Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Informatik

Clustering Recurrence Plots

Synopsis for a Master's Thesis

Author	Carl Witt
Email	wittcarl@student.hu-berlin.de
Matriculation Number	556313

Proposed Reviewers	Prof. Dr. Ulf Leser, Humboldt-Universität zu Berlin
	Dr. Mike Sips, GeoForschungsZentrum Potsdam

1 Introduction

Dynamic systems are ubiquitous: we find them in astronomy, weather, the financial market and in the human body. Two key observations for investigating such systems are that "(1) similar situations often evolve in a similar way; (2) some situations occur over and over again." [MCRTK07]. Using these properties of complex systems, we can for instance predict (sometimes correctly) that it is going to rain from cloud patterns or weather conditions we experienced before. Similarly, given a time series of observations of a system's parameters, states – or situations – of the system can be reconstructed. By analysing patterns of recurrence, insights about the system can be gained.

Perhaps the most simple way to find recurrent states is to compare each state to each other state. States are usually described as numerical vectors and can be compared e.g. using euclidean distance. This results in a similarity matrix, to which usually a threshold ϵ is applied. A similarity matrix element having a dissimilarity $\leq \epsilon$ is called a recurrence point. The visualization of the matrix is called a recurrence plot, as shown in the top right and the bottom of Figure 1. Typical structures formed by the recurrence points are

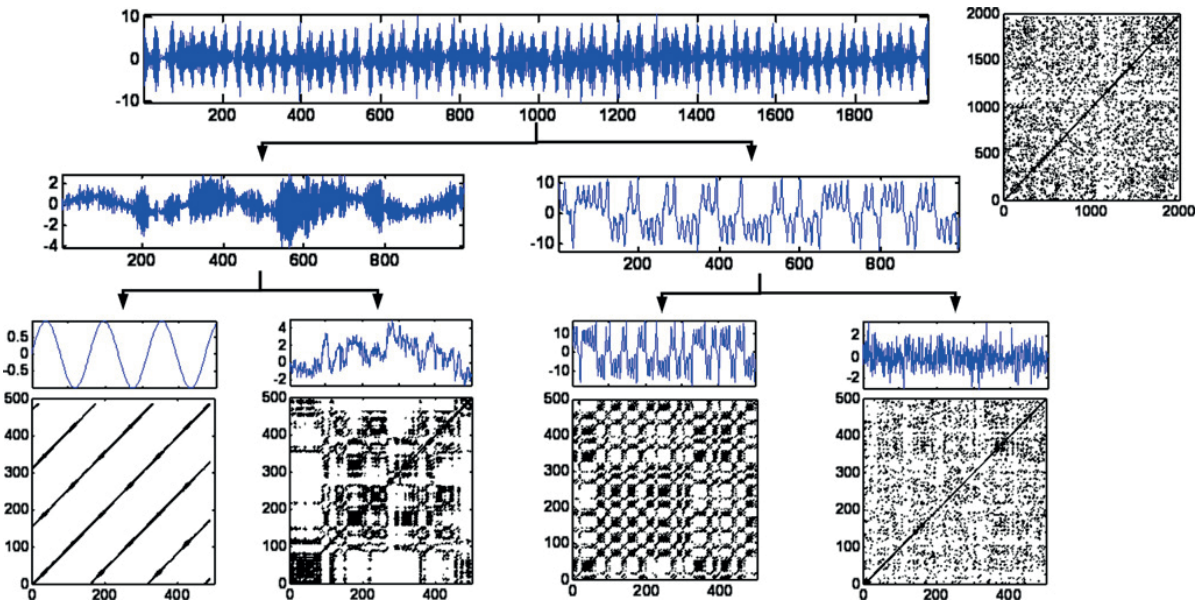

Figure 1: Recurrence Plots generated in a Wavelet Packet Decomposition scenario. Top: The original time series and its recurrence plot. Splitting the time series recursively in low (left) and high (right) frequency components results in the shown binary tree. Bottom: The recurrence plots of the leaf time series. [CY12]

diagonal and vertical lines. The former indicate that the system returns to a previous state and evolves in a similar manner again, the latter occur when the system stagnates. Recurrence quantification analysis (RQA) complements visual inspection with numerical measures that describe properties of the plot. For instance, the average length of the diagonal lines is a measure of how predictable a system is.

