PLA Data Reduction for Speeding Up Time Series Comparison

Bachir Boucheham
Department of Informatics, University of Skikda, Algeria

Abstract: *We consider comparison of two Piecewise Linear Approximation (PLA) data reduction methods, a recursive PLA segmentation technique (Douglas-Peucker Algorithm) and a sequential PLA-segmentation technique (FAN) when applied in prior of our previously developed time series alignment technique SEA, which was established as a very effective method. The outcome of these two combination are two new time series alignment methods: RecSEA and SeqSEA. The study shows that both RecSEA and SeqSEA perform alignments as good as those of SEA with important reductions in data (RecSEA: up to 60%, SeqSEA up to 80% samples reduction) and processing time(RecSEA: up to 85%, SeqSEA up to 95% time reduction) with respect to the SEA method. This makes both the two new methods more suitable for time series databases querying, searching and retrieval. Particularly, SeqSEA is significantly much faster than RecSEA for long time series.*

Keywords: *Pattern matchin, data reduction, time series comparison, time series alignment, datamining, data retrieval.*

Received May 17, 2010; accepted October 24, 2010

1. Introduction

Querying, searching, and mining of time series are important issues in many braches of science and technology. They are particularly useful in knowledge discovery, process monitoring and in diagnosis of systems generating these time series. All these applications share in common one basic operation: Comparison of two given time series patterns. That is, given two time series, the comparison operation consists in proposing a way to tell if these two patterns are similar in shape. One of the two time series stands in general for the reference (say X), whereas the second for the target (say Y). For instance, in speech recognition, X might be a known word or sound and Y a recorded word or sound to be compared with X. Other examples include economic time series and physiological time series where some particular patterns are significant to experts in each field.

Numerous methods have been proposed for time series comparison. However, the Dynamic Time Warping (DTW) [12] is the technique named by many as the most suitable comparison technique. The strength of the DTW is its ability in taking into account the time axis shift and/or scaling problems. It can also align time series of different lengths. The main weakness of DTW is its high computational complexity. In [4], we proposed an alternative time series comparison technique (SEA) that was shown to allow more accurate comparisons using less resource main memory and CPU time than DTW. In This study, we propose enhancement of the SEA method through introduction of a data reduction technique as a pre-processing step. The aim is to come up to methods that are as good as SEA in time series alignment and that are less data and time consuming. For this purpose, we consider two Piecewise Linear Approximation (PLA) data reduction techniques: A recursive segmentation method inspired from the Douglas-Peucker Algorithm (DPA) [3] and a sequential segmentation technique inspired from the classical FAN method [9]. These are classical PLA data reduction techniques in time series [5]. When combined to the SEA method, the outcome is two new time series comparison methods that we refer to respectively by RecSEA and SeqSEA.

Briefly, the two new methods work as follows. In both RecSEA and SeqSEA, there is first Dominant Points (DPs) extraction from the two time series to compare (X and Y). This step aims at extracting as few significant points as possible from the time series traces to perform the matching. In the second step, the SEA matching method is applied to the two sets of extracted DPs. Precisely, the DPs are sorted on magnitude, then, there is exchange of the two sets of DPs magnitudes. Finally, the two sets of DPs are brought to natural order and the two time series are reconstructed by the inverse data reduction algorithms.

The results of the comparison are evaluated numerically through the correlation factors of the two given time series, each with respect to its reconstructed correspondent. Comparison of RecSEA, SeqSEA and SEA on ECG time series show that both SeqSEA and RecSEA yield alignments that are as good as those of SEA, but both of them use much less samples and consume much less CPU time than SEA. The data reduction was about 60% for RecSEA and about 80% for SeqSEA with respect to the original samples necessary for SEA. The time reduction was up to 85%

for RecSEA and 95% for SeqSEA with respect the SEA execution time. However, the SeqSEA is much faster than RecSEA for long time series.

