# Scalable Pattern Recognition in Financial Time Series Data with Double-Cycled Value Based Data Representation

Kwan-Hua Sim, Kwan-Yong Sim, and Valliappan Raman

***Abstract*—Time series pattern recognition in motif discovery has emerged as one of the prominent primitives for data series mining and knowledge discovery. In the context of mining for scalable financial time series patterns, the repeated price patterns are not only unknown in advance, but they are also highly scalable in size, which require a more responsive elastic distance measure than lock-step distance measure. Nonetheless, the state-of-the-art motif discovery techniques are still operating on motif pair of equal length with lock-step Euclidean distance metric. Computational complexity in time series search is the bottleneck that not only refrains brute-force search with different length, but also deprives the implementation of more complex elastic distance measure. Hence, this study introduced a novel double-cycled Value Based Data Representation approach with inherent time transformation element from the data representation process. It aims to produce elastic measure comparable result by only using Euclidean distance metric, bypassing the computational complexity presence in elastic distance measure.**

***Index Terms*—Data representation, motif discovery, time series pattern recognition, time series data mining.**

## I. Introduction

Data analytic based on similarity of time series data is a critical tool in exploring and mining of information hidden in the data series. The immense volumes of time series data available today offers tremendous potential for knowledge discovery across every single domain knowledge [1]-[3].

In spite of the rapid technology advancement in data series pattern recognition, the state-of-the-art time series pattern recognition algorithms still require users to at least provide the length of the time series pattern [4]. This is based on the assumption that the users or domain experts are conversant with the pattern to be mined from the raw time series data.

The complication exacerbates in the area of hidden motif discovering for financial time series, where the time series patterns and scale are not only unknown in advance, but also the price data series patterns can be highly scalable. In financial time series analysis, patterns could exist over various range of scales due to the fact that financial time series is highly dimensional with multiple time frame [5]. However, this is not the case in other application domains, where same patterns with different scale are regarded as unfavourable and will invoke deterioration at the expense of the similarity measurement [4]. This could possibly the contribution to the lack of literature related to the mining of scalable pattern recognition in time series data.

Essentially, scalable time series pattern recognition poses three major challenges. First, distance measure between patterns with different scales. Typical lock-step distance measures are meant to measure motif pair with the same length, they are impotent in assessing motif pair with different lengths. Second, domain specific features in financial time series are poorly comprehended by typical lock-step Euclidean distance measure. Conversely, elastic distance measure such as Dynamic Time Warping with substantially more expensive computational cost appears to be more synonym in discerning domain specific features of financial time series data [6]. Unfortunately, state-of-the-art algorithms such as Matrix Profile are developed by leveraging the Euclidean distance properties in order to significantly reduce the complexity of the of search algorithm [4], [7], [8]. Third, a brute-force solution even by using state-of-the-art search algorithms to search through all possible subsequence lengths even in a medium size time series data is computationally untenable [4].

As an initial effort to make scalable time series pattern recognition a reality, this study focuses on formulating a novel value based time series data representation technique. It intends to better assimilates and represent the features presence in financial data series. Apart from achieving the native goal of dimension reduction, the proposed data representation technique will automatically embed elastic transformation of time element during the data representation process. This will allow the deployment of standard Euclidean distance measure in state-of-the-art search algorithm to remain intact, yet producing elastic distance measure compatible results.

## II. Background

### A. Time Complexity of Search Space

Time series are a collection of data values recorded in sequential time order. In the context of financial time series, it refers to the series of prices recorded from transactions that occur over certain period time in a particular financial instrument.

In a time series that consists of a sequence of $n$ real values, where $n$ is the length of $T$:

$$T = (t_1, \ldots, t_n), \quad t_i \in R \tag{1}$$

Manuscript received September 10, 2019; revised March 5, 2020.

K. H. Sim, K. Y. Sim, and V. Raman are with the Swinburne University of Technology Sarawak, Jalan Simpang Tiga, 93350 Kuching, Sarawak, Malaysia (e-mail: khsim@swinburne.edu.my, ksim@swinburne.edu.my, vraman@swinburne.edu.my).

doi: 10.18178/ijmlc.2020.10.3.951

A subsequence *S* is a subset time series begin at offset *a* in *T* with *w* contiguous values:

$$S(a,w) = (t_a, \ldots, t_{a+w-1}), \text{ with } 1 \le a \le n - w + 1 \quad (2)$$

