Computer Physics Communications 197 (2015) 118–127

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

Visualizing long vectors of measurements by use of the Hilbert curve

E. Estevez-Rams [a,b,*], C. Perez-Demydenko [b], B. Aragón Fernández [c], R. Lora-Serrano [d]

[a] Facultad de Física, University of Havana (IMRE), San Lazaro y L, CP 10400, La Habana, Cuba
[b] Instituto de Ciencias y Tecnología de Materiales, University of Havana (IMRE), San Lazaro y L, CP 10400, La Habana, Cuba
[c] Universidad de las Ciencias Informáticas (UCI), Carretera a San Antonio, Boyeros, La Habana, Cuba
[d] Universidade Federal de Uberlandia, AV. Joao Naves de Avila 2121- Campus Santa Monica, CEP 38408-144, Minas Gerais, Brazil

ARTICLE INFO

Article history:
Received 20 February 2013
Received in revised form
22 May 2015
Accepted 15 August 2015
Available online 3 September 2015

Keywords:
Hilbert curve
Visualization

ABSTRACT

The use of Hilbert curves to visualize massive vector of data is revisited following previous authors. The Hilbert curve mapping preserves locality and makes meaningful representation of the data. We call such visualization as Hilbert plots. The combination of a Hilbert plot with its Fourier transform allows to identify patterns in the underlying data sequence. The use of different granularity representation also allows to identify periodic intervals within the data. Data from different sources are presented: periodic, aperiodic, logistic map and 1/2-Ising model. A real data example from the study of heartbeat data is also discussed.

1. Introduction

A typical trait in many fields of science is the existence of a large amount of data resulting either from experimental measurements or from some other sources such as computer simulations. An archetypical example is that of a long time series, where the whole data set can be thought as a very large vector. Many combinatorial, statistical, informational theory tools among others have been developed to recognize or discover peculiar features in such data vectors. Common tasks can involve exploring regularities, extracting periodic intervals, asserting "randomness", among others. In spite of the availability of such powerful quantitative tools, the need of displaying the whole set of values in a meaningful way becomes relevant. The visual inspection of the data can help in identifying peculiarities and in discovering structures from where hypothesis can be advanced or decisions can be made, as to which tools are the most convenient to be applied to the actual data series. In a general setting, the one dimensional sequence is to long to make sense writing down all values of the string. The local inspection of the string of values can be useful in some contexts, but it has the serious shortcoming that the overall picture of the data can be missed. In local exploration, although features occurring in short periods can be identified, the logical arrangement of such features over the whole data set can remain hidden to the observer.

Previous authors have used the good locality preserving qualities of Hilbert curve for the purposes of data visualization in general [1] and in the context of DNA analysis [2]. Following such ideas, in this contribution we show the use of the Hilbert curve for displaying data vectors in a meaningful way for synthetic and real life data relevant to physics. We call such representation Hilbert plots and they display the whole of the data vector in a two dimensional (2D) plot that does not hinder the inspection of local intervals if needed. Furthermore, it will be shown that Hilbert plots combined with its Fourier transform, can be powerful enough to carry out some of the uses we have listed above to be expected from data visualization.

If we are seeking for some kind of optimality in visualizing data sets, neighborhood preserving and clustering are two good criteria for a large number of applications. Among the different one-to-one mappings between n-dimensional and one-dimensional (1D) spaces, Hilbert curve has the best neighborhood preserving properties [3–5]. It is a special case of space filling mapping known as Peano curves, that allows to map a 1D sequence into a multidimensional space and vice versa. The use of Hilbert curve has been extensively studied and used in a wide range of applications [6–8,2,9].

The display of one dimensional data series as 2D "images" is not new. A 'logical' approach would be the sequential partition of the data sequence into equal segments of values in order to build a display matrix using shades of gray or any other coloring scheme. Such approach has the serious drawback that far apart values in the original series, will end up close in the two dimensional plot, giving rise to false neighbors. The artificial creation of false

* Corresponding author at: Facultad de Física, University of Havana (IMRE), San Lazaro y L, CP 10400, La Habana, Cuba.
E-mail address: estevez@imre.oc.uh.cu (E. Estevez-Rams).

http://dx.doi.org/10.1016/j.cpc.2015.08.019
0010-4655/

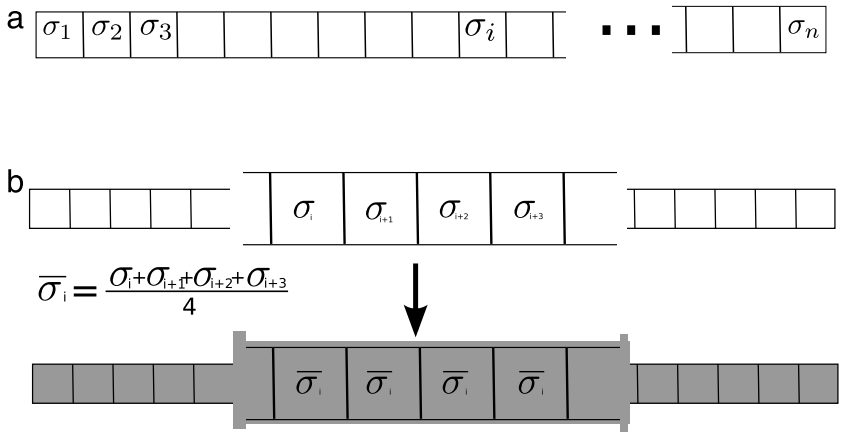

Fig. 1. A data sequence as a vector (a) of values σ_i indexed from left to right. (b) The mean value approximation of 4-granularity is constructed by a sliding window of length 4, where each non-overlapping partition $\sigma_{i;4}$ is substituted by a constant subsequence of the same length and value $\overline{\sigma}_i$.

