Dimensionality Reduction in Time Series: A PLA-Block-Sorting Method

Bachir Boucheham
Department of Informatics, University of Skikda , Algeria

Abstract: *We address the data reduction in time series problem through a combination of two newly developed algorithms. The first is a modified version of the Douglas-Peucker Algorithm (DPA) for short-term redundancy reduction. The second is an alternative to the classical statistic methods for long-term redundancy reduction and is based on block sorting. The block sorting technique is inspired from the quite recent Burrows and Wheeler Algorithm (BWA). The novel reduction scheme was applied to the ECG time series using the MITBIH public ECG database. Results show that the novel scheme is highly competitive with respect to the most performant existing techniques (SPIHT, TSVD, CCSP-ORD-VLC and others).*

Keywords: *Data reduction, time series, long-term compression,Douglas-Peucker algorithm, block sorting.*

Received February 2, 2006 ; accepted April 16, 2006

1. Introduction

Time series efficient representation is an important pre-processing task for analysis, prediction, data mining of specialized databases, storage and transmission over the network of this type of data. Data reduction in time series is in fact a dimensionality reduction problem. Let $X = (x_1, ... , x_N)$ be the time series vector of measures, with $x_i \in M$, where M is the space of measures. Thus, X belongs to the higher dimension space M^N. It is evidence that, when N is sufficiently large, which is the case for real world time series, efficient processing of this type of data requires some kind of data reduction. The problem of data reduction of set X is then to find another space of representation for the data, say E^k, $k < N$ and a mapping function f: $M^N \to E^k ; X \to Y = f (X)$; such that X can be reconstructed from Y within an admissible error of reconstruction.

The problem of dimensionality reduction in time series has been addressed mainly by transform methods (e. g., DFT [1], DWT [8]), the Truncated Singular Value Decomposition (TSVD) method (e. g., [15]) and the Piecewise Linear Approximation technique (PLA) (e. g., [10]-[11]). The PLA approach is particularly attractive for its simplicity and the availability of many inspiring related algorithms in many computer applications (computer vision, computer graphics, Geographic Information Systems (GIS)...). The PLA technique consists in the selection of a set of Characteristic Points (CPs) from X, based on some wisely predetermined rules. The set of selected CPs, $E^k = (x_j, t_j)$, j = 1..k, stands for the reduced form of representation of X and the mapping in this case is the linear interpolation function between consecutive CPs.

In [6], we have shown that, in a PLA scheme, there exist two main strategies for Short-Term Redundancy Reduction (STRR): The classically used sequential strategy and the recursive strategy, under different norms. Based on the Rate-Distortion (R-D) behavior and the execution time, the study showed that the recursive strategy is two times more efficient than the sequential strategy on the numerical level. However, in the resulting set, STRR only is considered in a PLA scheme. In the general case, the selected set of CPs is still redundant on the long-term, especially in the case of quasi-periodic time series.

Classically, the Long-Term Redundancy (LTR) is reduced by some entropy coding technique (e. g., *Huffman coding*). In this work, we propose a technique for taking into account the Long-Term Redundancy Reduction (LTRR), based on a block-sorting approach. The technique is inspired from the quite recent Burrows and Wheeler Algorithm (BWA) [7], combined to the Douglas-Peucker line simplification Algorithm (DPA) [9].

The remainder of this paper is organized as follows. In section 2, the main materials and methods used in this work are presented. In section 3, application of the developed technique to the ECG signal is illustrated. In section 4, obtained results are discussed. Finally, in section 5, concluding remarks and thoughts are presented.

2. Materials and Methods

2.1. PLA as a Solution to the STRR Problem

Given a discrete curve, formally expressed by the polyline $P = (p_i)$, $i = 1..N$, where $p_i = (x_i, y_i)$, with x_i the horizontal coordinate and y_i the vertical coordinate of p_i, the *PLA* of *P* consists in computing another polyline $Q = (q_j)$, $j = 1..K$, satisfying the following conditions:

1. *K<N.*
2. *q1 = p1 and qK = pN.*
3. *Let ||.,.|| be a distance between P and any PLA of it Q, then ||P, Q|| < ε, with ε > 0, a preset threshold on the tolerance of the approximation error.*

Figure 1 illustrates a polyline $P = (p_1, p_2, ..., p_{100})$, approximated with eight points $Q = (q_1, q_2,..., q_8)$.