Subsequences of constant length *w* can be excerpted from each offset in *T*, and each subsequence is measured against the sample sequence. That is, a sliding window is slide by one offset of step, with $1 \le step \le n - w$, and the subsequence *S(a,w)* is excerpted at offset *a*. There are a total of $((n-w) / step) + 1$ subsequences in *T* [6]:

$$windows(T, w, step) = \bigcup_{i=0}^{\frac{(n-w)}{step}} S(i.step + 1, w) \quad (3)$$

In subsequence to subsequence motif search, subsequences are extracted from a long query time series *Q* of length *n* and a long time series *C* of length *m*. As such, split *Q* into *c* subsequences of length *w*:

$$windows(Q, w, step) = \{ S(1, w), \ldots, S(c,w) \} \quad (4)$$

Next, sample sequence to subsequence matching is applied for each subsequence $S(i,w)$ in $C$ :

$$D_{subsub}(Q,C) = \sum_{S \in windows(Q,w,step)} D_{seqsub}(S,C) \quad (5)$$

The subsequences excerpted from Q could also be shifted or disjoint subsequences [6]. Subsequence to subsequence matching has a computational complexity of $O(w(n-w)(m-w))$ with Euclidean distance as the distance measure [6]. Hence, subsequence to subsequence motif search within the same time series *T* where *Q=C*, the computational complexity can be stated as $O(w(n\text{-}w)^2)$.

In the context of searching for hidden patterns in financial time series, both pattern and scale are unknown in advance. In the example illustrated in Fig. 1, a famous price pattern in technical analysis known as triangle are spotted multiple times in a financial time series data, and they appear in very different scale. Assuming a practitioner is unaware of the existence of such kind of price pattern in the time series data, thus the goal of the data mining task is to discover them without expecting any input from the practitioner.

Fig. 1. Repeated pattern in financial time series.

Ideally, a motif search for repeated patterns hidden in financial time series should require no input of parameter from user. The current state-of-the-art algorithms still require user to provide at least one parameter, which is the desired length of the motifs [9], [10]. Ironically, the ease of which motif discovery are performed with the latest technology advancement, this single motif length parameter value that depends solely on user's experience or intuition has appeared to be too great of a burden.

In order to search for scalable pattern recognition in time series data, the search has to be executed on motif pair that may occur in different lengths. Both lengths of a motif pair may range from any value within all the possible subsequence length in time series *T*. As such, it is a subsequence to subsequence motif search in a long query *Q* of length *n*, with the objective to search for similar time series pattern between subsequence length *w* and another subsequence length *v*, where both *w* and *v* may range from 2 to *n*-1, where $w \neq v$. This will explode the computational complexity to $O(n^4)$, which is computationally untenable even for a medium size data series.

### B. Distance Measure

Most motif discovery algorithms in time series data employ Euclidean distance measure, a lock-step distance measure with less expensive computation complexity. This is due to the fact that most time series data are naturally massive in size, and computational complexity of the search algorithm is always the dominant constraint. Even the state-of-the-art motif discovery algorithm of Matrix Profile model operates by exploiting the properties of Euclidean distance to significantly reduce the computational complexity into linear time, making it one of the fastest search algorithm for fixed length subsequence to subsequence search [4], [7], [9], [10].

In Euclidean distance measure, the distance between two time series $Q = (q1, \ldots, qn)$ and $C = (c1, \ldots, cn)$, both of length *n*, has computational complexity of *O(n)*, and can be formally stated as [6]:

$$D_{ED}(Q,C)\sqrt{\sum_i (q_i - c_i)^2} \quad (6)$$

Nevertheless, Euclidean distance metric applies linear alignment of the time axis, it is therefore futile to provide invariances on time axis, and incapable to measure time series with different length [11], [12]. Due to these shortcomings, Euclidean distance favor straight line data series patterns over more complex data series patterns. This has defeated the purpose of time series pattern recognition, especially in financial time series data with strong random walk property.

On the other hand, Dynamic Time Warping (DTW) is a well-known elastic similarity measure that commonly recognized as the best time series similarity measure [13]-[15]. DTW has the competency to accommodate distortions in the time axis by implementing elastic transformation on time axis to align similar shapes with differences in time alignment [6].

In financial time series price chart analysis, price patterns of time series *C* and time series *Q* in Fig. 2 are exhibiting as higher high and higher low followed by lower low price

patterns, hence they are categorized under the same price movement category according to the convention of technical analysis. Unfortunately, the difference of skewness in the price movement between both time series draws demotion in Euclidean distance metric. Undoubtedly, elastic distance measure such as DTW is clearly a more competent similarity measure under this context.

