EFDA–JET–CP(07)05/03

J. Vega, G. Rattà, A. Murari, P. Castro, S. Dormido-Canto, R. Dormido, G. Farias, A. Pereira, A. Portas, E. de la Luna, I. Pastor, J. Sánchez, R. Castro, M. Santos, H. Vargas and JET EFDA contributors

Recent Results on Structural Pattern Recognition for Fusion Massive 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 It is made available on the understanding that it may not be further circulated and extracts or references may not be published prior to publication of the original when applicable, or without the consent of the Publications Officer, EFDA, Culham Science Centre, Abingdon, Oxon,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."

Recent Results on Structural Pattern Recognition for Fusion Massive Databases

J. Vega[1], G. Rattà[1], A. Murari[2], P. Castro[1], S. Dormido-Canto[3], R. Dormido[3], G. Farias[3], A. Pereira[1], A. Portas[1], E. de la Luna[1], I. Pastor[1], J. Sánchez[3], R. Castro[1], M. Santos[4], H. Vargas[3]and JET EFDA contributors*

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

[1]AsociaciÛn EURATOM/CIEMAT para FusiÛn. Avda. Complutense, 22. 28040 Madrid, Spain
[2]Consorzio RFX-Associazione EURATOM ENEA per la Fusione. I-35127 Padua, Italy
[3]Dpto. de Inform·tica y Autom·tica. Universidad Nacional de EducaciÛn a Distancia. Madrid, Spain
[4]Dpto. de Arquitectura de Computadores y Autom·tica. Universidad Complutense. Madrid, Spain
** 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
IEEE International Symposium on Intelligent Signal Processing,
Alcala de Henares, Madrid, Spain
(3rd October 2007 - 5th October 2007)

ABSTRACT

Physics studies in fusion devices require statistical analyses of a large number of discharges. Given the complexity of the plasma and the non-linear interactions between the relevant parameters, connecting a physical phenomenon with the signal patterns that it generates can be quite demanding. Up to now, data retrieval has been typically accomplished by means of signal name and shot number. The search of the temporal segment to analyze has been carried out in a manual way. Manual searches in databases must be replaced by intelligent techniques to look for data in an automated way. Structural pattern recognition techniques have proven to be very efficient methods to index and retrieve data in JET and TJ-II databases. Waveforms and images can be accessed through several structural pattern recognition applications.

1. INTRODUCTION

Some plasma behaviors, as a result of unexpected events and instabilities, only become apparent in an intermittent way. This fact can complicate the interpretation of their physical nature and their potential effects on the plasma confinement. The starting point to analyze these phenomena is to find a number of occurrences high enough to formulate hypotheses with a sufficient statistical basis. The search of events is carried out in a manual way by means of visual data analysis. Visual inspection of signals allows the recognition of certain patterns that can be used to identify the presence of non-standard behaviors. The aim of this searching process is to determine both the shot number and the time interval where the patterns appear.

Nowadays, data retrieval methods can no longer be based on manual searches according to signal name and shot number. First, the pulse length of the experiments is increasing significantly. The longer the pulse, the more tedious is the manual pattern search. Second, the rapid increasing of imaging diagnostics should be considered. For example, fast cameras may acquire images with a rate of hundreds of frames per second and, therefore, the manual selection of a representative image for a particular event becomes a cumbersome procedure. Third, it should be noted that very large databases, with millions of signals (for example, TJ-II is a medium size device that acquires 500 signals/discharge and has stored about 18000 discharges) and Tbytes of data, have to be analyzed. For instance, JET may produce over 10 Gbytes of data per shot, although the typical rate is 5 Gbytes/discharge.

New models for data retrieval have to take advantage of the fact that fusion diagnostics produce similar signals for reproducible behaviors. This means that diagnostics translate physical properties into patterns with a correspondence between the plasma physical properties and the structural shapes that are generated in the signals. Therefore, this direct link allows the introduction of a new paradigm for data access. Instead of using the shot number as input parameter and the signal samples as output data, a more practical criterion would be to ask for a pattern and to receive the pulse numbers and the locations (time instant and/or spatial position) where the pattern appears.