The rest of this paper is organized as follows. In section 2, there is review of existing time series comparison methods. In section 3, the proposed RecSEA and SeqSEA methods are presented. In section 4, applications on the RecSEA, SeqSEA and the SEA methods are performed. In section 5, obtained results are discussed. Lastly, in section 6, a general conclusion is dressed.

2. Existing Time Series Comparison Methods

There are two categories of time series matching methods: Comparison techniques and alignment techniques. Let $X=(x_i)$, $i=1:n$ and $Y=(y_j)$, $j=1:m$, be two given time series. The first class renders a distance that reflects the degree of similarity between X and Y; whereas alignment methods perform a mapping between the points of X and those of Y. Of course, alignment methods render also a comparison measure. The euclidian distance equation 1 seems to be the first used time series comparison. This is basically due to its simplicity, but, also to its interesting linear temporal and spatial complexities. However, the Euclidian distance is reported to be very sensitive to the time axis shift and/or scaling and to local noise, which constitute the main difficulties in time series comparison/alignment. A DFT based method was proposed by Agrawal *et al.* [1] belongs also to this class. In this method, the two time series are mapped from the time domain to the frequency domain through the DFT. Then, the first most significant k coefficients of the DFT in each series are used in the comparison through the Euclidian distance. To all evidence, this technique compares the two time series, but does not align them. Shatkay and Zdonic [13] proposed also a comparison method. Their method consists in transforming first the raw data into characters defined over a limited alphabet and then in applying string matching techniques. However, the effectiveness of this method is limited, since data reduction by quantification of time series leads to important data distortion. A histogram based comparison method was proposed by Chen and Özsu [7]. Since histograms ignore the temporal dimension of the data sequences, this category of methods is capable of rendering only a global similarity measure. However, no alignment at the point-to-point level is possible in such methods. Another comparison method was proposed by Bozkaya *et al.* [6] and is referred to by Longest Common Sub-Sequence (LCSS). This method is based on a modified version of the edit distance [10]. This method allows non-linear mapping between the two time series. However, the threshold on the edit distance is very difficult to be set.

The other category of methods performs alignment between the two time series. The most popular and used time series alignment techniques is the DTW [12]. This technique has been used by many researchers in resolving many technological and scientific problems in numerous fields. As examples of DTW applications, one can cite speech recognition [2], music retrieval [16], ECG recognition [14] and general time series mining [11]. The main advantage of the DTW is its remarkable ability in taking into account the time axis shift and/or scaling problems. However, the DTW main problem resides in its high temporal complexity. SEA [4] is a method that was recently proposed for time series matching. It was shown to be more effective than DTW in the precision of matching. It also consumes less memory and time than DTW. In the next section, we propose two more effective time series comparison methods that we refer to by the acronyms RecSEA and SeqSEA. The new methods are based on combination of classical, yet very efficient, data reduction techniques, namely, FAN and Douglas-Peucker Algorithm, with our previously developed method SEA.

$$Euclidian(X,Y) = \sqrt{\sum_{i=1}^{i=n}\left(x_i - y_i\right)^2} \tag{1}$$

3. The RecSEA and SeqSEA Methods

As stated above, the proposed methods are both composed of two main step: Data reduction (with Douglas-Peucker or FAN algorithms) and Matching (with SEA).

3.1. PLA Data Reduction

There are many techniques for data reduction in time series. The curve simplification approach is one of the most used for such purpose. It consists briefly in what follows. Let $P=(p_i)$, $i=1..N$, where $p_i=(x1(i),x2(i))$, with $x1(i)$ the horizontal (temporal) coordinate and $x2(i)$ the vertical (magnitude) coordinate of p_i, be a given discrete curve. The simplification of P to a given precision ε, $\varepsilon>0$, a preset threshold on the tolerance of the approximation, consists in computing another polyline $Q=(q_j)$, $j=1..K$, satisfying the following conditions [15]:

a. $K<N$; (data reduction rule).
b. $q_1=p_1$ and $q_K=p_N$; (endpoints must coincide).
c. Let ||., .|| be a distance defined on discrete curves. Then $||P, Q|| <\varepsilon$.

