments are the extrema of the output data we want to achieve. In our case they correspond to extrema of *BM*, we select them to be *0* and *245*, what is visible in line 9. *255* is the maximum range of a byte value, which is a typical value for specifying maximum intensity of a color component. However, we reduce this value by 10 to avoid pure white color in the utility bitmaps. Such white color was, in our experiments, misinterpreted by the users as lack of information, instead of a maximum value. We recommend such limitation to the colormap range for any colormap that uses white for highest scalar values.

In lines 7 and 8 we iterate first over each high-utility subsequence and then over each element of the subsequence. For each element in line 9 we obtained a color representation of the element.

The resulting matrix *BM* can be best represented as matrix of squares filled by a color defined by the respective *(r,g,b)* values from the *BM* matrix organized in a Cartesian coordinate system. We recommend that *x* axis should be oriented, as usual, to the right, but *y* axis should be oriented downwards, which is the opposite of a typical orientation. The reason is to place the most important subsequence to the top of the bitmap.

V. VALIDATION WITH PAMAP2 DATASET

In this section, we apply the outlined approach to heart rate data from PAMAP2 dataset collected during three different physical activities for a number of subjects.

A. DATASET DESCRIPTION

The effectiveness of the proposed method was tested on two use cases from the *PAMAP2* dataset. For these use cases two pairs of heart rate (HR) signals were selected: (1) signals recorded during lying and ascending the stairs, and (2) signals recorded during ascending the stairs and during Nordic walking.

The first use case is supposed to provide an initial proof that the proposed method can provide a clear separation between the activities in the generated visualisations. For this reason selected HR signals significantly differ from each other. Heart rate patterns for the first use case are presented in Figure 2. The difference between the two traces can be easily observed, what should also be presented on the visualization produced

$$EUL_{cm} = \begin{bmatrix} (245, 245, 245) & (123, 123, 123) & (0, 0, 0) & (0, 0, 0) \\ (0, 0, 0) & (123, 123, 123) & (123, 123, 123) & (0, 0, 0) \\ (245, 245, 245) & (123, 123, 123) & (123, 123, 123) & (0, 0, 0) \end{bmatrix} \tag{11}$$

Algorithm 1 Outline of the Proposed Algorithm

```
Data: $TS := [T_{sub_1}, \dots, T_{sub_y}, \dots, T_{sub_Y}] \equiv [[v_{1,1}, v_{1,x}, v_{1,X}], [v_{y,1}, v_{y,x}, v_{y,X}] [v_{Y,1}, v_{Y,x}, v_{Y,X}]]$ a list of $Y$
      subsequences with highest utility, each of length $X$
Result: $BM := [[(r,g,b)_{1,1}, (r,g,b)_{1,x}, (r,g,b)_{1,X}], [(r,g,b)_{y,1}, (r,g,b)_{y,x}, (r,g,b)_{y,X}] [(r,g,b)_{Y,1}, (r,g,b)_{Y,x}, (r,g,b)_{Y,X}]]$
        a matrix of $Y$ rows and $X$ columns, each element consisting of 3 (red,
        green, and blue) scalar components of an additive color model

Function gray_cm(v, min_v, max_v, lowest, highest):
    scale = (highest − lowest) / (max_v − min_v);
    tval := scale*v + lowest − min_v*scale;
    return (tval, tval, tval);

$min_{TS}$ := min(TS);
$max_{TS}$ := max(TS);
for $y := 1$ to $Y$ do
    for $x := 1$ to $X$ do
        $BM_{y,x}$ := gray_cm($TS_{y,x}$, $min_{TS}$, $max_{TS}$, 0, 245);
    end
end
```

by the proposed algorithm. However, we can see on the plot that there are two parts of the HR traces close to each other: in the first 5 instances, and later in time instants between 15 and 18.

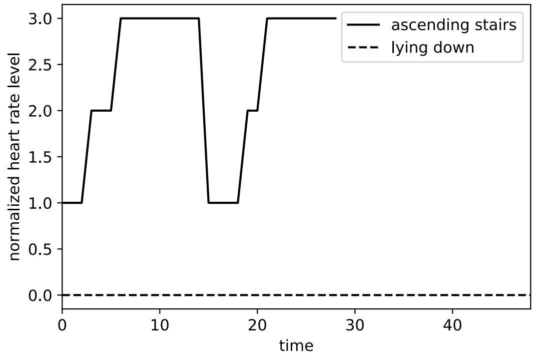

FIGURE 2. Heart rate trace during lying and ascending stairs.

