EFDA–JET–CP(08)01/05

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

Intelligent Technique to Search for Patterns within Images in Massive Databases

J. Vega[1], A. Murari[2], A. Pereira[1], A. Portas[1], P. Castro[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
HTPD High Temperature Plasma Diagnostic 2008, Albuquerque, New Mexico.
(11th May 2008 - 15th May 2008)

ABSTRACT

An image retrieval system for JET has been developed. The image database contains the images of the JET high speed visible camera. The system input is a pattern selected inside an image and the output is the group of frames (defined by their discharge numbers and time slices) that show patterns similar to the selected one. This approach is based on morphological pattern recognition and it should be emphasized that the pattern is found independently of its location in the frame. The technique encodes images into characters and, therefore, it transforms the pattern search into a character-matching problem.

1. INTRODUCTION

Data retrieval methods in present nuclear fusion devices can no longer be based on manual searches according to signal name and shot number. A new approach for data access has been proposed to look for data according to scientific and technical criteria [1]. This approach is founded on the fact that diagnostics generate reproducible signal shapes (morphological patterns) for similar physical behaviour. Therefore, the input parameter of a data searching system is a signal pattern and the outputs are the shot numbers and the time intervals where the pattern appears.

The structural shape of the signals has already been used in several data retrieval systems for the JET and TJ-II databases. An initial development was the searching of entire similar waveforms. In this case, the input pattern is a complete waveform and the outputs are the shot numbers where similar waveforms exist. Applications to TJ-II and JET can be found in [2] and [3] respectively.
The next step was to look for patterns inside waveforms. Graphical user applications were developed to select in a graphical way the pattern of interest within a waveform and to discover the shot numbers and the temporal instants where similar patterns can be found. The methodology is described in [4] and the application to the JET database is shown in [5].

Nowadays, image diagnostics are becoming very popular in magnetic confinement fusion as a consequence of the advanced capabilities of digital cameras and personal computers. Intelligent data retrieval methods are therefore also needed for accessing similar image patterns inside massive databases. Searching systems for entire images have been developed for both JET [6] and TJ-II [3]. Finally, the capability of finding patterns within video-images, independently of its location in the image (translation invariance) is also needed. This article shows the specific application to the JET visible high speed camera. Section 2 presents a description of the problem. Section 3 describes the JET searching system. In the end, the results and the perspectives of future work are the subject of section 4.

2. PROBLEM DESCRIPTION

The JET fast visible camera generates video movies with thousands of frames and more than 50000 pixels (10 bits) per image. Physical events like MARFEs, striations, filamentary structures or plasma-wall interaction are recognised by particular footprints (patterns) in the images. Due to the fact that

most studies in plasma physics require a statistical analysis, a typical situation takes place when one needs to know the set of discharges and the temporal intervals where specific plasma behaviour happened. To this end, visual data analysis tools should be developed to help locating the patterns of interest in the databases of camera frames. These tools would have to provide the following main capabilities: visualizing movies, choosing frames inside movies, selecting the pattern to search, and showing the list of shot numbers and time slices where similar patterns were found. It should be noted that the last point must be accomplished without missing analogous patterns and, for efficiency reasons, without having to traverse the whole database at the time of the searching process.

3. SEARCHING SYSTEM

To implement data retrieval algorithms for images, it seems natural to be inspired by pattern recognition techniques. The two key elements of any pattern recognition problem are object descriptions and similarity measures. The first one refers to the choice of parameters needed to properly describe the objects. Objects are represented by feature vectors that contain the object attributes of distinctive nature. Feature extraction has a twofold objective as it allows reducing the problem dimensionality and also translating the object descriptions into adequate representations to be understood by computers. The second element, the similarity measure, is a distance (in the mathematical sense) to compare how similar two objects are.

This section describes the particular implementation of the above elements to the JET searching system and it also explains the way to carry out the search of patterns.

3.1 PATTERN RECOGNITION SYSTEM

In the present case, the objects of the pattern recognition system are the video frames of the JET fast visible camera. The feature extraction process is accomplished in three steps: thresholding, 2D Haar wavelet transform and primitive computation. The thresholding phase can be seen as an intensity high pass filter. The technique looks for an adequate threshold to be applied to images in order to retain only the more significant shapes and to eliminate the vacuum vessel background to enhance plasma emissions (fig.1). The threshold value used here was 90% of the camera maximum intensity. To achieve an important dimensionality reduction, a bi-dimensional Haar wavelet transform is applied to images after the thresholding process. The resulting bi-dimensional array consists of the approximation coefficients of the transform at a specific decomposition level. The original size of the frames is 274x300 pixels and after the Haar transformation, the images are characterized in a lower dimensionality space. Tipically the dimensions are reduced to 32x32 and 16x16. It should be noted that images are represented in these cases by 1024 or 256 attributes respectively, only a 1.25% or 0.31% of the initial number of attributes.

