EFDA–JET–CP(08)04/12

J. Vega, A. Murari, A. Pereira, A. Portas, G.A. Rattá, R. Castro
and JET EFDA contributors

Overview of Intelligent Data Retrieval Methods for Waveforms and Images 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 It is made available on the understanding that it may not be further circulated and extracts or references may not be published prior to publication of the original when applicable, or without the consent of the Publications Officer, EFDA, Culham Science Centre, Abingdon, Oxon,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.”

Overview of Intelligent Data Retrieval Methods for Waveforms and Images in Massive Fusion Databases

J. Vega, A. Murari, A. Pereira, A. Portas, G.A. Rattá, R. Castro
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
25th Symposium on Fusion Technology, Rostock, Germany
(15th September 2008 - 19th September 2008)

ABSTRACT.
JET database contains more than 42 Tbytes of data (waveforms and images) and it doubles its size about every two years. ITER database is expected to be orders of magnitude above this quantity. Therefore, data access in such huge databases can no longer be efficiently based on shot number or temporal interval. Taking into account that diagnostics generate reproducible signal patterns (structural shapes) for similar physical behaviour, high level data access systems can be developed. In these systems, the input parameter is a pattern and the outputs are the shot numbers and the temporal locations where similar patterns appear inside the database. These pattern oriented techniques can be used for first data screening of any type of morphological aspect of waveforms and images. The article shows a new technique to look for similar images in huge databases in a fast an efficient way. Also, previous techniques to search for similar waveforms and to retrieve time-series data or images containing any kind of patterns are reviewed.

1. INTRODUCTION

Typically in fusion, diagnostics generate signals with similar structural shapes for plasmas with similar physical parameters. In general, visual inspection of signals has been a first screening method to look for similar physical events. The searching process is carried out on a signal by signal and shot by shot basis in a manual way. Of course, this is a tedious, inefficient and very time consuming procedure. Automated searching methods have been proposed where the search criteria are driven by physical meanings instead of signal name/discharge number [1]. With these methods, data retrieval use a signal pattern as input parameter and the outputs are the shot numbers and the time instants where the pattern appears.

This model of data retrieval can be extremely useful to access data in the massive databases of fusion devices. The most common signals in fusion are time-series and images. The former provide the temporal evolution of plasma properties during discharges. The latter are becoming a standard way of plasma diagnosis not only for physical studies (for example turbulence, heat load measurements and plasma-wall interactions) but also for safety of the device (for instance temperature of plasma facing components, identification of regions of interest and strike points).
Signals are grouped into collections for pattern oriented data retrieval. A signal collection is the complete set of recorded signals of the same type (for instance electron temperature, density or infrared images) for all the discharges of interest.

A pattern to be searched for can be either an entire signal (an entire waveform in the case of time-series data or an entire frame in the case of movies from visible or infrared cameras) or a set of adjacent points inside a signal (a subset of adjacent samples in a temporal evolution waveform or a subset of adjacent pixels in the case of images).

One of the main characteristics of the pattern searching techniques must be efficiency. In this perspective, 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 signals most likely to contain similar patterns.

To this end, the searching methods are based on a two step procedure. The first one can be seen as an indexation phase that identifies ‘similar signals’ (signals with similar structural shape) and builds a classification system to group them together. A similarity criterion is used to compare how similar two signals are. It requires the use of a distance (in the mathematical sense) to be used as a proximity measure between signals. The indexation phase is performed once per signal.

The second step is the actual searching process. Given a reference pattern to search, the classification system is used to directly retrieve from the database the signals that contain the most similar patterns.

This article proposes a new technique to retrieve similar entire images from databases (section 2). Also, other techniques for time-series and images are reviewed. In each particular case, the indexation system and the similarity measure are described. In addition, based on the gained experience in the use of data retrieval methods with structural pattern recognition systems, discussions about performances are presented: section 3 summarizes the search of entire waveforms; section 4 describes the data retrieval of time-series data containing a specified pattern; section 5 reviews the search of a reference pattern within images. Finally, section 6 sums up some applications of morphological pattern recognition to retrieve data by translating physical phenomena identification into a pattern recognition problem.

2. SEARCHING METHOD FOR ENTIRE SIMILAR IMAGES

The method has been carried out with images of the JET KL8 fast visible camera. This is an ultra high speed visible camera presently in operation in JET that can be devoted to analyzing pellets, influxes and instabilities. Movies contain thousands of frames (274×300 pixel resolution) per discharge and the storage per movie can be of Gbytes.