In real-world applications, there is often more than one signal present in the measurements of a time series. Wavelet analysis, more specifically wavelet packet decomposition, is one way to separate these signals. The process is sketched in Figure 1: The original time series is recursively decomposed into low and high frequency components, revealing different processes in it. When the resulting signals are still complex, advanced analysis tools like recurrence analysis are needed. The combination of both has been successfully applied to e.g. detect a type of heart disease [CY12]. It also makes recurrence analysis of long time series feasible, because analysing many short time series is far less expensive than analysing a very long time series. What is more, by separating signals, recurrence analysis becomes more informative. For example, in Figure 1, the recurrence plot of the original time series conveys little information. The recurrence plots of the separated signals reflect the underlying processes much better.

The downside is that this multiscale recurrence analysis produces many recurrence plots. Hundreds or thousands of plots might easily be generated, making visual inspection infeasible. To explore the different kinds of processes in the data, the proposed thesis suggests the use of clustering.

2 Related Work

We are not aware of any previous work on clustering recurrence plots. There are, however, three works that propose distance measures for recurrence plots, namely [Bel11] and [SDSIB13, SSB14]. In [Bel11], recurrence plots are represented as bit strings by concatenating their rows. The bzip2 compression rate of the concatenation of the bit strings of two plots is interpreted as their distance. A major drawback of serializing the plots is that the spatial neighbourhood of the recurrence points is compromised. Silva et al. [SDSIB13] use the mpeg-1 video compression algorithm to solve that problem. This compressor accounts for the spatial nature of the data, but in contrast to [Bel11], Silva et al. and Souza et al. do not work with thresholded recurrence plots. This is not desirable, since in recurrence analysis, states are usually either similar or dissimilar. It is not clear whether the techniques applied in [SDSIB13, SSB14] work well on thresholded plots.

By choosing an appropriate representation for a recurrence plot, other similarity measures can be used, as summarized in Table 1. The measures have been selected with regard to two assumptions. First, the representation should preserve the spatial nature of the data, considering the superior performance of [SDSIB13] over [Bel11]. Second, robust similarity measures either require alignment or feature extraction, to compensate distortions (like shifting a plot by a small offset) in the data.

Image representations are closest to the related work and offer a lot of similarity measures. On the other hand, when working with thresholded recurrence plots, the depth of the image is reduced to only two levels, rendering e.g. standard histogram techniques useless. In the domain of dynamic systems analysis, RQA measures are natural quantifications of the recurrence plots image content which also work on thresholded recurrence plots. Forming vectors of RQA measures would be a straightforward approach to reduce the problem to a classical feature vector clustering problem. A measure that

Recurrence Plot Representation	**Distance Measure**	**References**
Image	Compression Distance: bzip2	[Bel11]
	Compression Distance: mpeg-1 (CK-1)	[CK10, SDSIB13]
	Texture Features	[SSB14]
	RQA measures	[CY12]
	Spatiograms: Bhattacharyya distance	[COS07]
Statistical	2D Earth-Mover-Distance	[RTG00]
	Inter-rater agreement (Cohen's Kappa)	[Gwe08]
Graph	Contact-Map-Overlap	[CCI+04]
	Complex Network Measures	[DSD+11]
Time Series	Compression Distance	[SBVA06]
	Dynamic Time Warping	[KP00]

Table 1: Distance measures applicable to data structures that possibly qualify as representations of recurrence plots.

fits the highly discrete nature of the data is inter-rater agreement. Variants of this measure can be used to compare more than two matrices, which could be interesting during the clustering step. Graph representations have the advantage of exploiting not only spatial neighbourhood but a transitive closure of that relationship. In [DSD+11] it has been shown that recurrence plots can be represented as graphs and the according network measures, like the clustering coefficient, convey useful information. Instead of comparing features of the graph, the entire graph can be aligned. This is done in contact map overlap, which is a similarity measure for two thresholded similarity matrices that originates in structural biology. Since this problem is NP-hard, solutions must be found heuristically, e.g. using linear programming [CCI+04].

Finally, if an equivalence between a time series similarity measure and a recurrence plot similarity measure can be shown, then it would be possible to reduce the problem to a time series clustering problem. It is possible to reconstruct a time series from a recurrence plot, which could help in making the time series comparable. Dynamic time warping is a popular choice and has been used for clustering before. However, the problem differs from time series clustering in that recurrence plots explicitly reveal patterns of recurrent behaviour that are not apparent from the raw time series.

There is a vast amount of literature covering clustering and clustering evaluation. Due to the generality of the topic, it is not reviewed here. This thesis will focus on partitional clustering, since it is simple, popular, and has been applied to a wide variety of scenarios. Hierarchical clustering, probably being the other most popular technique, has the downside that comparison of dendrograms is a difficult problem. For comparing partitions, there exist several tried and tested indices, like the adjusted Rand index [HA85] or the more recent variation of information [Mei07].