As stated in the introduction section, we use a variant of the *DPA* [9]. This algorithm uses a recursive selection strategy, reducing gradually the distance between *P* and *Q* by the maximal possible amount under norm ||.,.|| at each selection. Our choice for the recursive approach is motivated by the excellent performance of this strategy at selection of most perceptually attractive points on the initial curve. By contrast, the classical *PLA* methods use a sequential strategy leading to selection of locally only significant points. The *DPA* main steps are as follows. The initial curve endpoints are first selected $(Q = [p_1, p_N])$. The next selected point, say $q_3 = (x_3, y_3)$, is s. t.:

$$\underset{q_3 \in [p_2 \;\; \cdots \;\; p_{N-1}]}{Max} d(q_3, \hat{q}_3) \tag{1}$$

where $\hat{q}_3$ is the vertical projection of point q_3 on polyline *Q* and *d* is a defined distance. Point q_3 is the most perceptible CP in the interval $]q_1,...,q_2[$. The process is then recursively repeated for the resulting sub-curves $[q_1, q_3]$ and $[q_3, q_2]$ until the condition ||P, Q|| < ε is met. In this study, we reduce the norm given by equation 2:

$$\|P, Q\| = \sqrt{\sum_{i=1}^{i=N} \left(y_i \;-\; \hat{y}_i\right)^2} \tag{2}$$

Then, for segment $[p_i,...,p_j]$ under process, the selected CP, say q_k, is s.t.:

$$q_k = Arg \underset{q_l \in]p_i,\dots,p_j[}{Max} d(q_l, \hat{q}_l) \tag{3}$$

with

$$d(q_k, \hat{q}_k) = |y_k \;-\; \hat{y}_k| \tag{4}$$

In equations (2) and (4), y_i is the magnitude of point q_i and $\hat{y}_i$ that of $\hat{q}_i$. Note that the so computed CPs are selected according to a binary tree of segmentation where the most perceptible points are selected in the upper levels. Figure 2 illustrates this property in the case reported in Figure 1. The approximation algorithm is formally described in Figure 3.

2.2. Block-Sorting: A New Solution to the LTRR Problem

Block sorting is quite a recent trend as fare as compression is concerned. The BWA [7] is one of the first compression algorithms using this technique. The original BWA is a lossless compression method, reported to yield excellent results on images, text and sound [14].

The main idea behind the BWA is computation of a reversible permutation of the original data that creates concentrations. These concentrations of data are successively coded by Run Length Encoding (RLE), MTF (move to front) techniques and finally, Huffman coding is applied.

The decoder proceeds in reverse order, which allows reconstruction of the initial permutation. The permutation in question is the last column of the N x N matrix obtained by cyclic shifting of the initial data, N - 1 times, sorted lexicographically. The reconstruction of the original data from the permutation is achieved through a well-established process.

Our proposition for taking into account the long-term redundancy available in a PLA output is as follows. The output (x_i, y_i), I = 1..K, of a specific STRR method is sorted on the y_i coordinate. Let the output of this step be (x_j', y_j'), j = 1.. K. This last curve is then approximated with the DPA of section 2.1 to reduce LTR. This yields another curve (x_m'', y_m''), m = 1..L, L < K and K time indexes of the STRR CPs. The compression ratio associated with the STRR step (first approximation), as expressed in terms of number of samples reduction is given by equation 5:

$$CR_0 = \frac{N}{2.K} \tag{5}$$

The compression ratio associated with the *LTRR* step (second approximation) is given by equation 6:

$$CR_1 = \frac{N}{K + 2.L} \tag{6}$$

Then, for L < K/2, $CR_1 > CR_0$, hence, gain in compression. Yet, CR_1 is upper bounded by $2CR_0$. Finally, for L < K/2 : $CR_0 < CR_1 < 2CR_0$.

The reconstruction of the original data is conducted as follows. Magnitude $\tilde{y}$ of the K CPs associated with the STRR compression is first computed using the K time indexes of the STRR and the L CPs of the LTRR through linear interpolation between successive CPs, using $(x_j')_{j = 1..K}$ and $(x_m'',y_m'')_{m = 1..L}$.