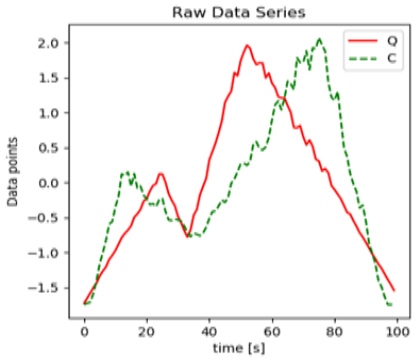



Fig. 2. Similar price pattern in financial time series.

DTW algorithm computes a cost matrix *M* that consists the distances between all pairs of values in time series *Q* and time series *C*. Then, DTW searches for minimum coast by traversal of the matrix *M* to derive optimal warping path. A warping path *p* in *M* is defined as a set of tuples that defines a traversal of the cost matrix, whereas i and j represent the data points of in *Q* and *C*. Thus, the DTW distance between two time series *Q* and *C* is defined as the path p through the cost matrix *M* with the minimal total distance [16]:

$$D_{DTW}(Q,C) = \min\left\{\sum_{(i,j)\in p}(q_i - c_i)^2 \mid p \in M\right\} \quad (7)$$

Regrettably, the distance matrix in DTW has the dimensionality of $n^2$, resulting a quadratic computational complexity of $O(n^2)$. This has made the implementation of elastic similarity measure almost impossible in time series pattern recognition. Apparently, certain application domains including financial time series analysis require elastic similarity measure. Obviously, this has been a deadlock in the pursuit of elastic similarity measure, while trying to maintain the complexity of the motif search algorithm in linear time.

### C. Fig Comparing Time Series with Different Scales

In the search for scalable time series patterns with different scales, similarity measurement has been one of the fundamental challenges. Patterns with similar visual shapes, but occur at different scales are often regarded as different patterns, and they draw punishment in Euclidean distance measure techniques. Besides, certain distance measure metrics will naturally favour patterns with shorter length, while other measurement techniques favour patterns with longer length [4].

In term of comparing motif pairs with variable length, various suggestions have been proposed by researchers over the past decade to neutrally compare two motif pairs with different length. This includes length-normalized Euclidean distance and factorize Euclidean distance [4].

It is essential to note that even the enhanced version of Euclidean distance search found in literatures are meant to compare two motif pairs of different length, each motif pair is still having two similar time series with the same length [4]. In other words, they are designed to compare two standard Euclidean distance measures of two motif pairs $D_x(T_{a,l}, T_{b,l})$ and $D_y(T_{i,g}, T_{j,g})$, motif pair *x* is found in time series *T* at data point *a* with length *l* and data point at *b* with the same length *l*. The Euclidean distance between motif pair *x*, $D_x$ is then compared with $D_y$, the Euclidean distance between motif pair *y*, which consists of time series *T* at data point *i* with length *g* and data point at *j* with same length of *g*, where $g \neq l$. In the context of comparing two time series patterns with different scale, $D_z(T_{a,m}, T_{b,n})$ involves time series *T* at data point *a* with length *m* and time series *T* at data point *b* with length *n*, where $m \neq n$. A standard Euclidean distance measure is non-functional in measuring the similarity of motif pair *z*.

There have been limited study done in comparing two time series patterns with different scale within the same motif pair, but this is the key element in scalable time series pattern recognition. Scalability of time series patterns may occur in both data value axis and time axis.

Apparently, scalability issue on data value axis can be instantly solved when z-normalization is performed on the data points [6]. However, scalability issue on time axis remains a challenge especially when the search is conducted using a sliding window across a huge search space.

## III. Value Based Data Representation

Financial time series pattern recognition has time complexity constraint by nature, this limits the option of distance metric to only lock-step distance measure, which possesses linear time complexity. Unfortunately, lock-step distance metric such as Euclidean distance measure has no time transformation property, yet these shortcomings are the most important constituents in scalable financial time series pattern recognition.

Instead of trying to reengineer elastic distance measure such as DTW to be less expensive computationally, this study intends to introduce a novel Value Based Data Representation technique at early stage of data representation, which the benefit will cascade down into distance measure at later stage. This ultimately allows standard Euclidean distance measure to produce result comparable to elastic similarity measurement.