The second use case is potentially more difficult as selected HR signals are fairly similar. Small variations result from differences in the exercise protocol followed by the test subjects. The heart rate patterns during Nordic walking and ascending the stairs are presented in Figure 3. Some differences are visible on the plots, but they are not as significant as in case of lying and ascending the stairs. The parts of the HR traces for instances from 15 to 18 might at first seem very different between the two activities. However, keeping in mind that in the proposed method the direct time correspondence is not considered, this part of the trace for ascending the stairs can be matched partially to the initial rise. Nevertheless, the proposed algorithm should also produce visualisation providing separation between the two examined activities, though it might not be as clear as in the first example.

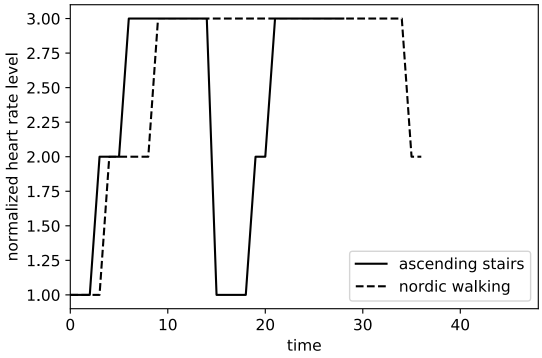

FIGURE 3. Heart rate trace during Nordic walking and ascending stairs.

Presented figures with HR trace show heart rate patterns for one selected participant. The resulting utility bitmaps include all the participants. Results produced by the proposed method are based only on the heart rate signal. More clear separation might be achieved with accelerometer data included.

The original method for finding most useful subsequences [15] does not assume that there is a direct time correspondence between two compared time series. Such an assumption would be simply too limiting. First of all, real-life data would seldom be perfectly aligned. Even in cases where alignment data could be available, it would require a significant effort to make use of it, except for a small subset of scenarios. Moreover, it would create issues with different length of datasets. Either large portion of the data would have to be discarded or multiple convolutions would have to be considered adding to computational complexity. As a result, we calculate the utility of a subsequence with respect to any potential positioning in the dataset.

B. RESULTS AND DISCUSSION

Figure 4 presents utility bitmaps generated applying outlined approach (first finding a set of most useful subsequences, then visualizing them in form of a bitmap) to an HR trace of 3 different subjects performing Nordic walking.

There is a clear similarity between between bitmap for each subject. Top 3 subsequences for Subject 1 and Subject 2 are identical and 4th subsequence for Subject 1 is the same as 5th subsequences for 2. Subject 3 shares the top subsequence with other 2 subjects, 2nd most important subsequence is the same as 4th most important for Subject 2, and 3rd most important subsequence is the same as 5th most important for Subject 1.

Figure 5 presents utility bitmaps generated based on the HR trace of 3 different subjects lying down.

For Subject 1 and 2 the whole utility bitmaps are in fact identical. For subject 3 1st and 5th subsequence are also the same. There are some small differences in 2nd, 3rd, and 4th pattern when comparing Subject 3 with the other two.

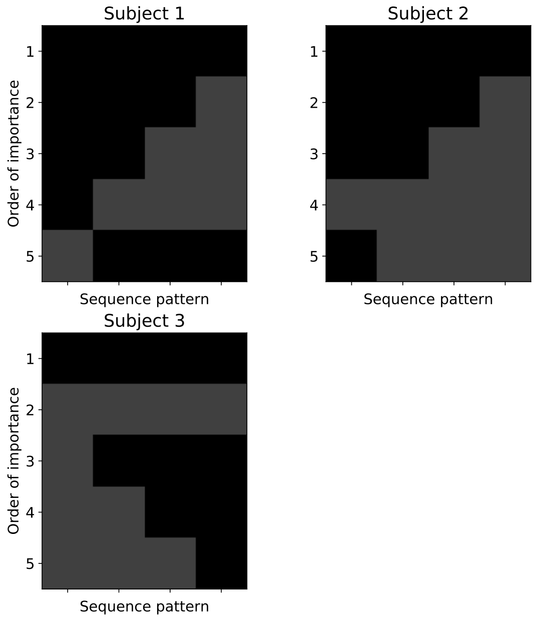

FIGURE 4. Utility bitmaps for Nordic walking.