A first approach for pattern oriented data retrieval is the use of the structural forms of the signals. The presence of characteristic patterns in waveforms (bumps, unexpected amplitude changes or

abrupt peaks) and images (high intensity zones or specific edge contours) convert structural pattern recognition techniques in optimal methods to attain an automated and efficient data access. Two very generic approaches, based on structural pattern recognition techniques, have been developed for general purpose data retrieval in fusion. First, the Entire Signal (ES) approach allows searching for similar images or waveforms [1] from a given one. An entire signal is a complete image or waveform. Complete waveforms are defined for the same temporal interval from a particular event. Examples are: a 40s interval from the plasma start, a 10s segment from the beginning of the neutral beam injection or 5s after an L-H transition. Secondly, another structural approach has been developed [2]: patterns in signals (PIS). PIS allows seeking for specific patterns within signals.
This article summarizes both techniques and shows specific applications to time-series data and images in databases of two different fusion devices: the TJ-II stellarator and the JET tokamak. TJ-II is a medium sized stellarator (heliac type) [3] located at CIEMAT, Madrid (Spain). JET [4] is the biggest fusion device in the world and it is located in Culham (UK).

Section II of the article introduces the notion of signal collections to put together comparable data. Section III summarizes the main concepts to consider in a general pattern recognition problem. Section IV is devoted to describing the application of structural pattern recognition techniques to Fusion databases. Sections V and VI review respectively the ES and PIS approaches. Section VII is a discussion.

2. SIGNAL COLLECTIONS

A signal is any kind of data that describes a particular measurement during a discharge and contains some information. Depending on the specific representation of the data, signals can be of several types. Firstly, we have bi-dimensional data. The samples are defined by ordered pairs (x, y). Temporal evolution signals are a particular case of bi-dimensional data where one of the coordinates is time (fig.1(a)). Secondly, contour maps are often encountered. They are 3-dimensional representations with two spatial coordinates and the corresponding amplitude (fig.1(b)). Thirdly, images are becoming every day more frequent. Each pixel is described by two spatial coordinates, a color intensity and a forth dimension to distinguish the red, green and blue color components (fig. 1(c)). Finally, it should be mentioned that as personal computers are providing more capabilities on computing and storage, video movies are becoming very popular diagnostic signals, and very promising results are being obtained with infrared and visible cameras in JET [5].

Signals are grouped into collections for pattern oriented data retrieval. A signal collection is the complete set of recorded signals for all the discharges of interest. As a first example, the plasma current collection is made up of all temporal evolution signals that provide the plasma current in a Tokamak. A second example of collection could be the set of movies of an infrared video camera. Generally speaking, any data representation to describe a particular measurement (usually on a shot to shot basis) is a collection. The existence of collections obeys to the need of grouping comparable data to make pattern searches easier through equivalent data representations.

3. MAIN CONCEPTS IN PATTERN RECOGNITION

Due to the fact that the new model for data retrieval should be pattern oriented, a straightforward approach would be the use of pattern recognition techniques [6] for data access. Pattern recognition is the scientific discipline dealing with methods for object description and classification. Therefore, two main concepts arise: object description and classification systems. A fundamental third concept in pattern recognition, independent of whatever approach we may follow, is the notion of similarity. Two objects are recognized as similar because they have valued common attributes.

It should be noted that a high dimensionality is an issue in pattern recognition techniques because the computational effort increases with the dimensionality of the problem. In Fusion massive databases, the dimensionality depends not only on the number of signals (millions of waveforms, images and video movies) but also on their size (millions of samples per waveform and hundreds of millions of pixels per movie).

3.1. OBJECT DESCRIPTION

Object description is the process by means of which a proper representation of the objects is achieved for classification purposes. The process consists of extracting features or attributes that are of distinctive nature. After feature extraction, objects are always represented by the corresponding feature vectors. The process has a double functionality. On the one hand, it translates characteristics of the objects into attributes that can be managed by a computer system. On the other hand, feature extraction is used to reduce the dimensionality of the problem as much as possible.

3.2. CLASSIFICATION SYSTEM

Classification systems are used to index the objects according to some criteria. This means the creation of different clusters (classes) to show the grouping in the data. Creating classifiers is a learning problem. Learning refers to some form of algorithm to assign each object to a cluster. There are two common types of learning problems, known as supervised learning and unsupervised learning. Supervised learning is used to estimate an unknown (input, output) mapping from known (input, output) samples. The term ‘supervised’ denotes the fact that output values for training samples are known. In the unsupervised learning scheme, only input samples are given to a learning system, and there is no notion of the output during the learning [7, 8].

3.3. SIMILARITY MEASURE

This concept is necessary to compare how similar two objects are. It requires the introduction of a distance between the features or attributes of the objects (in the mathematical sense) to be used as a proximity measure.