We use the DPA [8] for the RecSEA method and FAN [9] for the SeqSEA method for the points of Q determination. The Douglas-Peucker algorithm uses a recursive strategy in segmenting the time series curves, whereas the FAN algorithm uses a sequential strategy. As a result, both methods have strengths and weaknesses when applied as pre-processors for the

SEA method. For ease of notation, we refer to the data reduction step by DR (Douglas-Peucker or FAN). We also need to use the generic common acronym DRSEA for both RecSEA and SeqSEA. Following the DR process, we compute the data reduction ratio (in terms of samples reduction) by equation 2:

$$DRR(P,Q)=(1-2\frac{|Q|}{|P|})\ x\ 100\% \tag{2}$$

3.2. Matching

- *Step 1:* Both time series to align *X* and *Y* are passed through the DR procedure, described in sub-section 3.1. The outcome of processing a time series with the DR procedure is a set of (perceptually) significant points on the time series curve. Let *Xs* be the reduced set of points for *X* and *Ys* that of *Y*. Note here that each element of *Xs* (respectively *Ys*) is represented by two explicit coordinates: a temporal index Xs_1, and a magnitude value Xs_2 (respectively Ys_1 and Ys_2).
- *Step 2 Signature Establishment:* Time series $Xs=(Xs_1,Xs_2)_i$, $i=1:k$, which is initially ordered on the temporal value is reordered on the magnitude value. The obtained traces are referred to in this study as signature*(Xs)*. This operation is performed for both *Xs* and *Ys*. The obtained signature*(Xs)* and signature*(Ys)* will be used for the matching of *X* and *Y* through *Xs* and *Ys* alignment. The comparison is explained in the next step.
- *Step 3 Magnitude Exchange and Comparison:* In the third step of the matching, there is exchange of the magnitudes between the two time series *Xs* and *Ys*. That is, time series *Xs* will 'wear' the magnitude of time series *Ys* and vice versa. The comparison is then performed for each time series (e.g., *Xs*) with its reconstructed correspondent resulting from the exchange step (e.g., Xs_{Rec}), using the corr factor equation 3 as an objective criterion and visual inspection as a subjective criterion. Note here that the method handles time series of different lengths since the effective alignments and computed correlations are performed on equal length series (Xs, Xs_{rec}) and (Ys, Ys_{rec}).

$$Corr\ (X,\ Y) = \frac{cov(X,Y)^2}{var(X).var(Y)} \tag{3}$$

In the following, we present the generic DRSEA method in a more formal way. (Recall, DR stands for DPA or FAN).

3.3. The DRSEA Algorithm

Let $X=(X_1^{(i)},\ X_2^{(i)})$, $i=1..n$, *and* $Y=(Y_1^{(j)},\ Y_2^{(j)})$, $j=1..m$, *be the original time series to match, where:*

X_1: *is the temporal index of time series X;*
X_2: *is the magnitude index of time series X;*
Y_1: *is the temporal index of time series Y;*
Y_2: *is the magnitude index of time series Y;*

Let also Sort-on-Magnitude-Value be a procedure that sorts any input time series on the magnitude coordinate; and Sort-on-Temporal-Index the inverse procedure of Sort-on-Magnitude-Value.

Let also $DR(X,\varepsilon) \rightarrow Xs$ be the procedure that performs samples reduction on the given time series *X* curve by extraction of perceptually most significant points *Xs* to a given precision ε (sub-section 3.1).