An entire frame (defined by shot and frame number) is the input pattern and the outputs are the most similar entire images found within the database. The implementation has been optimized in three steps: feature extraction, similarity computation and efficient searching mechanism. As a previous step to feature extraction, the image content has to be condensed to just the most relevant structures. This is accomplished by means of thresholding. Thresholding is a fundamental technique applied in many types of image processing. It consists of an adequate threshold to be applied to images in order to retain only the most significant forms. For example, the aim of thresholding for the JET visible camera is to delete the vacuum vessel background to enhance plasma emissions.
The threshold image is a binary image (pixel values are 0 or 1). It is used as a mask in the feature extraction process. Applying the mask to an image has the effect of eliminating all pixel contents except the ones corresponding to the significant structures (fig.1). Of course, the relative value of the pixels above the threshold is not altered.

A bi-dimensional wavelet transform is used as feature extractor. The transform is applied to images after the thresholding process. Feature vectors are then made up of the approximations coefficients at a specific decomposition level. The original size of frames is 274×300 pixels and after feature extraction, the images are characterized in a lower dimensionality space: 5×5 pixels. It

should be noted that images are represented in this case by 25 attributes, only a 0.03% of the initial number of characteristics.

A classification system speeds up the search of similar images. A supervised clustering system has been developed to group together the frames whose feature vectors have exactly the same pixels with wavelet coefficients greater than 0. For example, one cluster consists of all frames with all coefficients being 0 (black images). Other cluster contains the frames whose unique pixel with coefficient greater than 0 is pixel (2,3). Another cluster is made up of all frames with positive coefficients in pixels (i, j), i = 2, 3, 4, j = 2, 3. Having into account that the number of characteristics is 25, the number of possible clusters is 2^{25}.

The concept of similarity is required to compare how similar two images are. Similarity measures have been computed by means of the Euclidean distance.

First tests of image searching have been carried out in a Windows Pentium computer with Matlab (fig.2). A total number of 135009 frames belonging to 71 discharges have been used and a threshold value of 0.9 has been selected. Given a target pattern, the waiting time to obtain the most similar frames (classification of the initial image into a cluster and similarity computation between the target pattern and the rest of images in the cluster) is about 1 ms per signal present in the cluster.

3. SEARCHING FOR ENTIRE WAVEFORMS

A specific development of this kind of systems can be found in [2]. Time series data are represented by means of the approximation coefficients of the Haar wavelet transform to a certain decomposition level. This approach allows treating the signals in a reduced dimensional space and retains the waveform time and frequency information. The similarity measure between two signals is computed as the absolute value of the normalised dot product of their Haar coefficients.

$$S_{uv} = \cos\alpha = \frac{|\mathbf{u}_w \cdot \mathbf{v}_w}{\mathbf{u}_w \cdot \mathbf{v}_w}, \quad 0 \quad S_{uv} \quad 1$$

The indexation system is based on a multi-layer classification system. The first layer is made up of the set of clusters that result after a first classification of the waveforms according to the discharge length. Some clusters contain a high number of waveforms showing different patterns and, therefore, they are sub-classified again according to a structural shape criterion that groups together waveforms whose similarity is above a specific value (fig 3). The new clusters form the second layer. Although the clustering refinement could continue in order to improve the classification system, no more than two layers have been used in the application of the method to different signal collections of the TJ-II stellarator and the JET tokamak databases [3].

The searching process is accomplished in two steps. Firstly, the initial signal is classified into one of the clusters in the classification system. Secondly, the similarity of the signal with all the existing ones in the cluster is computed. Outputs are ordered according to the similarity value.

This searching method is very efficient. The main computation time is spent in computing the similarity factor. As an average value in a personal computer, this time is 1 ms per signal present in

the cluster. In general, storage requirements for the classification systems are less than 20% of the signal collection storage.

4. SEARCHING FOR PATTERNS WITHIN WAVEFORMS

This approach allows the search of patterns inside time-series data. Patterns can be considered as composed of simpler sub-patterns. The most elementary ones are known as primitives. Primitives are represented by characters and, therefore, patterns are represented by strings of characters. The pattern recognition problem is converted into a pattern-matching problem. The indexation systems are based on database technologies that are optimised for efficient string retrieval.

Several primitive extractors have been tested. One of them consists of dividing the signals of a collection into segments of equal temporal length. Each segment is fitted with a straight line through a least square minimization process. Each segment is represented by a character depending on the slope of the fit [4]. This technique is called the slope method and was developed for the TJ-II database.

A variation of the technique was applied to the JET database (fig.4). In JET, the length of each temporal segment is variable (adaptive length primitives) [3].

Other adaptive length primitive methods use the concavity of the waveforms as feature extractor [5].