4. STRUCTURAL PATTERN RECOGNITION

Recorded signals from diagnostics are used to analyze the plasma physical properties. Such properties

can be identified by the presence of associated patterns (structural shapes) in the data. The recognition of structural shapes plays a central role to distinguish particular behaviors. Making use of this fact, computational methods can be developed to update the classical model of data retrieval with a new one based on searching for data according to physical criteria. The traditional method, based on asking for shot number and returning the signal samples, does not provide pattern locations. Patterns inside the signals have to be found by means of data inspection. A more powerful paradigm for data retrieval is founded on asking for patterns inside signal collections and obtaining the discharge numbers and the pattern location within the signals.

Taking into account the high dimensionality of Fusion databases, the main challenge to put into operation the new paradigm can be summarized with a single word: efficiency. In this context, efficiency means not to traverse the entire database when a specific pattern is searched, but to develop intelligent mechanisms to reduce the searching space just to the most probable signals of containing a similar pattern.

The crucial element to achieve efficiency in this structural pattern recognition problem is the classification system. In a linear approximation, looking for similar structural forms inside a signal collection means to compute the similarity measure between all pairs of feature vectors. However, the classification system allows the indexation of the feature vectors in such a way that clusters group similar objects. Therefore, each cluster represents a reduced searching space in which the more likely objects to be similar can be found.

To look for structural patterns inside a signal collection, some previous steps are required. Firstly, features to describe the signals must be chosen. Secondly, clustering criteria to group the data into convenient clusters have to be defined. Thirdly, a similarity measure is needed to be able to compare how similar (or dissimilar) two feature vectors are.

After the creation of the indexation system, the search of patterns can be carried out. Given a target pattern, the searching process of similar patterns is accomplished in three steps: feature extraction, feature vector classification and similarity factor computation. The former is essential to classify the target pattern into one of the existing clusters and, hence, the target pattern is grouped together with the more similar ones. The similarity measure is only computed between the target feature vector and the feature vectors of the cluster and, therefore, the search method avoids traversing the whole database.

As it was mentioned above, two different approaches were developed for data retrieval with structural pattern recognition techniques: the ES and the PIS methods. Table I summarizes applications of both techniques to several signal collections (waveforms and Thomson Scattering (TS) images) of different databases (JET and TJ-II).

5. ENTIRE SIGNAL APPROACH

The applications of the ES technique to JET and TJ-II waveforms use the Haar wavelet transform [9] as feature extractor. This selection allows a strong reduction of the dimensionality and it retains the waveform time and frequency information.

The indexation system is based on a multi-layer classification system whose clustering criteria may evolve in a flexible and dynamical way. Individual clusters can be split at any moment to reach an optimal classification. At present, two layers have been considered. The first one divides the collection into clusters that group shots with the same pulse length. Each first layer cluster can be split into several ones according to a structural shape criterion (fig.2). The figure shows how the cluster refinement produces groups with lesser number of signals. In this case, the grouping was carried out according to similarity values.

The similarity factor is the normalized inner product (NIP). Actually, the absolute value of this quantity has been chosen to measure the similarity of two feature vectors $\mathbf{u}_w$ and $\mathbf{v}_w$.

$$S_{\mathbf{uv}} = |\cos\alpha| = \frac{|\mathbf{u}_w \cdot \mathbf{v}_w|}{||\mathbf{u}_w|| \cdot ||\mathbf{v}_w||}, \; 0 \dagger S_{\mathbf{uv}} \dagger 1 \qquad (1)$$

This definition was taken by several reasons. First, the geometrical interpretation of the dot product is straightforward: two unitary vectors are equal if the inner product is 1. If the value is 0, both vectors are orthogonal and no similarity exists. Second, this NIP is independent of amplification factors. Signals differing exclusively in a gain factor are recognized as equal waveforms (similarity 1). Third, the NIP is also independent of signal polarities.

Figure 3 shows the application of the ES technique to a bolometry signal collection of the TJ-II database. The waveforms are raw data not calibrated in an absolute way. It should be emphasized that signals differing in gain and polarity are recognized as similar signals.

Java applications for data retrieval are accessible for concurrent execution from the TJ-II remote participation system (for TJ-II databases) and the JET data access environment (for JET databases).