The generic DRSEA algorithm is then as follows:

a. *Data reduction:*
 $DR(X,\varepsilon) \rightarrow Xs;$
 $DR(Y,\varepsilon) \rightarrow Ys;$

b. *Sorting on Magnitude:*
 - $X's=(X's_1,X's_2) \leftarrow$ *Sort-on-Magnitude-Value(Xs);*
 $X's_1=$*temporal-index(X's),* Xs_2: *magnitude-Value(Xs).*
 - $Y's=(Y's_1,Y's_2) \leftarrow$ *Sort-on-Magnitude-Value(Ys);*
 $Y's_1=$*temporal-index(Y's),* Ys_2: *magnitude-Value(Ys).*

c. *Normalization:* If $n \neq m$, then $X's_2$ and $Y's_2$ are normalized as described below.

d. *Signature Exchange:* There is exchange of the magnitudes between the two reduced time series:
 $X'' \leftarrow (X's_1,\ Y's_2)$: *X'' uses magnitudes of Y's and time indexes of X's.*
 $Y'' \leftarrow (Y's_1,\ X's_2)$: *Y'' uses magnitudes of X's and time indexes of Y's.*

e. *Reconstruction and Matching:* let Xs_{Rec} (resp. Ys_{Rec}) be the reconstructed time series as a result of reordering X'' and Y'' on their respective temporal index. Formally:
 $Xs_{rec}=$*Sort-on-Temporal-Index(X''): The reconstructed time series of Xs.*
 $Ys_{rec}=$*Sort-on-Temporal-Index(Y''): The reconstructed time series of Ys.*
 Note here that there is linear interpolation between the successive points of *Xs* and *Ys*.

f. *Times Series of Different Lengths:* using the same notations above, and assuming that $|X|=n \neq |Y|=m$, the comparison is performed by first applying a linear mapping between the two magnitudes $X's_2$ and $Y's_2$ and the corresponding time indexes $X's_1$ and $Y's_1$.

4. Comparison of RecSEA, SeqSEA and SEA

In this section, the data reduction based methods for time series alignment RecSEA and SeqSEA are compared to the original SEA method and to each other. The aim of the comparison is twofold. First, we establish that both RecSEA and SeqSEA are as good as SEA in aligning time series. For this objective we rely on the visual inspection of the alignments and on the correlation factor between each given time series (X or Y) and its respective reconstructed. The different experiments have been conducted on electrocardiogram (ECG) time series. This type of physiological data reflects heart activity. It is mainly a quasi-periodic signal where each period is composed of three main components: P wave, QRS complex and T wave. A normal two period ECG signal is presented in Figure 1. We shall mention also that all ECG data used in this study has been selected from the Massachusetts Institute of Technology-Beth Israel (MIT-BIH) ECG database. This is a collection of electrocardiograms recorded at 360Hz frequency and used by researchers as a reference for cross comparison of obtained results. However, the developed methods are useful for nearly any kind of time series.

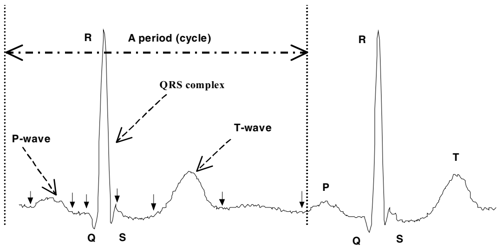

Figure 1. A typical ECG segment, where the sub-patterns are delimited by up-down arrows.

The figure shows a two period ECG. Each period is composed of three clinically significant basic patterns: P wave, QRS complex and T wave. Each period corresponds to one heart beat cycle.

The first experiment is presented in Figures 2, 3 and 4, where *X* and *Y* are segments from the same rec.111 of the MIT-BIH database (*X*: taken at the very beginning and *Y* taken at the very end of the record). Due to the nature of the ECG, these two segments, although presenting some local differences, are basically similar. In Figure 2, *X* and *Y* have been matched by the original SEA method. The original time series to match (*X* and *Y*) are given in the upper sub-plots. The middle plots present the original time series (*X* or *Y*) versus the respective reconstructed (*Xrec* or *Yrec*). The lower plots present the difference between the original time series (*X* or *Y*) and its reconstructed correspondent (*Xrec* or *Yrec*). From the middle plots and the lower plots, it can be seen that SEA did a very good job in aligning two quasi-similar time series, although containing many differences: local perturbations, time-magnitude axis shift/scaling, different number of periods and different lengths. Numerically, the (mean) correlation factor was 0.997 and the (total) matching time was 1.156 seconds. Figure 3 shows the same segments aligned by RecSEA. The results are presented in the same fashion as in Figure 2. It can be seen that RecSEA did an alignment as good as that of SEA. Numerically, the (mean) correlation factor was 0.993 and the matching time was 0.860 seconds. Since RecSEA is a data reduction based matching method, we computed the (mean) data reduction ratio which was 58.9%. That is RecSEA used only about 1/3 of the original samples to perform as good as SEA in much less time.