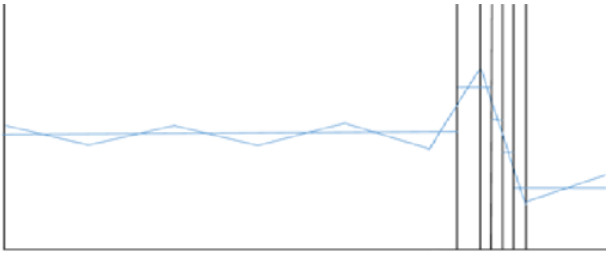


Fig. 3. Value-based data representation for financial data series.

As shown in Fig. 3, the proposed value-based data

representation technique performs dimensional reduction by summarizing prices based on the change of values between price data points against a pre-defined magnitude threshold (R). Notably, it does not base on a pre-defined time interval as practiced by majority of data representation techniques [17].

**Algorithm : Value-based Data Representation**

```
Input : Time series sequence X, Magnitude R
Output : vbDataRep[]

double startPoint = first data point of X
double vbDataRep []

loop each data point in X
        if (abs(startPoint - current data point) > R)
                vbDataRep.append(current data point)
                startPoint = current data point
        endif
endloop
```

Fig. 4. Pseudo-code of value-based data representation.

In traditional time series data representation, summarization of data points is done base on fix time interval, leads to potential wasting of many meaningless representations for areas with no signification movement of price data points. These data points could have just represented by one single price value instead of multiple representations with the same mean value. Contrary, the proposed Value Based Data Representation ensures price movement with significant changes to be better represented and captured in the dimension reduction process as illustrated in Fig. 3.

As explained in pseudo-code in Fig. 4, the key parameter for Value-based Data Representation algorithm is the threshold for magnitude of change (*R*) between price data points. The chosen threshold for price magnitude parameter has direct impact on the quality of the data representation and dimensional reduction outcome. Financial time series are naturally confounded by fluctuation between data points. Thus, the goal of Value Based Data Representation is to achieve dimensional reduction by preserving the major price movement pattern, while removing the noise of price fluctuation between price data points.

Fig. 5 presents the model of proposed scalable financial time series recognition, subsequence series of *q* and *c* are extracted from time series *T* in z-normalized form to maintain the original price pattern after normalization [6]. Since the proposed Value Based Data Representation utilizes the change of price data point to perform dimension reduction, it is therefore relatively vulnerable to noise in the raw data. Financial instruments with low liquidity tend to have sudden spike between price data points, which is not part of the major trend of price movement.

Therefore, a simply noise removal mechanism is instilled during the reading of subsequence data points for z-normalization process. Whenever a change between subsequent price data points surpasses one third of the maximum range of the subsequence data series, the sudden change of price may potentially be a spike that needs to be ignored in the representation. In such cases, this sudden exceptional change of price is required to persist for at least two consecutive data points at the same or greater level in order to be qualified as genuine price movement.

Subsequently, dimension reduction is performed on the normalized data points based on magnitude of change in data value. Nonetheless, the selection of threshold for magnitude parameter, *R* that is smaller than the price fluctuation range will fail to produce any degree of dimensional reduction in raw price data. Conversely, a relatively larger value of threshold for parameter magnitude will lead to the missing out of some significant price patterns during the data representation procedure.

Instead of relying on human expert to provide the most optimal value of price magnitude parameter, the proposed scalable financial time series pattern recognition model incorporates two pre-processing modules, namely First Level PAA module and Largest Change Identification module. These two modules are run on the raw time series data points in order to derive the most optimal threshold of *R* for magnitude parameter.

As such, both normalized subsequences that could be of different length are fed into first level of Piecewise Aggregate Approximation (PAA) pre-processing module. PAA is a common time series approximation technique that discretize data series into smaller number of discrete intervals with certain cardinality.

PAA approximates a time series *X* of length *m* into vector:

$$\overline{X} = (\bar{x}_1, \ldots, \bar{x}_m) \tag{8}$$

Any arbitrary length $M <= n$ where each of $\bar{x}_1$ is calculated as follows [16]:

$$\bar{x}_i = \frac{M}{n} \sum_{j=\frac{n}{M(i-1)}+1}^{(\frac{n}{M})i} x_j \tag{9}$$

The pre-processing of First Level PAA aims to equalize the length of both subsequence q and subsequence c. This is achieved by performing PAA on shorter subsequence to scale the length of the shorter subsequence series up to the length of the longer subsequence. This is in contradiction to typical PAA operations that commonly reduces the number of data points in the data series.