Computation times to complete a searching process were measured in two different computer environments with a single layer classification system based on discharge length. A first environment with 128 clusters was created with the Matlab software package on a Windows XP Pentium IV computer. The searching time is the sum of three different times: feature extraction of the target waveform, feature vector classification into one of the clusters of the classification system and NIP computation of the target feature vector with all feature vectors in the previous cluster. The first two times are negligible in comparison with the third one. A mean value can be established as 18 ms per signal present in the second step cluster. A second environment with 256 clusters of waveforms was tested at the JAC Linux Cluster at JET (a high performance cluster of 181 Athlon processor cores). The searching time was 1 ms/waveform in the cluster.

Concerning the additional storage requirement for the classification system, the worst case needed an extra 17% of the space filled by the signal collection.

The ES approach has been also applied to images. In particular, the collection is made up of the images from the TJ-II Thomson Scattering CCD camera. One image is collected per discharge and five different kinds of images are possible, depending of the type of measurement: CCD camera background, stray light, electron cyclotron heating (ECH) phase, neutral beam injection (NBI)

phase and cut off density (fig.4).

Feature extraction is accomplished by means of a Haar two-dimensional wavelet transform [9]. The classification system is based on a supervised clustering method with five classes (one per possible measurement). The similarity measure is computed with the Euclidean distance between feature vectors. In this particular case, the Euclidean distance provides a better discrimination than the NIP because it is not necessary to reach the 6^{th} or 7^{th} decimal in the similarity factor to distinguish images.

To search similar signals to a target one, the procedure performs feature extraction and the classification into one of the five clusters. It is attained by means of a linear discriminant function based on Support Vector Machines (SVM) in a one-versus-the-rest approximation. Similarities are computed with the feature vectors of the cluster.

6. PATTERNS IN SIGNALS APPROACH

This approach allows the search of patterns within time-series data. This is a big challenge in data retrieval taking into account the very large volume of Fusion databases.

Patterns can be considered as composed of simpler sub-patterns. The most elementary ones are known as primitives. Primitives are represented by characters, converting the pattern recognition problem into a pattern-matching problem.

The description of objects in this kind of pattern recognition systems is difficult to implement because there is no general solution for extracting structural features (primitives) from data. Primitive extractors can be developed to extract either the simplest and most generic primitives or the domain specific primitives that best support the subsequent searching task. The former are domain independent and the knowledge content is reduced to a minimum. The latter requires strong domain knowledge and this can be an issue for the wide application of the technique. Therefore, to solve general purpose needs, it is better to use domain independent feature extractors.

Bearing in mind that the feature extraction tries to reduce the problem dimensionality, any signal can be divided into segments of equal temporal length and each segment is fitted with a straight line through a least squares minimization process (fig. 5). Then, segments are encoded according to a discrete set of values (code alphabet). The definition of code alphabets enables the description of time-series data as strings, instead of representing the signals in terms of multidimensional data vectors. The labels of the segments are based on the slope of the straight lines (fig.5). For this reason, this method is called “the slope method”.

Due to the fact that waveforms are represented by strings, searching for patterns means looking for characters. Therefore, database technologies must help in the development of the classification system. One particular database model offers a unique combination of power, flexibility and universal acceptance: the relational model [10]. In addition, the relational model provides enough flexibility to retrieve combinations of data. For example, instead of searching for an exact match of slopes, it is easy to include in the query the search of adjacent slopes or even, the search of just the inverse polarity sequence (fig. 5). It should be mentioned that a relational database cannot be seen as a

clustering system in the most pure sense, but it is a very efficient indexing system to retrieve data. The application of this technique to the TJ-II database was developed with the Microsoft-Access relational database. It is discussed in [2] and a variant of this method was developed for JET databases. Looking for reducing the number of primitives to represent a signal, segments of variable temporal length were considered (fig.6). This length is defined by the number of samples to fit the signal with a straight line (least squares minimization), but maintaining the fit error lesser than a certain factor, F, depending on the waveform standard deviation, σ:

$$F = K \cdot \sigma \, , \; K = \text{constant} \tag{2}$$

Each new segment starts with the fit of three points to a straight line and samples are added (one by one) while the fit error is smaller than F. The temporal length (Δt) and the amplitude difference (ΔA) between the ends of each segment (fig.6) are stored to compute the similarity in pattern comparisons.

When selecting a pattern in a signal, (for instance a pattern made up of m characters '$C_1C_2\ldots C_m$'), the searching process queries to the relational database for this string and it returns all records containing '$C_1C_2\ldots C_m$'. At this point, it is necessary to sort the results by means of a similarity measure between the target pattern and the returned data.