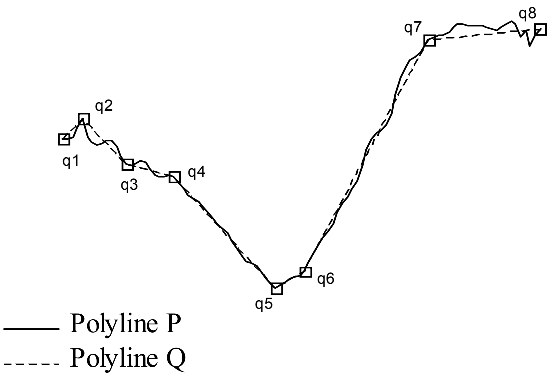

Figure 1. A PLA Q of polyline P.

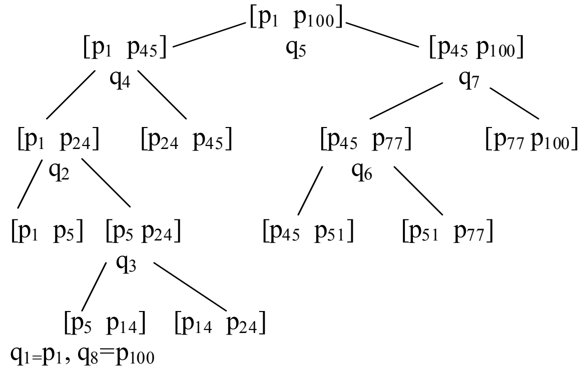

Figure 2. CPs selection binary tree (Figure 1).

```
Step0:  $Q \leftarrow \{p_1, p_N\}$;
        $DPA\ ([p_1, ..., p_N], \varepsilon)$ ;
        $Q \leftarrow Sort\ (Q)$;
        $M \leftarrow |Q|$;
Return (Q, M);
Procedure DPA ([$p_i$,...,$p_j$], $\varepsilon$)          $1 \le i < j \le N$
         If $\| [p_i, p_{i+1}, ..., p_j],[p_i, p_j] \| \ge \varepsilon$ Then
             $q_k \leftarrow Arg \underset{q_l \in ]p_i \cdots p_j[}{Max} d(q_l, \hat{q}_l)$
         $Q \leftarrow Q \cup \{q_k\}$;
         $DPA\ ([p_i, ..., q_k], \varepsilon)$;
         $DPA\ ([q_k, ..., p_j], \varepsilon)$;
         End f;
End
```

Figure 3. Modified Douglas-Peucker algorithm.

The result is a set of tuples $(x_j', \tilde{y}_j)_{j = 1..K}$. Sorting of these tuples on the first coordinate in ascending order yields the approximation of the K CPs of the *STR* compression. Finally, the reconstructed magnitude $\hat{y}$ as an approximation to the original magnitude y is realized by linear interpolation between successive CPs. This step yields the output tuples $(x_i, \hat{y}_i)_{i = 1..N}$.

3. Application to ECG Traces

We apply particularly the proposed method to the Electrocardiogram signal (ECG). The ECG is a biological signal reflecting the heart activity. Samples y_i of this signal represent the difference in potential as measured at the temporal index x_i between two electrodes positioned at specific positions on the body skin. Due to its quasi-periodic nature, a typical ECG signal is composed of a sequence of cardiac cycles. A normal cycle is itself composed of three clinically significant features, in this order: P wave, QRS complex and T wave. It may be interesting to mention that compression of the ECG has been under way during the last four decades.

Our proposed algorithm for LTR reduction is coupled as a post-processing step to the DPA output. The so-enhanced method is denoted herein DPA+. Evaluation of the DPA and DPA+ methods is performed on carefully selected records from the Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) public ECG database. The MIT-BIH database is a collection of 48 records sampled at 360 Hz. Each record is 30 minutes long and each sample is coded on 12 bits. This base serves as a cross-reference for researchers. The evaluation is performed on the numerical level through the compression ratios CR_0 (equation 5) for DPA and CR_1 (equation 6) for DPA+ and computation of the respective distortions upon reconstruction, expressed by the percent root difference PRD_0 for DPA and PRD_1 for DPA+, where PRD_0 is given by equation 7 and PRD_1 by equation 8, with $\bar{y}$ representing the mean original magnitude.