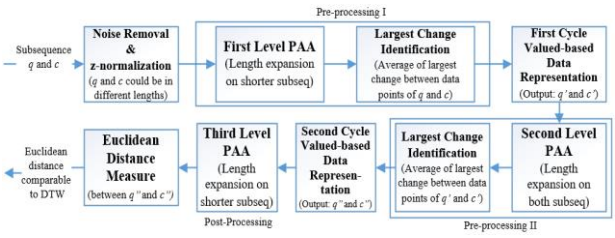



Fig. 5. Scalable financial time series pattern recognition model.

In elastic distance measure, more data points provide higher possibility for a finer matching of the most optimal distance measure between data points, but this also leads to the growth of time complexity in exponential manner as the

size of data points increases. Fortunately, scaling the length of subsequence data series up by using PAA only requires linear time complexity. Scaling up of subsequence length increases the size of data points in shorter subsequence series, and length equalization of both subsequces can be accomplished at the same time. In fact, scaling of subsequence's length either up or down between two subsequences with PAA has the same linear time complexity.

The second module of pre-processing is to identify the optimal threshold for a given subsequence data series based on the characteristic of price data fluctuation in the data series. This module removes the need of input from human expert and the ambiguity of setting the threshold value for magnitude parameter, warranting optimum output from Value Based Data Representation.

The idea is to find the largest change occurs between each price data points in a subsequence data series. This value eliminates all the noises possibly presence in the data series due to fluctuation between subsequent price data points. It is important to note that this module is performed after both subsequences of time series have been scaled into equal length through First Level of PAA. The average of the largest change between data points within both subsequece series is then used as the threshold value for magnitude parameter (*R*) to perform Value Based Data Representation.

Next, the proposed Value Based Data Representation is performed by reading the data points in subsequence data series in sequential order, a price data point will only be recorded if the accumulating change of price data points value exceeds the threshold *R* of magnitude parameter (refer Fig. 4). As a result, a new shorter time series sequences *q'* and *c'* are generated as the output from this data representation process. Essentially, the mechanism of skipping data points that do not fulfil the price value threshold has indirectly incorporating time transformation element into the output of Value Based Data Representation.

In order to fully harness the benefit of time transformation effect of the proposed data representation approach especial for more complex time series patterns, a second cycle of the same Valued Based Data Representation process is repeated on the subsequeces of *q'* and *c'*, which are the output of the First Cycle Value Based Data Representation.

However, there is no guarantee on the number of segments in each subsequence time series after the First Cycle Value Based Data Representation, thus it is very likely that both subsequences of *q'* and *c'* are in different length. Consequently, a Second Level PAA is carried-up as the first module in pre-processing II to scale up the length of both subsequences to be equal in length with the size of at least 100 data points. Experiment showed that a longer length size of beyond 100 data points has no significant effect on the outcome of Second Cycle Value Based Data Representation.

The Largest Change Identification module in Pre-processing II adopts the same search operations as Pre-processing I, except that the focus in this module is shifted into deriving a price magnitude threshold to ensures the major change of price data points is well represented.

Finally, the Second Cycle of Value Based Data Representation will be executed on subsequence series *q'* and *c'*. This cycle of data representation observes identical algorithm described in First Cycle Value Based Data Representation. The goal is to embed time transformation on any complex price patterns leftover from the first cycle of data representation process.

Similar to the output of First Cycle Valued-based Data Representation, the Second Cycle Value-based Data Representation produces subsequences *q''* and *c''* that could be of different lengths. Since Euclidean distance can only operate on two subsequences with equal size, Third Level PAA will have to be performed to equalize the length of both subsequences by expanding the length of the shorter subsequence, this aims to prevent any loss of information before Euclidean distance measure is employed.

Although the proposed scalable time series pattern recognition model employs multi-level PAA, double-cycle Value Based Data Representation may raise concerns over the overall time complexity, it is important to note that all the modules after the First Cycle Value-based Data Representation are operating on subsequences with fix length of less than one hundred data points in size. Hence, those modules are independent of the initial length of the original subsequence, and they have only constant time complexity.

## IV. Experiments and Discussion

In order to allow fair comparison across experiments that involve motifs of different lengths, a length-normalized Euclidean distance has to be adopted. Fundamentally, standard Euclidean distance has obvious bias toward time series pattern with shorter length, but length-normalized Euclidean distance favors longer time series patterns.