The similarity factor is defined through the mean value of NIPs over all the segments that form a pattern, where the NIPs are computed with the ordered pairs (Δt, ΔA) of each segment of the signals to compare.

$$\begin{aligned} S &= \frac{1}{m} \sum_{i=1}^{m} S_{u(i)v(i)}, \; m = \#\text{segments in the pattern} \\ u(i) &= (\Delta t_i, \Delta A_i)_{target\ pattern} \text{ and } v(i) = (\Delta t_i, \Delta A_i)_{retrieved\ data} \end{aligned} \tag{3}$$

With this definition the similarity is a real number between 0 (no similarity at all) and 1 (equal signals).

The slope method with variable temporal length segments is accessible in a concurrent way for multiple users from the JAC Linux Cluster of JET. It uses the PostgreSQL relational database (http://www.postgresql.org). A searching example is shown in figure 7. At the top, the target pattern appears. It is found with similarity 1 and similar patterns can be seen inside the other waveforms.

Computation time for data searching depends on the pattern to search and also on the flexibility level required in the query. Typical times are seconds. Additional storage requirement for the classification system is, in general, a small fraction of the space needed for the signal collection.

DISCUSSION

Structural pattern recognition techniques are an efficient way to implement a pattern oriented data retrieval paradigm.

The smoothing level to extract signal characteristics in the feature extraction process is related (in a direct way) to the degree of dimensionality reduction accomplished in the process. Therefore, fast events (like ELMS or MHD modes) require low smoothing levels. The cost for this is not to achieve high dimensionality reductions. As a consequence, higher additional storage for the classification system will be needed.

There is not a single criterion to develop classification systems. However, care should be taken to avoid the creation of clusters with only one or two shots.

ACKNOWLEDGMENTS

The authors wish to thank Prof. Sebastián Dormido Bencomo (UNED) and Prof. Jesús Manuel de la Cruz (UCM) for their constructive comments and help.

REFERENCES

[1] J. Vega et al. “Data mining technique for fast retrieval of similar waveforms in Fusion massive databases”. Sent to Fusion Engineering and Design.

[2] S. Dormido-Canto et al. Rev. Sci. Ins. **77** (2006) 10F514.

[3] C. Alejaldre et al. Plasma Phys. Controlled Fusion **41**, 1 (1999), A539.

[4] J. Pamela et al. “The JET Programme in Support of ITER.” SOFT Conference 2006, To be published in Fusion Engineering and Design.

[5] G. Vagliasindi et al. Application of Cellular Neural Network Methods to Real Time Image Analysis in Plasma Fusion.

[6] S. Theodoridis, K. Koutroumbas. *Pattern Recognition*, (2nd edition). Academic Press. 2003.

[7] R.O. Duda, P. E. Hart, D. G. Stork. *Pattern classification*, (2nd edition). John Wiley & Sons, INC. 2001.

[8] V. Cherkassky, F. Mulier. *Learning from data*. John Wiley & Sons, INC. 1998.

[9] Y. Nievergelt. *Wavelets made easy*. Birkhäuser. 1999.

[10] C.J. Date, H. Darwen. *Databases, Types and the Relational Model (3rd edition)*. Addison-Wesley. 2006.

Collection	Method	Feature extraction	Classification	Similarity measure
JET waveforms	ES	Haar wavelet	Multi-layer system	NIP
TJ-II waveforms	ES	Haar wavelet	Multi-layer system	NIP
TJ-II TS images	ES	2D Haar wavelet	Single-layer system	Euclidean distance
JET waveforms	PIS	Slopes	Relational database	NIP
TJ-II waveforms	PIS	Slopes	Relational database	Fitted error

Table 1. Summary of structural pattern recognition applications

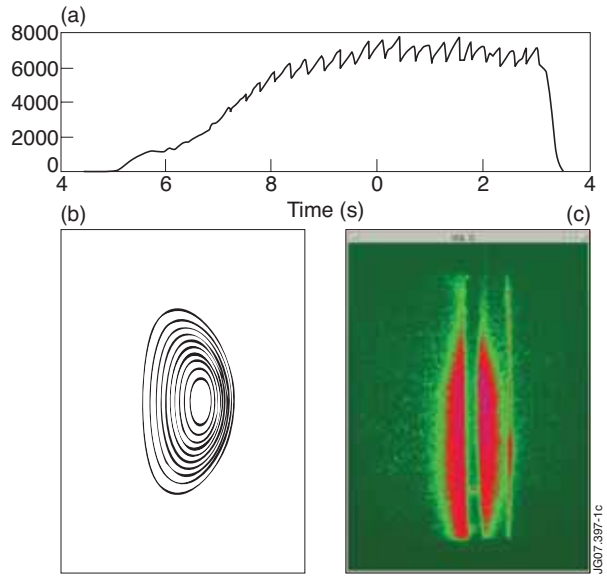

Figure 1: Examples of time-series data (a), contours (b), and images (c)