$$PRD_0 = \sqrt{\frac{\sum_{i=1}^{N} \left[y_i - \tilde{y}_i \right]^2}{\sum_{i=1}^{N} \left[y_i - \bar{y} \right]^2}} \text{x} 100\,\% \tag{7}$$

$$PRD_1 = \sqrt{\frac{\sum_{i=1}^{N} \left[y_i - \hat{y}_i \right]^2}{\sum_{i=1}^{N} \left[y_i - \bar{y} \right]^2}} \text{x} 100\,\% \tag{8}$$

The first application is an illustration on a segment of N = 4095 samples (about 11.38 seconds of recording) from the beginning of rec. 103. It is clear that the block size is an important factor for the proposed algorithm, as there is more long-term redundancy, as the block size is larger. But, since the samples are coded on 12 bits and the LTRR compression ratio CR_1 is expressed in terms of number of samples reduction, the largest allowed block size is $2^{12} = 4096$, in order to preserve the integrity of the CR_1 expression. This application is reported in Figure 4, where (original) is the input signal and (reconstructed) is the output signal. This segment was first approximated with $\varepsilon_1 = 5$, yielding K = 459 selected CPs. The compression ratio of the STRR step is then $CR_0 = 4.46{:}1$ with a distortion of $PRD_0 = 2.32\%$. Sorting on magnitude and approximation of the 459 CPs with $\varepsilon_2 = 2.25$ yielded 459 time indexes and L = 36 CPs. Note that, for all experiments, ε_2 is computed as a fraction of ε_1. The overall compression ratio is then CR_l = 7.71:1, for a new distortion of PRD_l = 2.46%. The gain in compression ratio is 73% for a gain in distortion of 6%.

In Figure 5, we report the R-D behavior of *STRR*-*DPA* versus *LTRR*-DPA+ algorithms. This plot shows the substantial R-D behavior enhancement as a result of the LTRR incorporation as a post-processing step for the PLA algorithm of Douglas and Peucker. For instance, for a common distortion of 4%, the compression ratio of the DP algorithm is about 6:1, whereas that of the DPA+ is 10.5:1. Likewise, for a common compression ratio of 10:1, the distortion of the DP algorithm is 9.5%, whereas it is about 3.5% for the DPA+ algorithm.

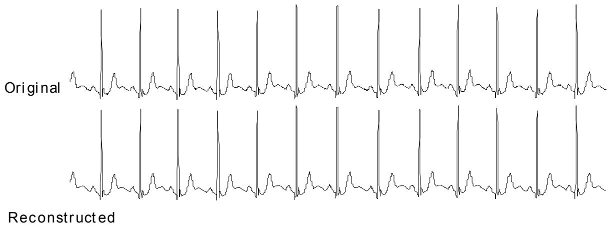

Figure 4. Illustration of the proposed method on a segment from rec. 103.

For the purpose of comparison, we have considered three of the most performant existing methods for ECG data reduction using three different approaches. The first method is a wavelet-SPIHT technique due to Lu *et al.* [12], the second is a TSVD technique due to Wei *et al.* [15] and the third is the Cardinality constrained shortest path, min max Operational Rate Distortion Variable Length Coding (CCSP min max ORD VLC) [13], an optimal PLA reduction scheme combining an optimal R-D coding technique to graph theory, and is due to *Nygaard et al.* Figure 6 shows the R-D behavior for eleven records used by Lu *et al.* The two curves report the mean PRD values at CR equal to 4:1, 5:1 8:1 10:1, 16:1 and 20:1 using 10 minutes recording from the beginning of each record. These plots clearly show that the two methods have comparable behavior with advantage to our method for CR < 10:1 and to the SPIHT method for CR > 10:1. Table 1 reports R-D performance of the DPA+ versus the TSVD method. This table clearly shows also the better performance of the DPA+ over the very sophisticated TSVD method for CR < 15. Table 2 shows the R-D behavior in terms of *bit rate*, expressed in bits per sample (bps) for the DPA+ and the CCSP-ORD-VLC optimal method. This table confirms also the better performance of the DPA+ over the CCSP for bit rates over 1 bps (CR < 10:1).

4. Discussion

Results show the effectiveness of the proposed method to efficiently reduce the short-term and the long-term redundancy as illustrated above. Results show that the novel block-sorting technique achieves higher long-term redundancy reduction for the same distortion than a PLA-STRR scheme. Obtained results show also that the proposed method is highly competitive with respect to the most performant LTRR compression techniques. Results confirm also our claim that the block-sorting technique could be a good alternative to the classically used entropy coding technique for lossy time series reduction schemes. Furthermore, additional compression is still possible with our method through an appropriate coding technique of the saved/sent samples.