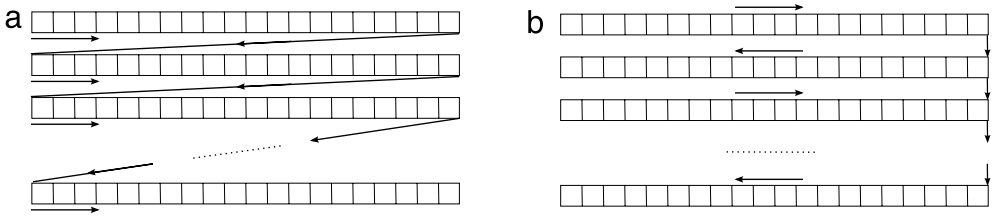

Fig. 2. Mapping of the data vector $\sigma_{1;n}$ to a 2D array. (a) Row scan from left to right with breaks at fixed size. (b) Row scan with interleaved reverse scan directions.

neighborhoods could lead to unfounded conclusions from the visual inspection of the plot. As serious as the former problem, the visualization devised in such a way, does not have nice clustering properties, and a substring of equal values will be displayed as a single line with the same color, hard to distinguish in a large array.

The remainder of the article is organized as follows. In Section 2 some notation and preliminary notions will be introduced. Granularity will be defined, as well as the dilation factor as a measure of locality preservation. The row scan will be described. In Section 3 we define the Hilbert curve construction, and analyze some of its properties. Hilbert plots will then be defined. In Section 4 we discuss the use of Hilbert plots for visualizing pure periodic data and interleaved periodic sequences; aperiodic data as those resulting from fixed morphism (such as Thue–Morse and Fibonacci sequences); data from the binary partition of non-linear maps such as the logistic map; and finally, data from the 1/2-Ising model with first neighborhood interactions. Conclusions follow.

2. Preliminaries

Consider a sequence of values organized in a one dimensional string Σ of length $n = |\Sigma|$. Each value σ_i will be indexed by i, giving its position in the string from left to right (Fig. 1a). The distance between two positions i and j in the string will be the absolute value of the difference between their indexes $|i - j|$. A substring of length l and starting at index i will be denoted by $\sigma_{i;l}$ ($\equiv\sigma_i\sigma_{i+1}\ldots\sigma_{i+l-1}$).

In what follows it will be important to define granularity over a sequence Σ. The l-granularity representation of string Σ, is the mean approximation over substrings of length l. If a partition of the string Σ into non overlapping substrings $\sigma_{i;l}$ is performed, then for each substring a mean value can be calculated:

$$\overline{\sigma_i^l} = \frac{1}{l}\sum_{j=0}^{l-1}\sigma_{i+j}, \tag{1}$$

and the following substitution is made (Fig. 1b)

$$\sigma_{i;l} \longrightarrow \overline{\sigma}_{i;l}, \tag{2}$$

where

$$\overline{\sigma}_{i;l} = \overline{\sigma_i^l\sigma_i^l\sigma_i^l}\ldots\overline{\sigma_i^l} \tag{3}$$

is the mean approximation of the original substring.

1-granularity representation of the sequence is, by construction, equal to the original Σ, which will also be called the faithful representation. For $l > 1$, the l-granularity of a sequence gives a coarser view of Σ. For a periodic sequence of period p, the lp-granularity representation of Σ (l a natural number) will result in a constant vector with components having the mean value over the periodic unit.

The simplest possible mapping of a string to a two dimensional (2D) array is to fix a length $ymax$ and scan the sequence left to right, making a row break at every multiple of $ymax$ as shown in Fig. 2a. The value σ_i will be mapped to the array position given by $(i \bmod ymax, i + \lfloor i/ymax \rfloor)$, where $\lfloor x \rfloor$ is the largest integer less than or equal to x. If two neighboring values σ_i, σ_{i+1} are not split by a row break ($i \bmod ymax \neq 0$), they will be nearest neighbors in the 2D mapping. Yet, from Fig. 2a, it is clearly seen that from the four nearest neighbors of any interior point in the 2D mapping only two of them will actually be consecutive values in the original Σ sequence, while the other two (from the row above and below) can be far away in the Σ sequence. A slight improvement to avoid the loss of locality at the row boundaries can be achieved by the scan shown in Fig. 2b, but the loss of locality from the above and below row remains a problem.

In any scan scheme, some loss of locality is unavoidable. In the 1D sequence each σ_i will have two nearest neighbors, while in the 2D mapping, for any interior point, a coordination of four neighbors is achieved. So at least half of the nearest neighbors of σ_i in the 2D mapping will not correspond to nearest neighbors in the 1D sequence. One will hope that all neighbors will still be close enough to the σ_i in the Σ string, and some meaningful clustering of close values is then achieved in the 2D mapping.