Best results are achieved with temporal segments of equal length. Typical searching times are seconds and the amount of storage for the primitives is a very small fraction of the raw data.
These techniques provide very good results when the pattern has a reduced number of primitives (number of characters < 20). As the number of primitives in the pattern increases, a single mismatch in one character avoids the recognition of the similarity. To overcome this difficulty and to identify long patterns, a new method based on using just two primitives has proven to be successful [6].

5. SEARCHING FOR PATTERNS INSIDE IMAGES

This method is fully described in [7]. Patterns are found independently of their location in the images (invariance to translations).

The technique has been applied to images of the JET fast visible camera. The feature extraction process is similar to the one describe in section 2 for entire images: thresholding and bi-dimensional Haar wavelet transform. In the present case, the Haar transform reduces the dimensionality up to 16x16 pixels that means that images are represented by 256 attributes. Each attribute is represented by a single letter depending on the wavelet coefficient and according to the procedure described in figure 5. The number of possible letters (primitives) is variable but presently 4 letters are considered. Due to the fact that feature vectors are made up of letters, the recognition task is again turned into a pattern matching problem. The searching process needs a powerful indexation system and, therefore, database technologies can be used. In particular, the system developed for JET uses the PostgreSQL relational database.

A Matlab application allows selecting patterns inside waveforms and similar patterns are found typically in times of 1 minute for a database with 26000 frames. The classification system demands very low additional storage for the letter database (0.1 % of the size of the movies).

6. STRUCTURAL PATTERN RECOGNITION FOR SPECIFIC PHYSICAL PHENOMENA AT JET

The previous sections show how to use the morphological information of the signals to develop general purpose data retrieval systems. If domain dependent knowledge is added to the structural information, particular data retrieval methods for specific physical phenomena can be developed. This has been applied in JET to look for both cut-offs in ECE heterodyne radiometer signals and L-H transitions [8].

The first case is an example of physical phenomena recognition by just a single pattern in temperature waveforms. The second case is an application of multiple pattern recognition to identify regime transitions combining the analysis of density and D_α signals.

ACKNOWLEDGEMENTS

This work was partially funded by the Spanish Ministry of Science and Innovation under the Project No. ENE2008-02894/FTN. This work, supported by the European Communities under the contract of Association between EURATOM/CIEMAT, was carried out within the framework of the European Fusion Development Agreement. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

REFERENCES

[1]. J. Vega and JET EFDA Contributors. Intelligent methods for data retrieval in fusion databases. Fusion Engineering and Design. **83** (2008) 382-386.

[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. Fusion Engineering and Design. **83** (2008) 132-139.

[3]. J. Vega, G. A. Rattá, A. Murari, P. Castro, S. Dormido-Canto, R. Dormido et al. Recent results on structural pattern recognition for Fusion massive databases. Proc. of the IEEE International Symposium on Intelligent Signal Processing. ISBN: 1-4244-0829-6 (2007) 949-954.

[4]. 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.

[5]. S. Dormido-Canto, G. Farias, R. Dormido, J. Vega, J. Sánchez, N. Duro et al. Structural pattern recognition methods based on string comparison for fusion databases. Fusion Engineering and Design. **83** (2008) 421-424.

[6]. A. Pereira, J. Vega, A. Portas, R. Castro, A. Murari and JET-EFDA Contributors. Optimized Search Strategies to Improve Structural Pattern Recognition Techniques. Proc. of the 8th FLINS Conference on Computational Intelligence in Decision and Control.

[7]. J. Vega, A. Murari, A. Pereira, A. Portas, P. Castro and JET-EFDA Contributors. Intelligent Technique to Search for Patterns within Images in Massive Databases. Review of Scientific Instruments (in press).

[8]. 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. Fusion Engineering and Design. **83** (2008) 467-470.