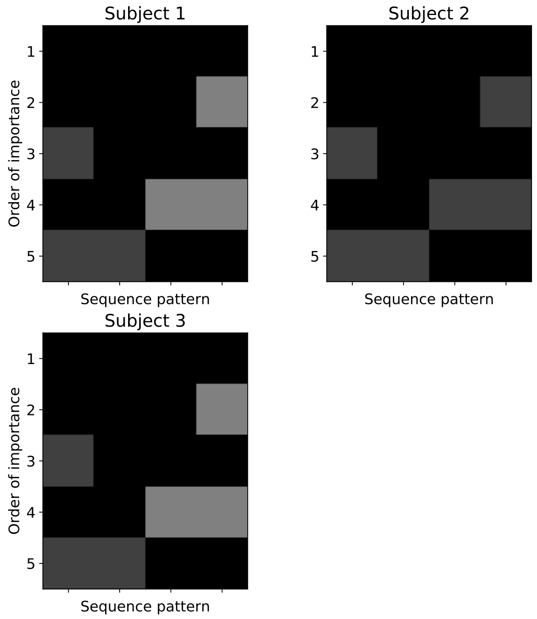

FIGURE 6. Utility bitmaps for ascending stairs.

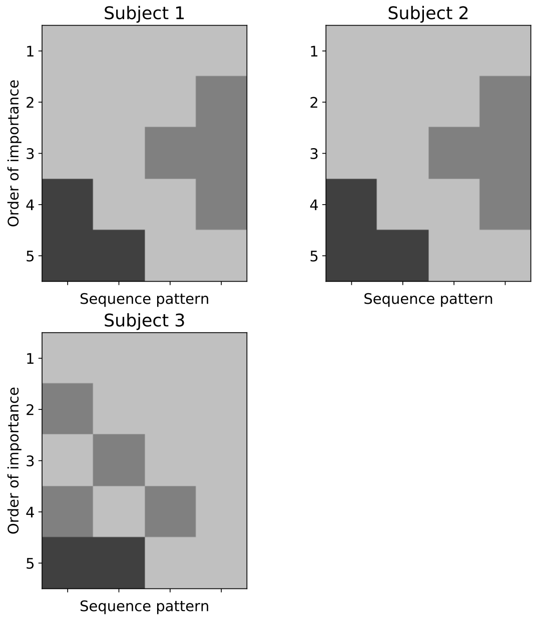

FIGURE 5. Utility bitmaps for lying down.

When comparing to utility bitmaps for Nordic walking the differences are clear and there is no doubt about grouping of activities despite some differences in bitmaps between subjects.

Figure 6 show utility bitmaps generated applying outlined approach (first finding a set of most useful subsequences, then visualizing them in form of a bitmap) to an hr trace of 3 different participants ascending stairs.

In this case bitmaps for part 1 and 3 are identical. For 2 we observe small differences on the right side of patterns 2 and 4, but they are really minor.

When comparing to utility bitmaps for lying the differences are as clear as in the earlier case and there is no doubt about grouping of activities despite some differences in bitmaps between participants.

In a more difficult scenario, when comparing to utility bitmaps for Nordic walking a clear separation is anyway visible. There is an overlap for some subsequences, in particular 1st subsequence, but the whole bitmap is noticeably different. Both on individual and group level.

VI. VALIDATION WITH ECG DATASET

In this section a set of ECG data collected using Biopac Student Lab during Ruffier's test are analyzed. Unlike data discussed in Section V here the signals were collected from persons performing the same physical exercise, however the subjects are characterized by different level of cardiovascular fitness.

A. DATASET DESCRIPTION

The aim of the experiment performed at LUT was to determine heart rate changes in time on the basis of the recorded signals (ECG) during the cardiac stress test. Data were acquired from 5 male volunteers. Each participant showed different tolerance to physical effort due to the type of work performed. There were no contraindications for carrying out performance/cardiac stress tests as none of the examined persons had diagnosed cardiovascular diseases (and in particular problems with heart activity). Before conducting the tests, they were informed about the course and purpose of the experiment, and signed the permission for analysis of their registered signals.