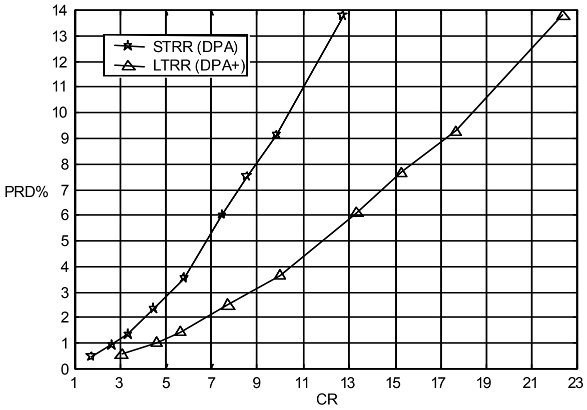

Figure 5. R-D behavior of the STRR (DP algorithm, stars) and DPA+ (STRR-LTRR, triangles) for a block of data from rec. 103.

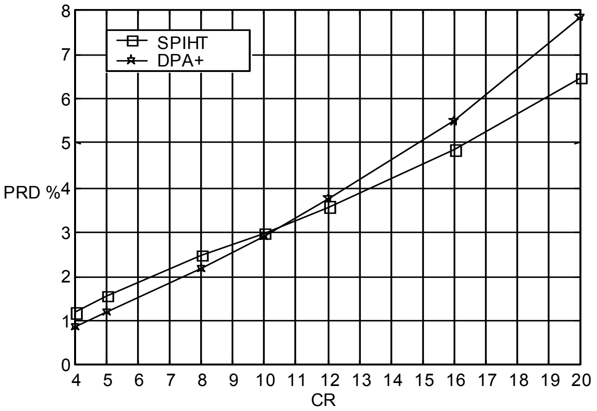

Figure 6. Mean R-D behavior of DPA+ (Stars) and SPIHT (squares) for 10 minutes recording from recs. 100, 101, 102, 103, 107, 109, 115, 117 and 119.

Table 1. DPA+ vs. TSVD: CR-PRD behavior. The condition indicates the estimated interval where the DPA+ or TSVD shows better RD behavior.

Rec.#	DPA+ best for	TSVD best for
100	CR < 10:1	CR > 10:1
101	CR < 15:1	CR > 15:1
109	CR < 10:1	CR > 10:1
111	CR < 15:1	CR > 15:1
117	CR < 8:1	CR > 8:1
119	CR < 12:1	CR > 12:1
213	CR < 23:1	CR > 23:1

Table 2. CCSP-ORD-VLC vs. DPA+ : R-D behavior.

Rec.	ORD-VLC	DPA+
100	Bit rate < 1	Bit rate > 1.3
202	Bit rate < 0.7	Bit rate > 1
207	Bit rate < 0.7	Bit rate > 0.9
214	Bit rate < 0.8	Bit rate > 1.1
203	Bit rate < 1.1	/
203	Bit rate < 1.5	/

5. Conclusion

A novel method for dimensionality reduction in time series for the purpose of analysis, prediction, data mining, storage and transmission over the network of this type of data has been proposed. The novel method is based on a PLA-block sorting combination. Although, the method is quite a general-purpose one-dimensional time series data reduction technique, it is more efficient for quasi-periodic signals. Accordingly, it was implemented as a post-processing step for the DPA method in the specific case of the ECG signal, a highly quasi-periodic time series. Results of the enhanced DPA+ method confirm the substantial improvement of the compression ratio-distortion behavior with respect to that of the DPA method. Results grant the proposed LTRR compression technique a comfortable ranking amongst performant existing methods. We stress the important point that through this work, we proposed an alternate method to the entropy coding technique for long-term redundancy reduction. Note also that we used the Douglas-Peucker algorithm as a unified tool for the resolution of many time series problems in previous works [2-6].

Acknowledgements

The author wishes to thank Professor Mohamed Batouche, Mentouri University of Constantine, Algeria, for interesting discussions and encouragements and Doctor Youcef Ferdi, 20 Août 55 University of Skikda, Algeria, for support and material related to this work.

This work has been partially supported by The Ministry of Higher Education and Scientific Research of the Algerian Government, project #J2101/02/03/05.

