EFDA–JET–CP(07)04/11

J. Vega, A. Murari, G.A. Rattà, P. Castro, A.Pereira, A. Portas
and JET EFDA contributors

Structural Pattern Recognition Techniques for Data Retrieval in Massive Fusion Databases

“This document is intended for publication in the open literature. It is made available on the understanding that it may not be further circulated and extracts or references may not be published prior to publication of the original when applicable, or without the consent of the Publications Officer, EFDA, Culham Science Centre, Abingdon, Oxon, OX14 3DB, UK.”

“Enquiries about Copyright and reproduction should be addressed to the Publications Officer, EFDA, Culham Science Centre, Abingdon, Oxon, OX14 3DB, UK.”

Structural Pattern Recognition Techniques for Data Retrieval in Massive Fusion Databases

J. Vega[1], A. Murari[2], G.A. Rattà[1], P. Castro[1], A. Pereira[1], A. Portas[1]
and JET EFDA contributors*

JET-EFDA, Culham Science Centre, OX14 3DB, Abingdon, UK

[1]Asociación EURATOM/CIEMAT para Fusión. Avda. Complutense, 22. 28040 Madrid, Spain
[2]Consorzio RFX-Associazione EURATOM ENEA per la Fusione. I-35127 Padua, Italy
** See annex of M.L. Watkins et al, "Overview of JET Results ",*
(Proc. 21 st IAEA Fusion Energy Conference, Chengdu, China (2006)).

Preprint of Paper to be submitted for publication in Proceedings of the
Inernational Workshop on Burning Plasma Diagnostics, Villa Monastero,Varenna, Italy.
(24th September 2007 - 28th September 2007)

ABSTRACT.
Diagnostics of present day reactor class fusion experiments, like the Joint European Torus (JET), generate thousands of signals (time series and video images) in each discharge. There is a direct correspondence between the physical phenomena taking place in the plasma and the set of structural shapes (patterns) that they form in the signals: bumps, unexpected amplitude changes, abrupt peaks, periodic components, high intensity zones or specific edge contours. A major difficulty related to data analysis is the identification, in a rapid and automated way, of a set of discharges with comparable behavior, *i.e.* discharges with "similar" patterns. Pattern recognition techniques are efficient tools to search for similar structural forms within the database in a fast an intelligent way. To this end, classification systems must be developed to be used as indexation methods to directly fetch the more similar patterns.

1. INTRODUCTION

So far, data retrieval in fusion databases has been based on a classical model of shot number for input and signal samples for output. However, high level tools to look for data according to scientific and technical criteria are a main challenge for present (JET) and future fusion devices (W7X and ITER). A new paradigm for data access has been proposed not based on shot number but on pattern recognition [1]. The basis for this rests on the fact that diagnostics translate physical properties into patterns with a direct correspondence between the physical behavior of the plasma and the structural shapes that are generated in the signals. The new paradigm means to evolve the input/output data retrieval mechanisms towards a new archetype in which the input is a pattern and the outputs are the shot numbers and the locations (temporal and/or spatial) where the pattern appears.

Pattern recognition is the scientific discipline whose aim is the classification of objects into a number of categories or classes. Key elements to implement pattern recognition systems are the signal representations (feature vectors contain the object attributes of distinctive nature), the classification systems (grouping similar objects into clusters) and the similarity measures (to compare how similar two objects are).

The crucial element to achieve efficiency when searching for similar signals is the classification method. It allows the reduction of the searching space just to the most probable signals of containing a similar pattern. Each cluster represents a reduced searching space in which the objects more likely to be similar can be found.

2. STRUCTURAL PATTERN RECOGNITION IN FUSION

Different pattern recognition systems can be developed for diverse signal collections. A signal collection is the complete set of recorded signals (individual waveforms or images) for all the discharges of interest. Two very general approaches to search for data are considered: the Entire Signal Search (ESS) and the "Pattern In Signal" Search (PISS). An entire signal is a complete image or waveform. ESS allows the searching of similar entire signals within the database. PISS

permits seeking for specific patterns inside signals through the whole database.

Feature extraction processes, classification systems and similarity measures are developed for general purpose data retrieval methods. Given a target signal, the searching process of similar waveforms begins with the classification of the initial signal into one of the clusters of the classification system. Then, the similarity measure is only computed between the signals in the cluster. In this way, the waveforms can be ordered according to the similarity value.

3. ENTIRE SIGNAL SEARCH

3.1 TEMPORAL EVOLUTION SIGNALS

Due to the fact that temporal evolution signals are the dominant waveform type in fusion, attention was focused on this kind of signals.