Before the study, each person was assigned the appropriate level of a 4-grade scale that determined physical condition based on the self-assessment, profession and performed interview (1 - poor, 2-average, 3-good, 4-very good).

- Person 1: age 47, fireman, physical condition: 3;
- Person 2: age 22, physical/construction worker, physical condition: 4;
- Person 3: age 19, student physically active before starting studies, physical condition: 2
- Person 4: age 23, office worker, hobby: playing handball, physical condition: 2
- Person 5: age 30, computer scientist, physical condition: 1

Performed experiment was based on the Ruffier's test. The subjects performed 30 sit-ups during one minute. Each of the respondents was asked to wear appropriate clothing that did not limit movements during exercise. The time of each participants' ECG signal registration lasted 210 seconds and was divided into 3 stages:

1) Resting in a sitting position for 60 seconds (heart rate measurement at the beginning of time measurement).
2) Effort (30 sit-ups) in 60 seconds.
3) Rest in a sitting position for 90 seconds (heart rate measurement immediately after exercise and after 1 minute of rest).

Additionally pulse measurement was performed by examined person (they declared that they could measure their heart rate).

The Biopac Student Lab (BSL) system was used for measuring the pulse in specified time periods and for continuous ECG signal recording. To monitor the heart activity, three bipolar leads were used according to the Einthoven triangle, in which the electrodes are placed on the limbs (2 on the upper limbs and 1 on the lower left limb). This arrangement of electrodes makes it possible to obtain a high-quality signal at rest. However, during effort, when both the lower and upper limbs move, such arrangement of the electrodes would affect the signal quality due to the large amount of interference in the recorded signal. For this reason, the electrodes from the upper limbs were moved subclavian cavities, while the electrode from the lower limb was moved below the navel line. This allowed to limit the movement of the electrodes, and thus reduce the amount of noise in the collected signal.

During the conducted tests, electrical activity of the heart (ECG) was recorded using Biopac Student Lab system. Analog signals were converted to digital ones with a sampling frequency of 200 Hz. The signal was amplified 1000 times and preprocessed by using a low-pass filter with a cut-off frequency of 150 Hz. The recorded data was saved in the '.txt' file. Each of the recorded signals has been analyzed according to the following algorithm:

1) Removal of cardiac isoelectric line flow by using median filter.
2) Determination of local minima and maxima (initial detection of QRS complexes).
3) Limit the distance between detected maxima and minima to less than 10 samples.
4) Limit the maximum and minimum amplitude (threshold value selected directly for the registered signal).
5) Determination of RR intervals
6) Determination of HR changes in time according to the equation HR = 60 / RR.
7) Median filtration of obtained signal.

Obtained waveforms are presented in Figure 7. For person 1 it can be seen that after the start of exercise, there is a rapid increase in heart rate, but it does not reach a very high value. It means that the person has rather good cardiovascular fitness. Even during exercise, the frequency of heart beats gradually decreases, but during the minute of rest it does not reach the initial value. For person 2, heart rate during exercise increases with lower intensity than for person 1, however, it reaches much higher values. After exercise, the maximum value is maintained for about 5 seconds. After a minute of rest the heart rate does not return to the initial value. The interview conducted before the start of the study suggested that this person has very good physical fitness. The results of the study do confirm that. This person, being a construction worker, regularly performs strength training, but it does not improve cardiovascular fitness as opposed to aerobic training. For person 3, a sharp increase in heart rate can be observed after the start of exercise, after the end of exercise, for another 7 seconds, the heart rate continues to increase. Next we observe a short, sharp drop of HR followed by the slow decrease of the heart rate. In the case of person 4, during the effort, the heart rate increases to about 120bpm and after the exercise it falls quite rapidly. For person 5, it can be seen that after the start of the exercise, the heart rate increases to over 160 beats per minute, and after the end of the exercise, the value decreases, however, it does not return to the value at the beginning of the test.

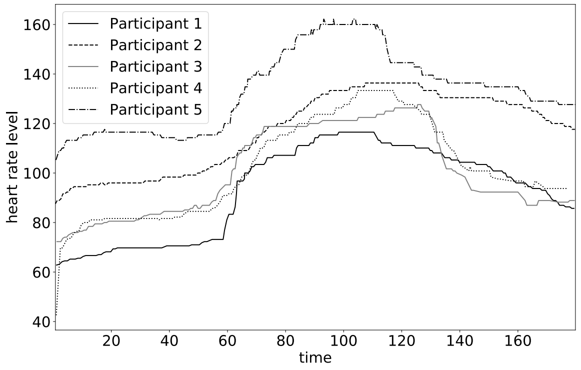

FIGURE 7. Heart rate trace during the experiment.