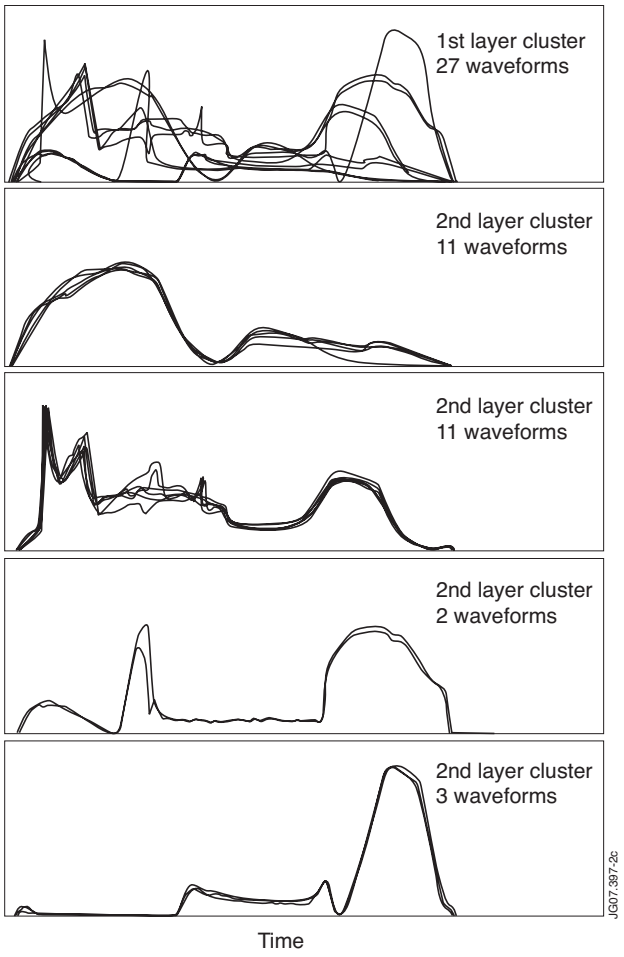

Figure 2: Example of a cluster splitting with Electron Cyclotron Emission (ECE) waveforms from the JET database

Similarity	Shot
1.00000	13774
0.99931	13775
0.99778	13770
0.99774	13764
0.99711	12881
0.99676	13760
0.99664	13773
0.99617	13777
0.99577	13761
0.99563	12957