- Feature extraction
  - Haar wavelet transform: $\mathbf{u} = (u_1, u_2,...u_M) \Leftrightarrow \mathbf{h_u} = (h_1, h_2,...h_N)$, $N << M$
    - Retains most relevant time/frequency information
    - Strong dimensionality reduction
  - Feature vector: $\mathbf{h_u} = (h_1, h_2,...h_N)$
- Classification system: multilayer classification system (fig.1)
- Similarity measure
  - Normalized inner product: $S_{\mathbf{uv}} = |\cos\alpha| = \frac{|\mathbf{h_u} \cdot \mathbf{h_v}|}{\|\mathbf{h_u}\| \cdot \|\mathbf{h_v}\|}$, $0 \dagger S_{\mathbf{uv}} \dagger 1$

A fully description of the ESS approach to the TJ-II stellarator database can be found in [2]. Applications to the JET database were shown in [3].

3.2. IMAGES

- Feature extraction
  - Thresholding + 2D Haar wavelet transform
    - Thresholding: intensity high-pass filter
      - $\mathbf{u} = (u_{ij})$, $i = 1,...M$, $j = 1,...M$ ($u_{ij} = 0$ *if* ($u_{ij} < threshold$)
  - 2D Haar: $\mathbf{U} = (u_{ij}) \Leftrightarrow \mathbf{H_U} = (h_{ij})$, $i = 1,...N$, $j = 1,...N$, $N < M$
  - Feature vector: $\mathbf{H_U} = (h_{ij})$, $i = 1,...N$, $j = 1,...N$
- Classification system
  - 2^{NxN} clusters defined by pixels with simultaneously intensity > 0
- Similarity measure
  - Euclidean distance: $d(\mathbf{U}, \mathbf{V}) = \|\mathbf{H_U} - \mathbf{H_V}\|_{Euclidean}$

Results with this technique were given in [4].

4. "PATTERN IN SIGNAL" SEARCH

4.1 TEMPORAL EVOLUTION SIGNALS

- Feature extraction

- Haar wavelet transform + delta calculation + primitive computation
  - Haar: $\mathbf{u} = (u_1, u_2,...u_M) \Leftrightarrow \mathbf{h_u} = (h_1, h_2,...h_N)$, $N << M$
  - Delta calculation: $\delta_i = h_{i+1} - h_i, i = 1,...N-1$
  - Primitive: label assignment according to (δ_i/Δ) (fig.2)
- Feature vector: string of characters **v** = *"eeabedd"*

- Classification system
  - Relational database
    - Recognition problem ⇒ pattern-matching problem
- Similarity measure
  - Euclidean distance: $d(\mathbf{v}, \mathbf{w}) = \sqrt{\sum_{i=1}^{patternlength} (\delta_{vi} - \delta_{wi})^2}$

Similar techniques not based on the delta transformation but on the slope of a linear fitting of segmented waveforms (both equal length segments [5] and variable length segments [3]) were previously developed.

4.2 IMAGES

- Feature extraction
  - Thresholding + 2D Haar wavelet transform + primitive computation
    - Thresholding: intensity high-pass filter
      $\mathbf{u} = (u_{ij})$, $i = 1,...M$, $j = 1,...M$ ($u_{ij} = 0$ *if* ($u_{ij} < threshold$)
    - 2D Haar: $\mathbf{U} = (u_{ij}) \Leftrightarrow \mathbf{H_U} = (h_{ij})$, $i = 1,...N$, $j = 1,...N$, $N < M$
    - Primitive: label assignment according to $(\delta_i/\Delta h)$ (fig.3)
  - Feature vector: 2D string of characters
- Classification system
  - Relational database
    - Recognition problem ⇒ pattern-matching problem
- Similarity measure
  - Euclidean distance: $d(\mathbf{U}, \mathbf{W}) = \left[\|\mathbf{H_U} - \mathbf{H_W}\|_{Euclidean} \right]_{pattern\ row\ columns}$

This technique is going to be applied to visible fast cameras in JET.

REFERENCES

[1]. J. Vega and JET EFDA Contributors. "Intelligent methods for data retrieval in fusion databases". 6th IAEA Technical Meeting on Control, Data Acquisition and Remote Participation for Fusion Research. 4-8 June 2007. Inuyama (Japan) (http://tm2007.nifs.ac.jp/). (Sent to Fusion Engineering and Design).

[2]. J. Vega, A. Pereira, A. Portas, S. Dormido-Canto, G. Farias, R. Dormido et al. "Data mining technique for fast retrieval of similar waveforms in Fusion massive databases". Sent to Fusion Engineering and Design.