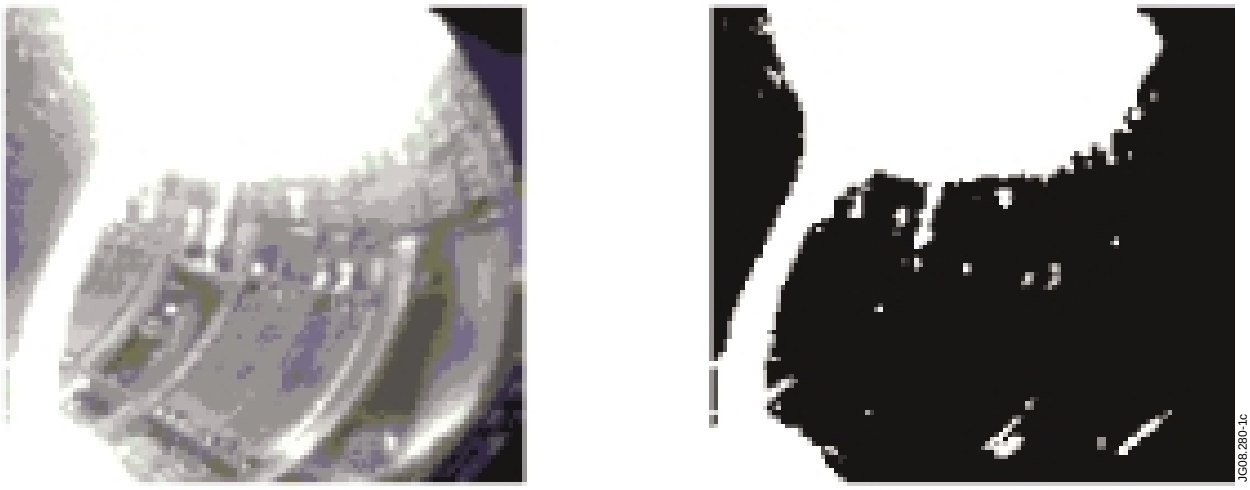

Figure 1: KL8 images. Raw frame (left) and after thresholding (right).

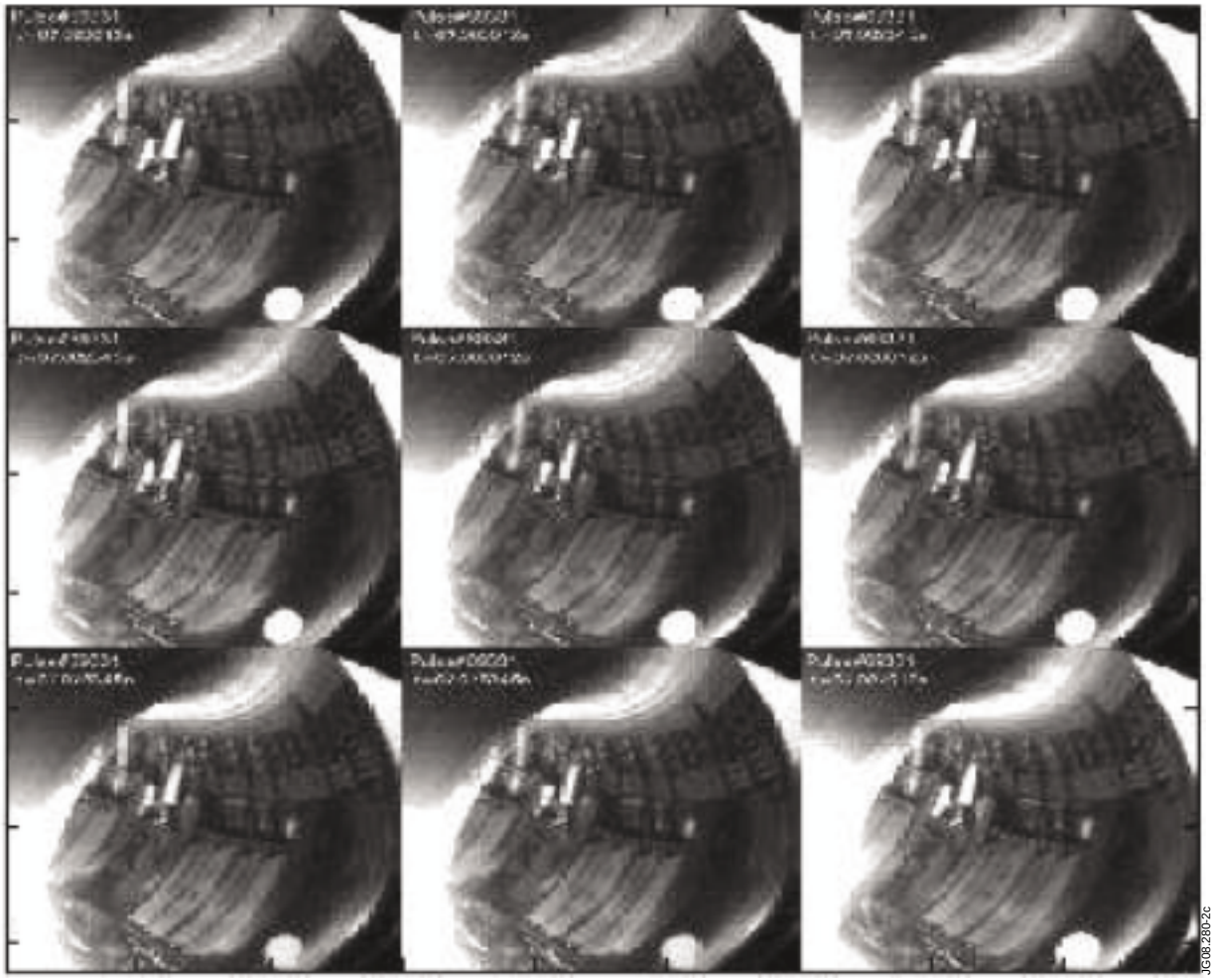

Figure 2: Target pattern (top-left) and similar frames.