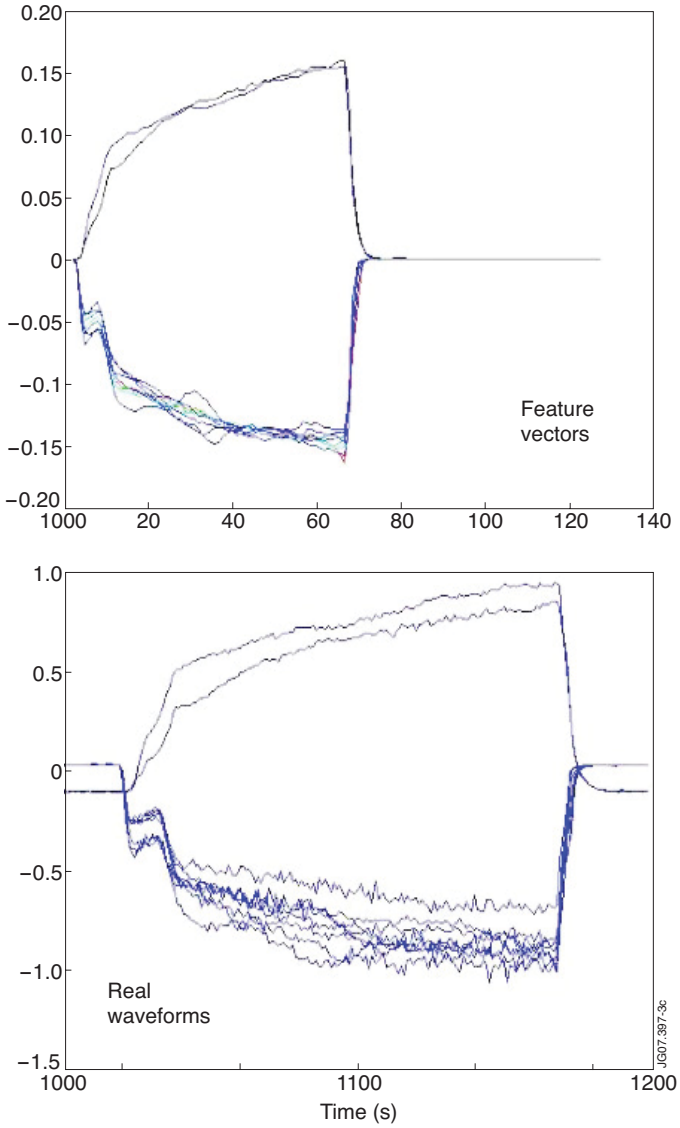

Figure 3: ES technique applied to a TJ-II database collection.
Waveforms whose difference is gain factor or polarity are recognised as similar

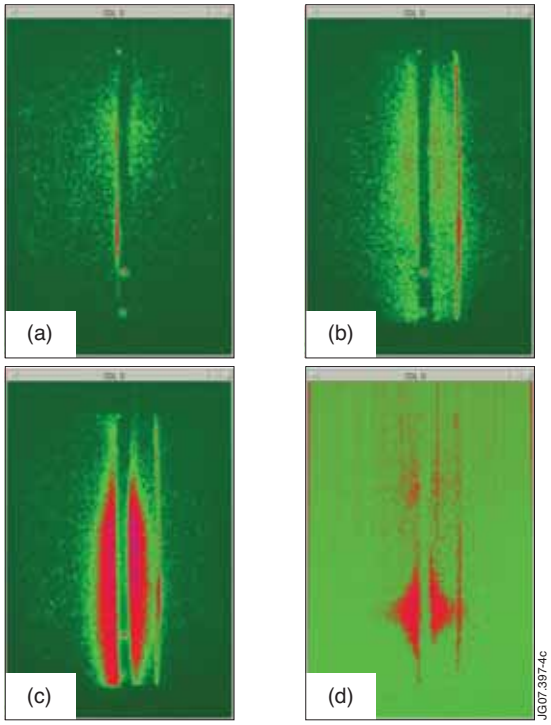

Figure 4: CCD camera images corresponding to spray light (a), ECH phase (b), NBI phase (c) and cut off density (d).