B. RESULTS AND DISCUSSION

Utility bitmaps were generated for the examined persons in the period of resting after the sit-ups session. We can observe clear patters for 3 groups, what was not visible in the plots in time domain. Results suggest that persons 2 and 5 are in poor physical condition, while persons 1, 3 and 4 are in good physical condition, however person 1 was significantly older than the rest of the participants what resulted in his qualifying for a separate group.

In Figure 8 we can see utility bitmaps for persons 3 and 4. The most significant elements show high values (light color of elements) in both cases. It represents high HR right after the exercise period. Next rows, in the order of importance, show rather rapid transitions between the high and low HR values. This is characteristic for persons with good overall physical and cardiovascular fitness.

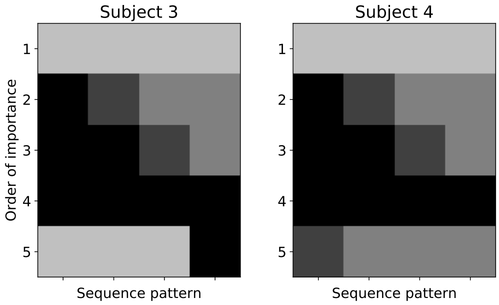

FIGURE 8. **Utility bitmaps for participants three and four.**

Figure 9 presents utility bitmaps for persons 2 and 5. They are exactly the same, showing the pattern of relatively low HR values being the most significant. It suggests a trend of small changes of HR in the resting period after exercise session. Other segments, apart from the elements in the second row, also show minor changes what is a sign of slow process of returning to the low heart rate values. Such patter is characteristic for persons with poor cardiovascular fitness.

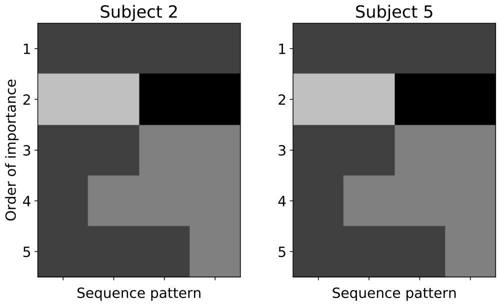

FIGURE 9. **Utility bitmaps for participants two and five.**

Figure 10 shows the results for person 1. The most significant components in the signal are those with very low values as this person does not achieve high HR values throughout the whole Ruffier's test. In general there are no high values present in this utility bitmap. It implies that the person is in good physical shape and is characterized by high efficiency of the circulatory system. Lower HR values are characteristic for older persons, therefore we can see that this utility bitmap, generated for person aged 47, is significantly different than bitmaps obtained for fit persons in their early 20s.

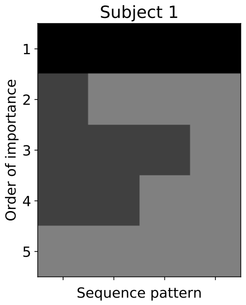

FIGURE 10. **Utility bitmap for participant one.**

VII. CONCLUSION AND FUTURE WORK

In this paper, we proposed an approach to visualization of generic utility of sequential patterns. Based on algorithm for finding generic utility of sequential patterns and an approach inspired by time series bitmaps.

We provided a proof of the effectiveness of our approach using *PAMAP2 Physical Activity Monitoring Data Set* [3], [4], an open dataset from the UCI Machine Learning Repository and an ECG dataset collected using Biopac Student Lab during Ruffier's test.

For PAMAP2 dataset, utility bitmaps allow for immediate separation of various physical activities. Variation between participants are present, but do not overshadow differences between the activity types.

For the ECG dataset, utility bitmaps immediately indicate age and fitness differences between the participants, even thought this information was not available to the algorithm. In both cases, partial similarity in bitmaps can be traced back to partial similarity in activities or participants generating the data.

Based on these tests the approach seems to be promising for exploratory analysis of large collections of long time series and possibly other sequential patterns such as distance series common in sports data analysis and depth series common in petroleum engineering.