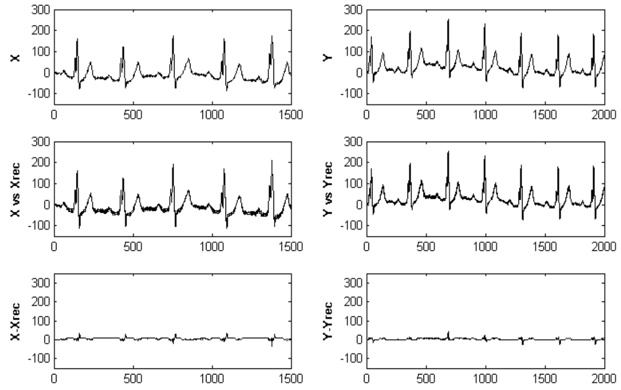

Figure 2. Results of matching two similar ECG time series by SEA. Upper plots: Orginal time series (*X* and *Y*); middle plots: Original plots (*X* or *Y*) versus its reconstructed time series (*Xrec* or *Yrec*); lower plots: Difference between the original time series (*X* or *Y*) and its reconstructed correspondent (*Xrec* or *Yrec*).

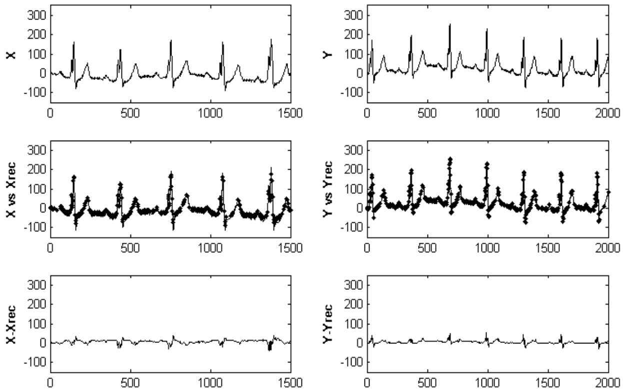

Figure 3. Results of matching the same two similar ECG time series (Figure 2) by RecSEA. Upper plots: Orginal time series (*X* and *Y*); middle plots: Original plots (*X* or *Y*) versus its reconstructed time series (*Xrec* or *Yrec*); the computed DPs are plotted as small dots; lower plots: Difference between the original time series (*X* or *Y*) and its reconstructed correspondent (*Xrec* or *Yrec*).

Figure 4 shows the same segments aligned by SeqSEA. The results are presented in the same fashion as in Figure 2. It can be seen that SeqSEA did also an alignment as good as that of SEA. Numerically, the (mean) correlation factor was 0.990 and the matching time was 0.750 seconds. Since SeqSEA is also a data reduction based matching method, we computed the (mean) data reduction ratio which was 80.2%. That is SeqSEA used only about 1/5 of the original samples to perform as good as SEA in much less time. For the purpose of establishing the relation between the different numerical factors (matching time, correlation factor and data reduction ratio) for RecSEa and SeqSEA with respect to SEA, we applied the three methods on segments taken from the same record (rec.111) of increasing sizes (1000 to 50000). The summary of the obtained results are reported in Figures 5 and 6. Figure 5 reports the (mean) correlation factors of the three methods as a function of the time series length, whereas Figure 6 reports the percentage of time reduction and the percentage of data reduction for both RecSEA and SeqSEA with respect to SEA.