As a consequence, a fair comparison regardless of the time series length can only be achieved by factorizing the Euclidean distance by *sqrt*(*1/L*), where *L* is the length of the time series sequence. Study has shown that *sqrt*(*1/L*) correction factor provides a near to perfect invariant distance over different lengths [4]. Thus, max-normalized by the square root of length Euclidean distance, that has the range of 0 to 1 is deployed throughout all the experiments in this study.

The evaluation on the competency of double-cycled Value Based Data Representation begins with two time series motifs consist of subsequence *Q* and *C*. Both data series demonstrate bi-directional price pattern, but different in skewness as illustrated in Fig. 6. Data series motif 2 exhibits bigger skewness difference than data series motif 1. As a result, the max-normalized Euclidean distance for data series motif 2 has deteriorated almost three times from 0.241 recorded in data series motif 1 to 0.726 in data series motif 2, while DTW distance only degenerates for about two folds from 0.038 in data series motif 1 to 0.077 for data series motif 2. This comparison clearly distinguishes the superiority of elastic distance measure such as DTW over common lock-step distance measure represented by Euclidean distance in dealing with the presence of time variability exists within the original data series.

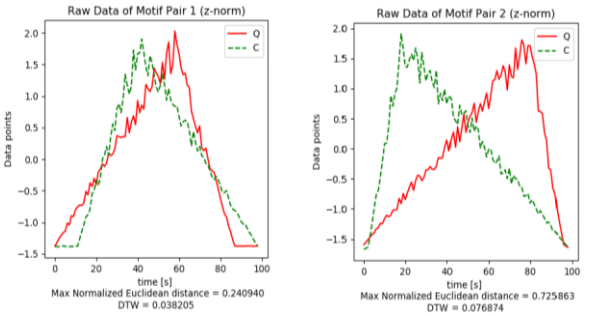



Fig. 6. Raw data of bi-directional motif pair with different skewness.

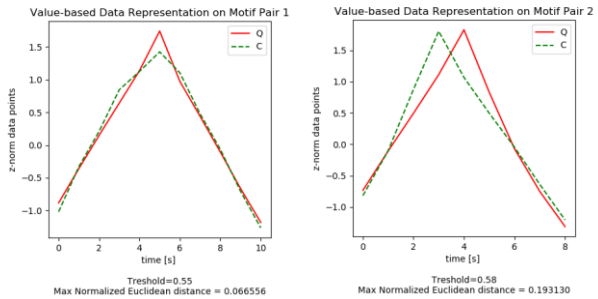



Fig. 7. Result of Value-based Data Representation on motif 1 & 2.

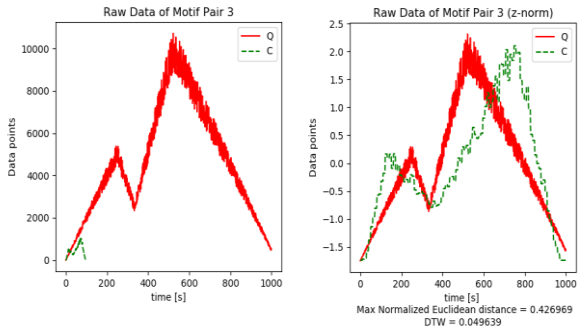



Fig. 8. Raw data of motif pair with scalable multi-directional pattern.

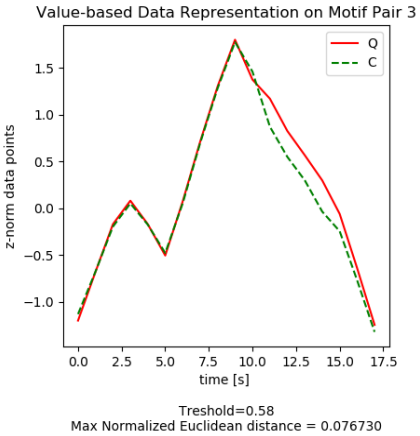



Fig. 9. Result on scalable multi-directional pattern in motif pair 3.

Subsequently, the proposed Value-based Data Representation is deployed on the two same data series motifs. As observed in Fig. 7, max-normalized Euclidean distance has tremendously improved to 0.067 for data series motif 1, more than 75% improvement compares to the original data series. Similarly, the alignment of both data series in motif 2 has significantly enhanced and led to over 70% improvement in max-normalized Euclidean distance, it drops from 0.726 to as low as 0.193. Importantly, this result shows that the proposed Value Based Data Representation handles time variability in data series proficiently, making elastic measure comparable result possible by just using light-weight lock-step Euclidean distance measure.