3 Goal of this Thesis

The goal of this thesis is to evaluate four similarity measures for recurrence plots from Table 1. An applied and empirical approach is proposed, centred around a synthetic data set that captures combinations of prototypical properties of recurrence plots in the analysis of dynamical systems domain. Typical systems in this domain include the Lorenz System, the Rössler System, autoregressive processes or noise.

Additionally, the performance of the similarity measures on real-world data sets shall be evaluated. The *Bern-Barcelona EEG database* [ASR12] contains two classes of time series, those recorded during an epileptic seizure and those recorded during normal brain activity. Andrzejak et al. have developed statistical tests based on recurrence properties that are correlated with the two classes. The outcomes of these tests might be taken as an additional structure on the data.

The *UCR time series data set* [KZH+11] can be used as another data set with known categories. It is used in the two works closest related to the proposed thesis, [SDSIB13, SSB14]. Although these works focus on time series classification, comparing to their results could be a contribution to the current research in this area.

In a first step the chosen similarity measures sshall be compared to each other, based on these data sets. The following questions should be answered:

- Which similarity measures yield similar results?
- How do the similarity measures relate to RQA measures?
- How do the similarity measures behave when comparing noise with non-noise plots?
- How expensive are the different similarity measures? Can they be lower bounded?

Once a basic understanding of the different similarity measures has been developed, the interplay with the clustering algorithm shall be evaluated.

- Is there a clustering tendency for a given similarity measure at all?
- Which similarity measures are able to reveal known clustering structures?
- Do different similarity measures lead to different clusterings?
- How valid are the clusterings, according to internal, external and relative criteria?

The overall objective is to outline needs and solutions for recurrence plot clustering scenarios. More specifically, how different similarity measures capture domain relevant properties of recurrence plots.

References

[ASR12] R G Andrzejak, K Schindler, and C Rummel. Nonrandomness, nonlinear dependence, and nonstationarity of electroencephalographic recordings from epilepsy patients. *Physical Review E*, 86(4), 2012.

[Bel11] Juan P Bello. Measuring Structural Similarity in Music. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2013–2025, September 2011.

[CCI+04] Alberto Caprara, Robert Carr, Sorin Istrail, Giuseppe Lancia, and Brian Walenz. 1001 optimal PDB structure alignments: integer programming methods for finding the maximum contact map overlap. *Journal of Computational Biology*, 11(1):27–52, 2004.

[CK10] Bilson J L Campana and Eamonn J Keogh. A compression-based distance measure for texture. *Statistical Analysis and Data Mining*, 3(6), December 2010.

[COS07] C O Conaire, N E O'Connor, and A F Smeaton. An Improved Spatiogram Similarity Measure for Robust Object Localisation. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007.

[CY12] Yun Chen and Hui Yang. Multiscale recurrence analysis of long-term nonlinear and nonstationary time series. *Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena*, 45(7):978–987, July 2012.

[DSD+11] Reik V Donner, Michael Small, Jonathan F Donges, Norbert Marwan, Yong Zou, Ruoxi Xiang, and Jürgen Kurths. Recurrence-based time series analysis by means of complex network methods. *International Journal of Bifurcation and Chaos*, 21(04):1019–1046, 2011.

[Gwe08] Kilem L Gwet. Intrarater reliability. *Wiley encyclopedia of clinical trials*, 2008.

[HA85] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[KP00] Eamonn J Keogh and Michael J Pazzani. Scaling up dynamic time warping for datamining applications. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Request Permissions, August 2000.

[KZH+11] Eamonn Keogh, Q Zhu, Y Hao, X Xi, L Wei, and C A Ratanamahatana. The UCR Time Series Classification/Clustering Homepage: `www.cs.ucr.edu/~eamonn/time_series_data/`, 2011.

[MCRTK07] Norbert Marwan, M Carmen Romano, Marco Thiel, and Jürgen Kurths. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5):237–329, 2007.

[Mei07] Marina Meilă. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, May 2007.

[RTG00] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2), November 2000.

[SBVA06] C Costa Santos, J Bernardes, P Vitanyi, and L Antunes. Clustering fetal heart rate tracings by compression. *arXiv.org*, December 2006.

[SDSIB13] D F Silva, V M A De Souza, and G E A P A Data Mining ICDM Batista. Time Series Classification Using Compression Distance of Recurrence Plots. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, 2013.

[SSB14] Vinicius M A Souza, Diego F Silva, and Gustavo E A P A Batista. Extracting Texture Features for Time Series Classification. In *2014 22nd International Conference on Pattern Recognition (ICPR)*, pages 1425–1430. IEEE, 2014.