Possible application of utility bitmaps is EEG signal analysis for detection of various brain reactions to particular stimuli among large groups of patients. People react in different way to visual, auditory, olfactory or other sensory stimuli, depending on their personal characteristics, experience or associations. Changes of power spectral density of particular brainwaves in time are often hard to detect and interpret when presented in a form of waveforms in time. Utility bitmaps can be helpful in detection of specific patterns and dependencies between the type of stimuli and patients' characteristics.

The proposed method can be extended to analysis of video sequences, in everyday life applications, including object detection and tracking, or in medical imaging involving long sequences, such as videoendoscopy. One of the possible applications is videoplethysmography (VPG), where, depending on the personal characteristics of the patient, crucial information can be included in different color spaces, channels or combinations of channels.

REFERENCES

[1] W. Gan, J. Chun-Wei Lin, P. Fournier-Viger, H.-C. Chao, V. S. Tseng, and P. S. Yu, "A survey of utility-oriented pattern mining," 2018, *arXiv:1805.10511*. [Online]. Available: http://arxiv.org/abs/1805.10511

[2] C. C. Aggarwal, M. A. Bhuiyan, and M. A. Hasan, "Frequent pattern mining algorithms: A survey," in *Frequent Pattern Mining*, C. C. Aggarwal J. Han, Eds. Cham, Switzerland: Springer, 2014, pp. 19–64.

[3] A. Reiss and D. Stricker, "Creating and benchmarking a new dataset for physical activity monitoring," in *Proc. 5th Int. Conf. Pervas. Technol. Rel. Assistive Environ. (PETRA)*, New York, NY, USA, 2012, p. 40.

[4] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *Proc. 16th Int. Symp. Wearable Comput.*, Jun. 2012, pp. 108–109.

[5] P. Fournier-Viger, J. C.-W. Lin, R. U. Kiran, Y. S. Koh, and R. Thomas, "A survey of sequential pattern mining," *Data Sci. Pattern Recognit.*, vol. 1, no. 1, pp. 54–77, 2017.

[6] E. Keogh, S. Lonardi, B. Y. Chiu, "Finding surprising patterns in a time series database in linear time and space," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, 2002, pp. 550–556.

[7] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2250–2267, Sep. 2014.

[8] S. Seiler and J. E. Sjursen, "Effect of work duration on physiological and rating scale of perceived exertion responses during self-paced interval training," *Scandin. J. Med. Sci. Sports*, vol. 14, no. 5, pp. 318–325, Oct. 2004.

[9] H. Tanaka, K. D. Monahan, and D. R. Seals, "Age-predicted maximal heart rate revisited," *J. Amer. College Cardiol.*, vol. 37, no. 1, pp. 153–156, Jan. 2001.

[10] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intell. Data Anal.*, vol. 11, no. 5, pp. 561–580, Oct. 2007.

[11] S. Malinowski, T. Guyet, R. Quiniou, and R. Tavenard, "1d-SAX: A novel symbolic representation for time series," in *Advances in Intelligent Data Analysis XII*. 2013, pp. 273–284.

[12] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, "Searching and mining trillions of time series subsequences under dynamic time warping," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2012, pp. 262–270.

[13] D. Dua and C. Graff, "UCI machine learning repository," School Inf. Comput. Sci., Univ. California, Irvine, CA, USA, Tech. Rep., 2017. [Online]. Available: https://archive.ics.uci.edu/ml/citation_policy.html

[14] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: A novel symbolic representation of time series," *Data Mining Knowl. Discovery*, vol. 15, no. 2, pp. 107–144, Aug. 2007.

[15] T. Wiktorski and J. C.-W. Lin, "Approximate approach to finding generic utility of sequential patterns," in *Proc. Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2019, pp. 1029–1034.

[16] N. Kumar, V. N. Lolla, E. Keogh, S. Lonardi, C. A. Ratanamahatana, and L. Wei, "Time-series bitmaps: A practical visualization tool for working with large time series databases," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2005, pp. 531–535.

[17] E. Keogh, L. Wei, X. Xi, S. Lonardi, J. Shieh, and S. Sirowy, "Intelligent icons: Integrating lite-weight data mining and visualization into GUI operating systems," in *Proc. 6th Int. Conf. Data Mining (ICDM)*, Dec. 2006, pp. 912–916.