The proposed Value Based Data Representation is further assessed with a more complex price pattern in highly scalable context. Diagram on the left in Fig. 8 presents a raw data series of motif pair 3, which is not only substantially different in scale, but also demonstrates multi-directional price pattern. In this motif pair, data series *Q* is ten times larger in scale compares to data series *C*, but both poses similar price pattern with different skewness. Diagram on the right outlines the similarity of price pattern in this motif pair after z-normalization of the raw data series. The length of *C* has also been expanded to allow Euclidean measure calculation. The max-normalized Euclidean distance between these two normalized raw time series has lodged a reading of 0.427.

Double-cycled Value Based Data Representation is also applied to the normalized motif pair 3 and the outcome is presented in Fig. 9, advanced alignment of time variability has ameliorated the max-normalized Euclidean distance all the way to only 0.076. The proposed Value-based Data Representation continues to exhibit its supremacy in incorporating time transformation into multiple directional data series data, even the original time series subsequences are excessively vary in scale.

In order to assess further the robustness of the proposed Value Based Data Representation approach in real world application under the context of real market condition for financial time series, daily price data series of Dow Jones Industry Average (DJI), Gold Futures (XAUUSD), Bitcoin (BTC/USD) and WTI Crude Oil Futures from 2002 to 2018 were used. Four types of price series patterns are extracted from these financial instruments, namely mono-directional price pattern, bi-directional price pattern, multi-directional price pattern and randomized price pattern.

Table I summarize the results of the experiments. Value Based Data Representation recorded the best performance in Bi-directional price pattern with a substantial improvement of 37.44% on overall Euclidean distance. The average Euclidean distance after implementing the proposed data representation has been reduced to 0.284, obvious decrements from the average Euclidean distance of 0.454 logged from the original data series. This aligns with the expectation that bi-directional price movement allows Value Based Data Representation to harness the maximum benefit of time transformation between the two original data series.

TABLE I: RESULTS OF FINANCIAL INSTRUMENTS

| **Price Pattern** | **Comparison of Different Price Patterns** | | |
|---|---|---|---|
| | ***ED on raw data*** | ***ED of VBDRep*** | ***Improvement (%)*** |
| Mono-directional | 0.248 | 0.162 | 34.68 |
| Bi-directioinal | 0.454 | 0.284 | 37.44 |
| Multi-directional | 0.519 | 0.368 | 29.11 |
| Random | 0.725 | 0.636 | 12.28 |

It is worthwhile to note that the proposed Value Based Data Representation has achieved not only obvious improvement in the reading of Euclidean for bi-directional price pattern, but the improvement is also comparable to mono-directional price pattern. This is indeed a significant

accomplishment as conventional Euclidean measure has a strong bias toward mono-directional price movement [6].

Moreover, the Value Based Data Representation approach posted an overall of 29.11% reduction in Euclidean distance measure for multi-directional price pattern, despite of the surge in the complexity of multi-directional price pattern.

Though time series subsequences with randomized price patterns inevitably recorded the worst Euclidean distance of 0.725, the proposed data representation still managed to contribute to an improvement of 12.28% on the average Euclidean distance measure in that category.

## V. Conclusion

The proposed Value-based data representation possesses the element of time transformation that allows elastic distance measure comparable output by simply using standard Euclidean measure. The algorithm of this data representation approach only requires an update to the largest change between data point whenever a new data point is added to the subsequence during a sliding window search. Recalculation of the whole subsequence is only required when the data point that produces the largest change is dropped after the sliding window shift. Since the PAA in the pre-processing module still requires recalculation of the whole subsequence in case the sliding window is run on the shorter subsequence of the motif search, future research will explore other dimension reduction technique that does not entail the recalculation of the entire subsequence for subsequence length adjustment. Future works will also attempt to implement the proposed data representation approach into conventional time series search algorithm such as k-nearest neighbor (KNN) or Support Vector Machine (SVM), where elastic distance measure is required. The ultimately goal is to reduce the computational complexity of the search process, making the search of hidden patterns in time series data with elastic distance measure a reality.

## Conflict of Interest

The authors declare no conflict of interest.

## Author Contributions

Kwan-Hua conducted the research, analyzed the data and wrote the paper; Kwan-Yong and Valliappan assisted in the pre-processing of data, and provided technical advice on the testing and experiment; all authors had approved the final version.