[3]. G.A. Ratt·, J. Vega, A. Pereira, A. Portas, E. De la Luna, S. Dormido-Canto et al. "First applications

of structural pattern recognition methods to the investigation of specific physical phenomena at JET". 6th IAEA Technical Meeting on Control, Data Acquisition and Remote Participation for Fusion Research. 4-8 June 2007. Inuyama (Japan) (http://tm2007.nifs.ac.jp/). (Sent to Fusion Engineering and Design).

[4]. D. Raju, J. Vega, P. Castro, G.A. Rattà, A. Murari, G. Vagliasindi and JET EFDA Contributors. "Structural pattern recognition for image processing in fusion plasmas". 6th IAEA Technical Meeting on Control, Data Acquisition and Remote Participation for Fusion Research. 4-8 June 2007. Inuyama (Japan) (http://tm2007.nifs.ac.jp/).

[5]. S. Dormido-Canto, G. Farias, J. Vega, R. Dormido, J. Sánchez, M. Santos et al. "Search and retrieval of plasma waveforms: structural pattern recognition approach". Rev. Sci. Ins. **77** (2006) 10F514.

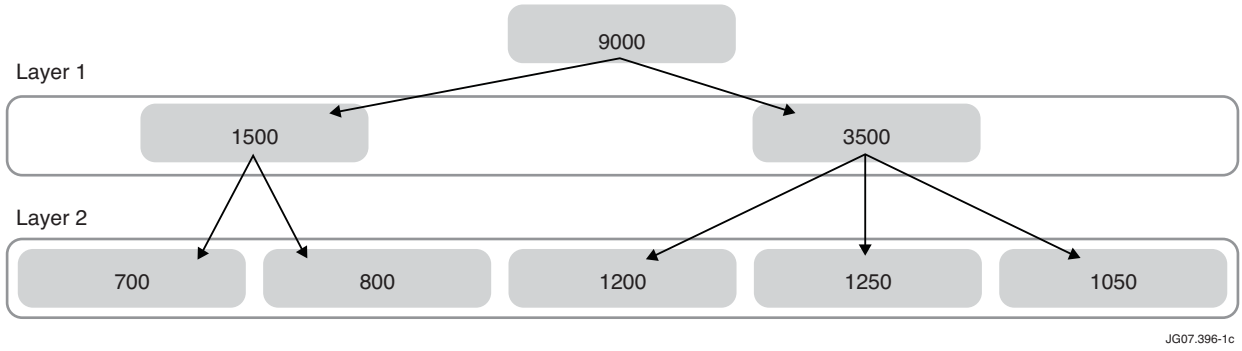

Figure 1: Different clustering criteria can be defined between layers. Clusters are split into smaller ones through the different layers. Numbers represent the number of signals in the clusters. Properties of the multi-tier scheme are described in [2].

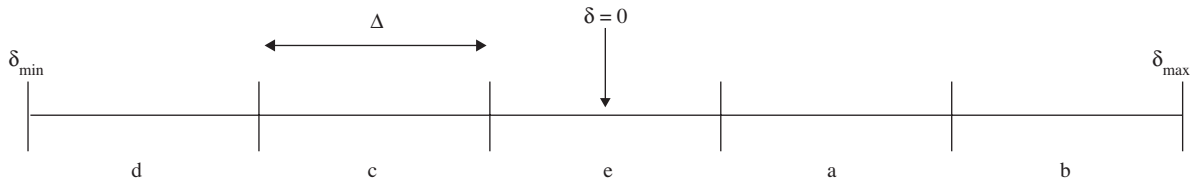

Figure 2: Labels (five values in this example: d, c, e, a, b) are assigned depending on the delta value between consecutive points of the Haar transform.

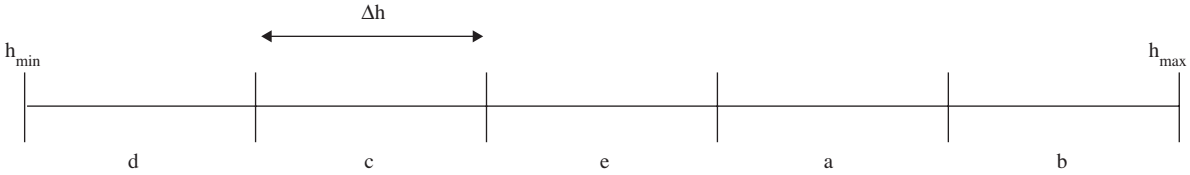

Figure 3: Labels (five values in this example: d, c, e, a, b) in the pixels are assigned depending on the values of the Haar transform.