[18] M. F. Barnsley, *Fractals Everywhere*. New York, NY, USA: Academic, 2014.

[19] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh, "Experimental comparison of representation methods and distance measures for time series data," *Data Mining Knowl. Discovery*, vol. 26, no. 2, pp. 275–309, Mar. 2013.

[20] E. Keogh E. X. Xi, L. Wei, and C. Ratanamahatana. (2006). *The UCR Time Series Dataset*. [Online]. Available: http://www.cs.ucr.edu/~eamonn/time_series_data/

[21] A. Camerra, T. Palpanas, J. Shieh, and E. Keogh, "ISAX 2.0: indexing and mining one billion time series," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 58–67.

[22] B. Lkhagva, Y. Suzuki, and K. Kawagoe, "New time series data representation ESAX for financial applications," in *Proc. 22nd Int. Conf. Data Eng. Workshops (ICDEW)*, Apr. 2006, p. x115.

[23] R. Chan, Q. Yang, and Y.-D. Shen, "Mining high utility itemsets," in *Proc. 3rd IEEE Int. Conf. Data Mining*, Nov. 2003, pp. 19–26.

[24] H. Yao, H. J. Hamilton, and C. J. Butz, "A foundational approach to mining itemset utilities from databases," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2004, pp. 211–225.

[25] Y. Liu, W. K. Liao, and A. Choudhary, "A two-phase algorithm for fast discovery of high utility itemsets," in *Advances in Knowledge Discovery and Data Mining*. Berlin, Germany: Springer, 2005, pp. 689–695.

[26] M. Liu and J. Qu, "Mining high utility itemsets without candidate generation," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2012, pp. 55–64.

[27] C.-W. Lin, T.-P. Hong, and W.-H. Lu, "An effective tree structure for mining high utility itemsets," *Expert Syst. Appl.*, vol. 38, no. 6, pp. 7419–7424, Jun. 2011.

[28] J. C.-W. Lin, L. Yang, P. Fournier-Viger, and T.-P. Hong, "Mining of skyline patterns by considering both frequent and utility constraints," *Eng. Appl. Artif. Intell.*, vol. 77, pp. 229–238, Jan. 2019.

[29] W. Gan, J. C.-W. Lin, P. Fournier-Viger, H.-C. Chao, and P. S. Yu, "HUOPM: High-utility occupancy pattern mining," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1195–1208, Mar. 2020.

[30] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proc. 11th Int. Conf. Data Eng.*, Mar. 1995, pp. 3–14.

[31] W. Gan, J. C.-W. Lin, P. Fournier-Viger, H.-C. Chao, and S. P. Yu, "A survey of parallel sequential pattern mining," *ACM Trans. Knowl. Discov. Data*, vol. 13, no. 3, p. 25, Jun. 2019.

[32] J. C.-W. Lin, T. Li, M. Pirouz, J. Zhang, and P. Fournier-Viger, "High average-utility sequential pattern mining based on uncertain databases," *Knowl. Inf. Syst.*, vol. 62, no. 3, pp. 1199–1228, Mar. 2020.

[33] C. F. Ahmed, S. K. Tanbeer, B.-S. Jeong, and Y.-K. Lee, "Efficient tree structures for high utility pattern mining in incremental databases," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 12, pp. 1708–1721, Dec. 2009.

[34] V. S. Tseng, B.-E. Shie, C.-W. Wu, and P. S. Yu, "Efficient algorithms for mining high utility itemsets from transactional databases," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 8, pp. 1772–1786, Aug. 2013.

[35] M. Roham, A. R. Gabrielyan, and N. Archer, "A systematic review of knowledge visualization approaches using big data methodology for clinical decision support," in *Advances in Intelligent and Personalized Clinical Decision Support Systems*, 1st ed. Rijeka, Croatia: IntechOpen, 2019.

[36] M. Adnan, M. Just, and L. Baillie, "Investigating time series visualisations to improve the user experience," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2016, pp. 5444—-5455.

[37] W. Jentner and D. A. Keim, "Visualization and visual analytic techniques for patterns," *High-Utility Pattern Mining: Theory, Algorithms and Applications*. Cham, Switzerland: Springer, 2019.