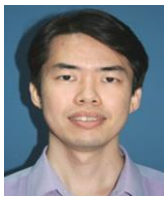
## Acknowledgment

This work was funded by Fundamental Research Grant Scheme (FRGS), Grant No: FRGS/1/2015/ICT04/SWIN/02/1.

## References

[1] T. Palpanas, "Data series management: The road to big sequence analytics," *SIGMOD Record 44*, vol. 2, pp. 47–52, 2015.
[2] U. Raza, A. Camerra, A. L. Murphy, T. Palpanas, and G. P. Picco, "Practical data prediction for real-world wireless sensor networks," *IEEE Trans. Knowl. Data Eng*, 2015.
[3] K. Mirylenka, V. Christophides, T. Palpanas, I. Pefkianakis, and M. May, "Characterizing home device usage from wireless traffic time series," *EDBT*, 2016.
[4] M. Linardi, Y. Zhu, T. Palpanas, and E. Keogh, *VALMOD - Scalable Discovery of Variable-Length Motifs in Data Series*, 2018.
[5] P. J. Kaufman, *Trading Systems and Methods*, John Wiley & Sons, New Jersey, pp. 80-91, 2013.
[6] P. Schäfer, "Scalable time series similarity search for data analytics," Ph.D. Thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät, pp. 16-19, 2015.
[7] C. C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. F. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. Keogh, "All Pairs similarity joins for time series: A unifying view that includes motifs," *Discords and Shapelets*, 2016.
[8] Y. Zhu, C. C. M. Yeh, Z. Zimmerman, K. Kamgar, and E. Keogh, *SCRIMP++: Time Series Motif Discovery at Interactive Speed*, 2018.
[9] Y. Zhu, M. Imamura, D. Nikovski *et al*., "Introducing time series chains: a new primitive for time series data mining," *Knowledge and Information Systems*, vol. 60, no. 1135, 2019.
[10] A. D. Hoang and E. Keogh, "A generic technique to incorporate domain knowledge into motif discovery," in *Proc. the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 125-134.
[11] P. Schäfer. (2014). The BOSS is concerned with time series classification in the presence of noise. [Online]. Available: https://www2.informatik.hu-berlin.de/~schaefpa/boss.pdf
[12] P. Schäfer, "Experiencing the shotgun distance for time series analysis," *Transactions on Machine Learning and Data Mining*, vol. 7, no. 1, pp. 3–25, 2014.
[13] H. Ding, "Querying and mining of time series data: Experimental comparison of representations and distance measures," *VLDB*, 2008.
[14] A. Bagnall and J. Lines, "An experimental evaluation of nearest neighbour time series classification," *arXiv Preprint*, 2014.
[15] J. Lines and A. Bagnall, "Time series classification with ensembles of elastic distance measures," *Data Min. Knowl. Disc*., pp. 1–28, 2014.
[16] J. Lin, S. Williamson, K. D. Borne, and D. DeBarr, "Pattern recognition in time series," *Advances in Machine Learning and Data Mining for Astronomy*, pp. 617-647, 2012.
[17] K. H. Sim, K. Y. Sim, and N. Bong, "Dynamic time interval data representation in scalable financial time series pattern recognition," in *Proc. 2018 2nd International Conference on Computer Science and Artificial Intelligence*, 2018, pp. 120-125.





**K. H. Sim** was born in Kuching in 1975. He received his BCompSci (Hons) from University Malaysia Sabah in 1999 and MSc. (IT) in 2001 from University Malaysia Sarawak. He is currently a lecturer at the School of ICT, Faculty of Engineering, Computing and Science, Swinburne University of Technology Sarawak, Malaysia. His has keen research interests in financial time series analysis, data mining and statistical analysis.

**K. Y. Sim** was born in Kuching in 1976. He received his Beng (hons) from the National University of Malaysia in 1999, and masters of computer science from University of Malaya, Malaysia in 2001. He received his doctorate of philosophy from Swinburne University of Technology, Melbourne. He is currently a senior lecturer and the head at the School of Engineering, Faculty of Engineering, Computing and Science, Swinburne University of Technology, Sarawak Campus, Malaysia. His research interests include software testing and for embedded system testing.

**R. Valliappan** was born in India. He completed his bachelor of engineering in computer science in 2002. He has completed his master of science in 2005 and PhD in 2015 from University Sains Malaysia. Currently he is working as a senior lecturer at Swinburne University of Technology Sarawak, Kuching, Malaysia. His research interests are in medical imaging, data analytics and Internet of Things. Dr. Valliappan has worked as team to acquire many external research grants and published papers in impact factor journals.