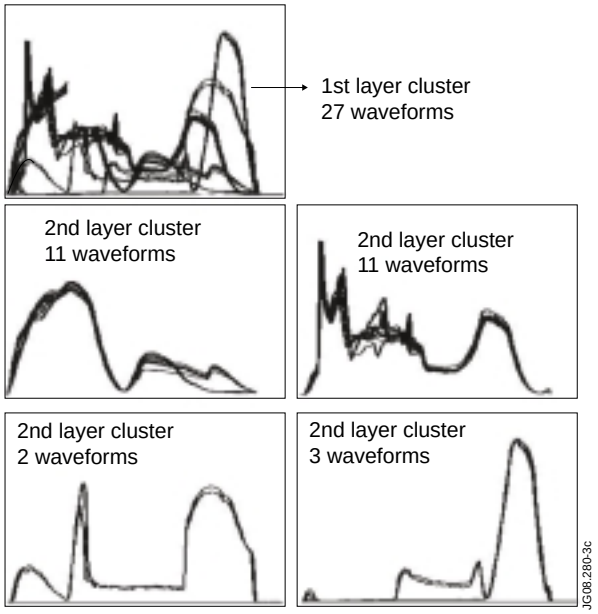

Figure 3: Example of a cluster splitting with Electron Cyclotron Emission (ECE) waveforms from the JET database.

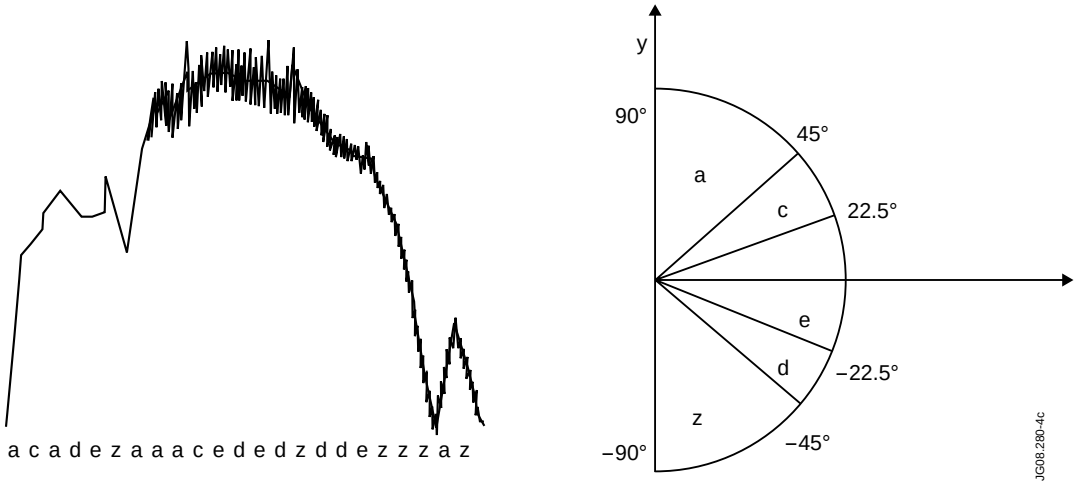

Figure 4: The slope method. Signals are fitted with straight lines and the labels are the slopes of the straight lines.

Thresholding: $U = (u_{ij})$, i = l,..., 274, j =1,...300
Haar: $U = (u_{ij}) \Leftrightarrow H_U = (h_{kl})$, k =1,..., 16, l = 1,...16
Label assignment: rounding toward infinity (h_{kl} / Δh)
Feature vectors: 2D string of characters {a, b, c, d}
$C = (c_{mn})$, m = 1,..., 16, n=1,...16

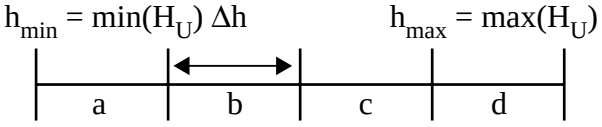

Figure 5: Label assignment after reducing the dimensionality to 16×16 and considering 4 primitives.