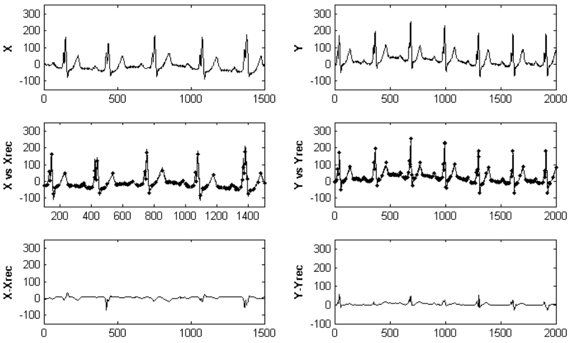

Figure 4. Results of matching the same two similar ECG time series (Figures 2 and 3) by SeqSEA. Upper plots: Orginal time series (*X* and *Y*); middle plots: Original plots (*X* or *Y*) versus its reconstructed time series (*Xrec* or *Yrec*); the computed DPs are plotted as small dots; lower plots: Difference between the original time series (*X* or *Y*) and its reconstructed correspondent (*Xrec* or *Yrec*).

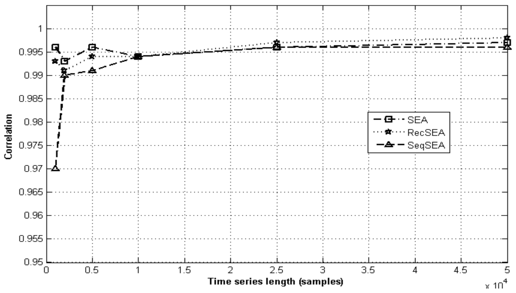

Figure 5. Correlation for SEA (square), RecSEA (stars) and SeqSEA (triangles) as a function of time series length.

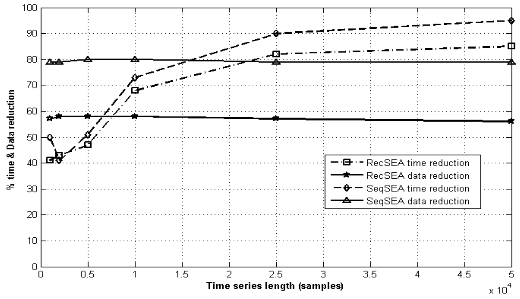

Figure 6. Data reduction and time reduction for RecSEA (resp. stars and squares) and SeqSEA (resp. triangles and pentagones).

Clearly, the correlation factors of the three methods are very close to each other and to the perfect 1 value. On the other hand, both RecSEA and SeqSEA show important data and time reduction with respect to SEA. But it is important to note also that SeqSEA performs better than RecSEA both on data reduction (about 80% versus 60%) and matching times reduction (40-95% versus 40-85%).

5. Discussion

The obtained results show that the PLA data reduction step was very effective in enhancing the original SEA method in the sense that the proposed data reduction based matching techniques perform as good as the SEA method, whereas they consume much less samples and time. It is worth noting the existence of other ideas for speeding up time series matching techniques. For instance, Junkui and Yuanzhen [10] used the early abandon strategy to accelerate DTW [12]. They report interesting results regarding time reduction. However, this technique does not target data reduction. Our results suggest that RecSEA and SeqSEA would be very effective time series alignment and comparison techniques. Particularly, since RecSEA and SeqSEA allow both important data reduction and time reduction, they both would be especially very suitable tools for large databases searching, querying and mining. Our results show also that the sequential data reduction strategy performs better than the recursive one in terms of data and time reduction. In particular, SeqSEA is more effective for dealing with very long time series as a whole entity to match. This last finding could be even expanded in future works to better take advantage of the SeqSEA method.

6. Conclusions

In this study, we designed two novel PLA-data reduction based techniques for time series comparison. The data reduction step was injected into the existing time series matching method SEA. The study showed the effectiveness of this step in the sense that the obtained two methods are as good as the original SEA method in aligning time series but are more efficient.

For future works, we plan to investigate other data reduction techniques combined with SEA and other time series alignment techniques.