TOMASZ WIKTORSKI received the Ph.D. degree from the University of Stavanger, Norway, in 2011, including a scholarship at Stanford University, USA, and the M.Sc. degree from the Lodz University of Technology, Poland, in 2007, including a scholarship at the University of Alicante, Spain.

In 2007, he was a Research Intern at Carnegie Mellon University, USA. He is currently an Associate Professor with the Department of Electrical Engineering and Computer Science, University of Stavanger, where he is also a Study Program Manager for Computer Science and Data Science. His work focuses on data science and data-intensive computing, including in particular machine learning and deep learning approaches to the analysis of time series. He was a Principal Investigator at UiS in the EC-funded EDISON project that developed common data science framework to support standardized education across European institutions. He lead work packages and deliverables in several EC and NFR-funded projects, largest having over 110PM.

ALEKSANDRA KRÓLAK received the Ph.D. degree in computer science from the Lodz University of Technology, Poland, in 2009.

Since 2009, she has been an Assistant Professor with the Institute of Electronics, LUT, Poland. Her research interests include processing and analysis of biomedical signals and images, development of human–computer interaction systems, and aspects of accessibility in ICT.

Dr. Królak's awards and honors include three gold medals for inventions: the Salon International des Inventions, Geneva, in 2010; the Korean International Women's Inventions Exposition, Seoul, in 2010; the IV International Warsaw Invention Show, in 2010; and distinction or development of eye-blink controlled human–computer interaction system from Polish Ministry of Science and Higher Education, in 2011.

KAROLINA ROSIŃSKA received the B.S. degrees in biomedical engineering and material engineering and the M.S. degree in Biomedical Engineering from the Lodz University of Technology, Poland, in 2017, 2018, and 2018, respectively, where she is currently pursuing the Ph.D. degree in material engineering.

Mrs. Rosińska awards and honors include 1st place in the XXXIII competition of the Łódź Federation Council SNT-NOT for the best master's thesis at the Lodz University of Technology, in the 2017/2018 academic year, the student of the Year 2015/2016 and 2016/2017, the Rector's scholarship for the best students for the academic year 2014/2015, 2015/2016, 2016/2017, 2017/2018, and the Ericpol scholarship as part of the Employers Scholarships Scholarship Program in Łódź.

PAWEL STRUMILLO (Senior Member, IEEE) is currently the Head of the Institute of Electronics and holds the position of a Full Professor with the Lodz University of Technology (TUL). From 1991 to 1993, he was with the University of Strathclyde (under the EU Copernicus programme) where he defended his Ph.D. thesis, in 1993, devoted to ECG signal processing. He has published more than 200 scientific articles and three books. He has supervised eight Ph.D. candidates in electronics, computer science, and biomedical engineering disciplines. In 2007, he was a member of the panel, Development scenarios of medical technologies in Poland, working in the framework of the National Foresight Programme, Poland, in 2020. In 2008, he co-developed new undergraduate course Biomedical Engineering taught in English. His current research interests include medical electronics, processing of biosignals, soft computing methods, and human–system interaction systems. In recent years, with his co-workers, he has built award winning assistive technologies for the disabled, e.g., an electronic travel aid for the blind awarded by the Polish Agency for Enterprise Development in 2012, commercialized with Orange Labs.

Prof. Strumillo was a member of the steering committee of the HORIZON 2020 project on aiding the visually impaired, from 2015 to 2017. He is a member of the Biocybernetics and Biomedical Engineering Committee of the Polish Academy of Sciences. He is also engaged in the COST Action (CA 18110), and serves on the Management Committee.

JERRY CHUN-WEI LIN (Senior Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, in 2010. He is currently a Full Professor with the Department of Computer Science, Electrical Engineering, and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway. He has published more than 300 research articles in refereed journals, such as the IEEE Transactions on Knowledge and Data Engineering (TKDE), the IEEE Transactions on Cybernetics (TCYB), ACM TKDD, and ACM TDS, and international conferences, such as the IEEE ICDE, IEEE ICDM, PKDD, and PAKDD. His research interests include data mining, soft computing, social computing, artificial intelligence and machine learning, and privacy-preserving and security technologies. He is the IET Fellow and a Senior Member of ACM. He is also the Project Co-Leader of well-known SPMF: An Open-Source Data Mining Library, which is a toolkit offering multiple types of data mining algorithms. He also serves as the Editor-in-Chief for the *International Journal of Data Science and Pattern Recognition*.