The last step is called primitive computation. Patterns can be considered as composed of simpler sub-patterns and the most elementary ones are known as primitives. In the present approach, the primitives are related to each one of the 1024 (or 256) attributes after the wavelet transform. Each

attribute is represented by a single letter depending on the wavelet coefficient and according to the procedure described in fig. 2. Therefore, the feature vectors that describe the frames are matrixes of $N_{rows}xN_{colums}$ letters (32x32 or 16x16). Results in this article will be shown with both 4 primitives and 2 primitives.

Due to the fact that the components of the feature vectors are letters, the recognition task is turned into a pattern matching problem. Hence, the searching process needs a powerful indexation system and, therefore, database technologies can be used. In particular, the system presented in this article uses a relational database: PostgreSQL (http://www.postgresql.org).

Four different searching systems have been developed: 32x32 and 16x16 letters with both 4 and 2 primitives. In each case, feature vectors are stored in a single table. The primary key consists of the shot number and frame order inside a movie. In addition, there are a number of columns equal to N_{rows} that is the number of rows after dimensionality reduction. Each table column stores $N_{columns}$ characters, *i.e.* the number of columns after dimensionality reduction. Figure 3 shows an example in which feature vectors are represented by a 4x2 matrix and three primitives.

3.2 PATTERN SEARCHES

The graphical user interface to look for patterns is a Matlab application. The user selects the shot number and the movie is reproduced. Several controls allow the rewinding/forwarding of the frames. A particular frame can be selected and a pattern is chosen by displacing the mouse through the area of interest of the image.

The searching process is accomplished in two phases: feature extraction and pattern search. The former processes the selected frame to get its corresponding feature vector exactly in the same way that images have been processed to create the relational database: thresholding, Haar transform and primitive computation. The pixels selected by the user have their equivalent in the transform space and their letter content defines the pattern to search. The pattern search is carried out by means of the SQL database language with a SELECT instruction. It should be emphasised that the search of a pattern implies to find the first row of the pattern in a column of the database and then to search for the remaining rows in consecutive columns and in equivalent positions (figure 4). The outputs are the shot numbers and the frame identifications where similar patterns appear. The output is ordered according to the similarity value. It is computed with the wavelet coefficients of the feature vectors as the Euclidean distance between the initial pattern and the retrieved ones. Therefore, the lesser value the more similar are the patterns.

4. RESULTS AND FUTURE WORK

About 26000 frames with an initial storage of 6 Gbytes have been used to develop the four classification systems mentioned above. Searching processes have been carried out with different pattern sizes. Typically, less than half an image has been selected. The conclusions from the searching systems can be summarised in 6 points:

1. The technique works and similar patterns are retrieved. A target pattern can exist in the letter database or not. If it exist, the pattern is always found with the maximum similarity (Euclidean distance = 0).
2. The classification systems demand very low additional storage for the letter databases. The amounts were 26 Mbytes for the 32x32 case and 7 Mbytes for the 16x16 case, a very small fraction of the movies size (just a 0.4% and 0.1% respectively).
3. With a resolution of 32x32 coefficients, typically few similar patterns are found and they show big similarities (small Euclidean distances). With a poorer resolution (16x16) more patterns are found with lesser similarities (big Euclidean distances), as it was expected. The latter has a lesser computational cost.
4. The comparison of results with 4 and 2 primitives gives a similar situation to the previous point and much more patterns are retrieved with the second system. This is easily understood taking into account that reducing the number of primitives provides lesser resolution in the process of assigning primitives, just as it happens for instance when digitizing waveforms from 12 to 10 bits. The CPU time in the searching process with 2 primitives can be 10 times greater than the one with 4 primitives.
5. In the worst cases, searching times with 2 primitives were 10 minutes and typical times are minutes. This result suggests the development of batch processes to look for similarities. These processes should be executed as detached processes from the visual user interface. Communication mechanisms to notify the end of the executions and the availability of results should therefore be implemented.
6. The high searching times are directly related to the amount of data in the letter database. In general, the output of a searching process shows several consecutive frames containing similar patterns. The presence of consecutive frames in the output can be considered as redundant. The important point is the detection of a similar pattern around a time instant. Therefore, the inclusion of similar consecutive frames in the letter database should be avoided. This can be also part of future work to filter the frames which are encoded in the letter database.

ACKNOWLEDGEMENT

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. “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, J. Sánchez, N. Duro,

M. Santos, E. Sánchez, G. Pajares. “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, G. Farias, A. Pereira, A. Portas, E. De la Luna, I. Pastor, J. Sánchez, N. Duro, R. Castro, M. Santos, H. Vargas. “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. S·nchez, M. Santos, J. A. Martín, G. Pajares. “Search and retrieval of plasma waveforms: structural pattern recognition approach”. Rev. Sci. Ins. **77** (2006) 10F514.