References

[1] Agrawal R., Faloutsos C., and Swami A., ‘‘*Efficient Similarity Search in Sequence Ddatabases*,’’ *in Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms*, Chicago, USA, pp. 69-84, 1993.

[2] Boucheham B., Ferdi Y., and Batouche M., “A Characteristic Points Unified Approach to ECG Analysis and Compression,” *in Proceedings of IEEE EMBS Asian Pacific Conference on Biomedical Engineering (APBME)*, Keyhanna, Japan, October 2003.

[3] Boucheham B., Ferdi Y., and Batouche M., “A Dominant Points Based Method for ECG Main Features Detection and Modeling,” *in Proceedings of the International Federation for Medicine & Biology Engineering (IFMBE), EMBEC’02*, Vienna, Austria, vol. 1, pp. 452-453, December 2002.

[4] Boucheham B., Ferdi Y., and Batouche M., “Anchor Points Based Method for QRS Detection in Noisy ECG Records,” *in Proceedings of IFMBE, World Congress on Medical Physics & Biomedical Engineering*, Sydney, Australia, 2003.

[5] Boucheham B., Ferdi Y., and Batouche M., “Piecewise Linear Correction of ECG Baseline Wander: A Curve Simplification Approach,” *Computer Methods and Programs in Biomedicine,* vol. 78, no. 1, pp. 1-10, 2005.

[6] Boucheham B., Ferdi Y., and Batouche M., “Recursive Versus Sequential Multiple Error Measures Reduction: A Curve Simplification Approach to ECG Data Compression,” *Computer Methods and Programs in Biomedicine*, vol. 81, no. 2, pp. 162-173, 2006.

[7] Burrows M. and Wheeler D. J., “A Block Sorting Lossless Compression Algorithm,” *SRC Research Report 124*, Digital Systems Research Center, Palo Alto, CA, May 1994.

[8] Chan K. P. and Fu A. C., “Efficient Time Series Matching by Wavelets,” *in Proceedings of the International Conference on Data Engineering (ICDE)*, Sydney, Australia, pp. 126-133, 1999.

[9] Douglas D. H. and Peucker T. K., “Algorithms for the Reduction of the Number of Points Required to Represent A Digitized Line or its Caricature,” *Canadian Cartographer*, vol. 10, no. 2, pp. 112-122, 1973.

[10] Korn F., Jagadish H. V., and Faloutous C., “Efficient Supporting *Ad Hoc* Queries in Large Databases of Time Series,” *SIGMOD 1997, in Proceedings of ACM SIGMOD, International Conference on Management of Data*, Tucson, Arizona, USA, pp. 289-300, May 1997.

[11] Li C., Yu P., and Castelli V., “MALM: A Framework for Mining Sequence Databases at Multiple Abstraction Levels,” *in Proceedings of the 9thInternational Conference on Information and Knowledge Management,* pp. 267-272, May 1997.

[12] Lu Z., Kim D. Y., and Pearlman W. A., “Wavelet Compression of ECG Signals by the Set Partitioning In Hierarchical Trees Algorithm,” *IEEE Transactions Biomedical Engineering*, vol. 47, pp. 849-856, 2000.

[13] Nygaard R., Melnikov G., and Katsaggelos A. K., “A Rate Distortion Optimal ECG Coding Algorithm,” *IEEE Transactions on Biomedical. Engineering*, vol. 48, no. 1, pp. 28-40, 2001.

[14] Salomon D., *Data Compression: The Complete Reference*, Springer-Verlag, New York, 2000.

[15] Wei J., Chuang C. J., Chou N. K., and Jan G. J., "ECG Data Compression Using Truncated Singular Value Decomposition," *IEEE Transactions on Information Technology in Biomedicine,* vol. 5, no. 4, pp. 290-299, 2001.

Bachir Boucheham received the engineering Bachelor degree in computer science in 1984 from the University of mentouri, Constantine, Algeria, his Master degree in computer science from the University of Minnesota, Minnesota, USA, in 1987, and the PhD degree in computer science in 2005 from the University of Mentouri Constantine, Algeria. In 1987, he worked as research specialist at the University of Minnesota Hospital, School of Dentistry, Department of Orthodontics. Since 1988, he is a senior lecturer at the University of Skikda, Algeria, in the field of computer science. He is a member of the LRES laboratory, University of Skikda, Algeria. His main research interests include pattern recognition, image and signal processing, and compression.