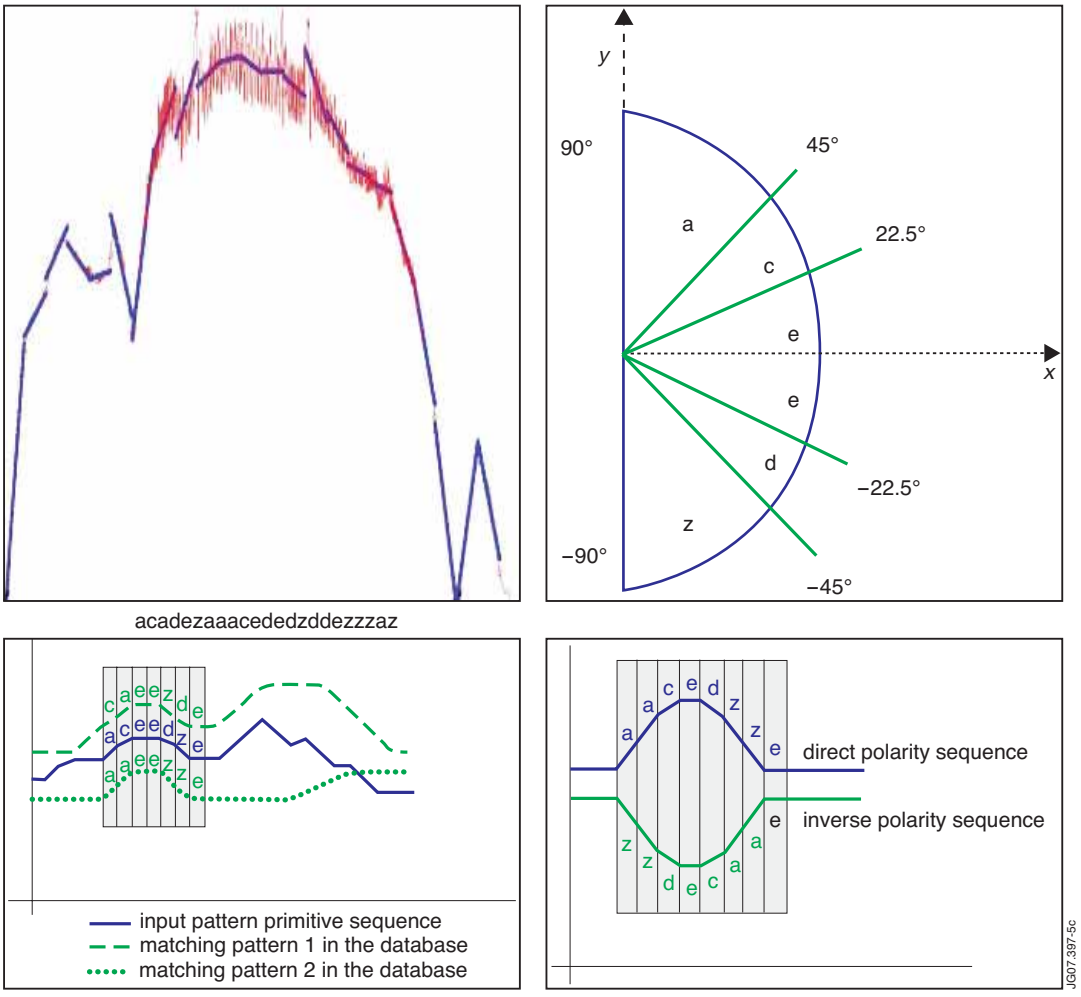

Figure 5: The slope method. Signals are fitted with straight lines and the labels are the slopes of the straight lines.

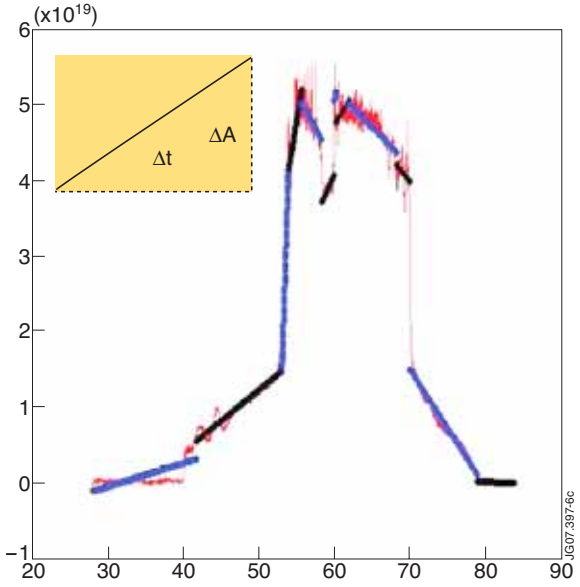

Figure 6: The slope method with segments of variable temporal length

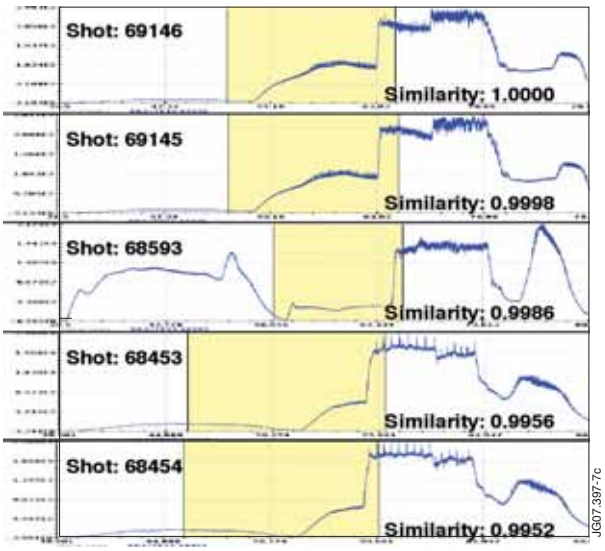

Figure 7: Search of similar patterns in JET. The waveforms correspond to ECE signals to measure electron temperature. Variable temporal length primitives were used. Note that the all patterns found follow the same behaviour but during different time: 1) a fall; 2) a fast slope (more abrupt in Pulse No: 68593 but codified with the same code); 3) a flat zone; 4) a fast rise; 5) a new flat top