Acknowledgements

This work has been partially supported by the ministry of higher education and scientific research of the algerian government through a MESRS-CNEPRU project.

References

[1] Agrawal R., Faloutsos C., and Swami A., "Efficient Similarity Search in Sequence Databases," *in Proceedings of 4^{th} Information Conference of Foundations of Data Organization and Algorithms*, USA, pp. 69-84, 1993.

[2] Berndt D. and Clifford J., "Using Dynamic Time Warping to Find Patterns in Time Series," *in Proceedings of KDD: AAAI Workshop on Knowledge Discovery in Databases*, Washington, pp. 359-370, 1994.

[3] Boucheham B., "Dimensionality Reduction in Time Series: A PLA-Block Sorting Method," *The International Arab Journal of Information Technology*, vol. 4, no. 4, pp. 307-312, 2007.

[4] Boucheham B., "Matching of Quasi-Periodic Time Series Patterns by Exchange of Block-Sorting Signatures," *Journal of Pattern*

Recognition Letters, vol. 29, no. 4, pp. 501-514, USA, 2008.

[5] Boucheham B., "ShaLTeRR: A Contribution to Short and Long-Term Redundancy Reduction in Digital Signals," *Journal of Signal Processing*, vol. 87, no. 10, pp. 2336-2347, 2007.

[6] Bozkaya T., Yazdani N., and Ozsoyoglu Z., and Glu M., "Matching and Indexing Sequences of Different Lengths," *in Proceedings of The 6th Internaitonal Confernce on Information and Knowledge Management*, USA, pp. 128-135, 1997.

[7] Chen L. and Özsu M., "Similarity Based Retrieval of Time Series Data Using Multi-Scale Histograms," *Technical Report*, School of Computer Science, University of Waterloo, 2003.

[8] Douglas D. and Peucker T., "Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or Its Caricature," *The International Journal for Geographic Information and Geovisualization*, vol. 10, no. 2, pp. 112-122, 1973.

[9] Gardenhire L., "Redundancy Reduction the Key to Adaptive Telemetry," *in Proceedings of National Telemetry Conference*, CA, pp. 1-16, 1964.

[10] Junkui L. and Yuanzhen W., "Early Abandon to Accelerate Exact Dynamic Time Warping," *The International Arab Journal of Information Technology*, vol. 6, no. 2, pp. 144-152, 2009.

[11] Keogh E. and Ratanamahatana C., "Exact Indexing of Dynamic Time Warping," *Journal of Knowledge and Information Systems*, vol. 7, no. 3, pp. 358-386, 2005.

[12] Kruskall J. and Liberman M., "The Symmetric Time Warping Algorithm: From Continuous to Discrete", *in Proceedings of Time Warps, String Edits and Macromolecules*, Stanford, pp. 125-161, 1983.

[13] Shatkay H. and Zdonik S., "Approximate Queries and Representations for Large Data Sequences," *in Proceedings of the 12th International Conference on Data Engineering*, New Orleans, pp. 536-545, 1996.

[14] Tuzcu V. and Nas S., "Dynamic Time Warping As A Novel Tool in Pattern Recognition of ECG Changes in Heart Rhythm Disturbances," *in Proceedings of IEEE International Conference on Systems,* Hawaii, pp. 182-185, 2005.

[15] Wagner R. and Fisher M., "The String to String Correction Problem," *Journal of the ACM*, vol. 21, no. 1, pp. 168-173, 1974.

[16] Zhu Y. and Shasha D., "Warping Indexes With Envelope Transforms for Query by Humming," *in Proceedings of ACM International Conference on Management of Data*, USA, pp. 181-192, 2003.

Bachir Boucheham received the Doctor of Science in 2005 and the HDR Post-Doctoral Degree in 2009 in computer science from the University of Constantine, Algeria. He is currently an associate professor of computer science at the University of Skikda, Algeria. Dr Boucheham is a member of the LRES laboratory, University of Skikda, Algeria. His main research interests include pattern recognition, image and signal processing.