[5]. G.A. Ratt·, J. Vega, A. Pereira, A. Portas, E. De la Luna, S. Dormido-Canto, G. Farias, R. Dormido, J. Sánchez, N. Duro, H. Vargas, M. Santos, G. Pajares, A. Murari and JET EFDA Contributors. “First applications of structural pattern recognition methods to the investigation of specific physical phenomena at JET”. Fusion Engineering and Design. **83** (2008) 467-470.

[6]. J. Vega, G.A. Rattá, A. Murari, P. Castro, S. Dormido-Canto, R. Dormido, G. Farias, A. Pereira, A. Portas, E. De la Luna, I. Pastor, J. Sánchez, N. Duro, R. Castro, M. Santos, H. Vargas. “Recent results on structural pattern recognition for Fusion massive databases”. Sent to IEEE Transactions on Instrumentation and Measurement.

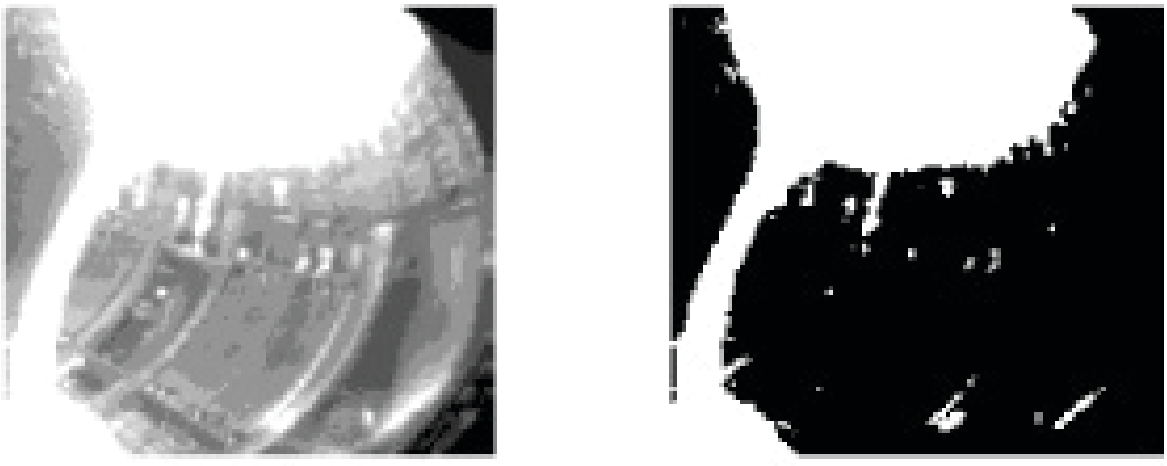

Figure 1: Filtered image after thresholding.

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

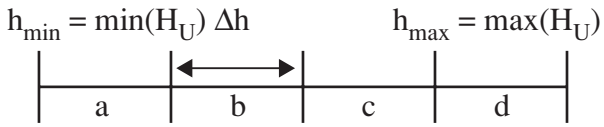

Figure 2: Label assignment after reducing the dimensionality to 16×16 and considering 4 primitives.

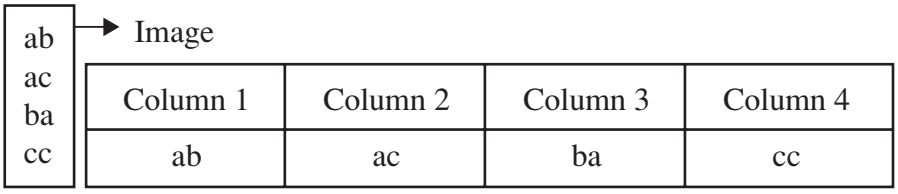

Figure 3: Example of an image in a table with 4×2 characters and 3 primitives (a, b, c).

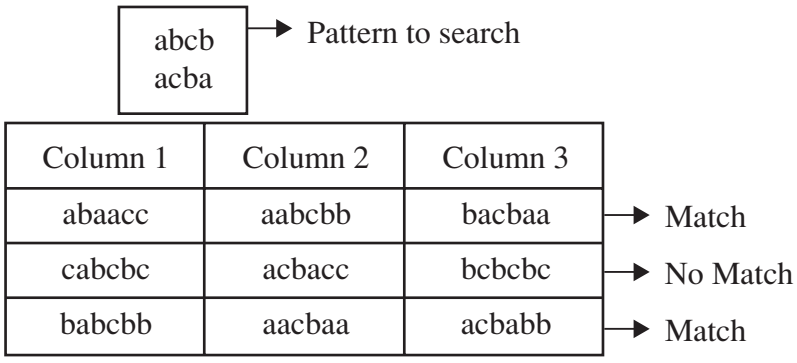

Figure 4: Example of matches for images of 3×6 attributes and a pattern of 2×4 characters.