AUTOMATIC ELM LOCATION IN JET USING A UNIVERSAL MULTI-EVENT LOCATOR

S. GONZÁLEZ,[a]* J. VEGA,[a] A. MURARI,[b] A. PEREIRA,[a] M. BEURSKENS,[c] and JET-EFDA CONTRIBUTORS†

JET-EFDA, Culham Science Center, Abingdon OX14 3DB, United Kingdom
[a]*Asociación EURATOM CIEMAT para Fusión, Madrid, Spain*
[b]*Consorzio RFX - Associazione EURATOM ENEA per la Fusione, Padova, Italy*
[c]*EURATOM/CCFE, Culham Science Center, Abingdon OX14 3DB, United Kingdom*

Received March 15, 2010
Accepted for Publication May 10, 2010

Massive amounts of data generated by fusion machines (such as JET) require developing automatic methods for data analysis. Edge-localized modes (ELMs) are instabilities occurring in the edge of H-mode plasmas. The aim of this work is to develop an automatic off-line method for identifying and locating ELMs. This method uses Universal Multi-Event Locator (UMEL) as the event locator. The combination of information from D_α emission and diamagnetic energy allows the recognition of single ELMs. This paper shows the way in which waveforms of a wide range of discharges can be treated and how UMEL is applied in order to identify and locate ELMs independently of the signal amplitudes. A large database of more than 1200 discharges has been used to test the performance of the method obtaining 226 751 ELMs.

KEYWORDS: *edge-localized modes, universal multi-event locator, support vector regression*

I. INTRODUCTION

In modern fusion machines such as JET, machine safety is one of the most important factors to take into consideration. All the risks should be mitigated during plasma operation as much as possible to increase the operational safety.

Edge-localized modes (ELMs) are instabilities occurring in the edge of H-mode plasmas. These events can be dangerous in high-performance scenarios of fusion machines. In order to minimize the potential problems caused by ELMs, the statistical knowledge of these phenomena should be improved. Most ELMs occurring in tokamak experiments are not indexed, so the information provided is not easily used or is totally wasted. Nowadays, the location process is not automatic. In the context of this paper, "automatic location" has a specific meaning. First, it should be emphasized that no publications exist (to our knowledge) on a general methodology to locate ELMs. Visual data analysis (VDA) to identify ELMs consists of recognizing typical peaks in the H_α/D_α signals that are synchronous with a drop in the stored diamagnetic energy. In software applications, generally, people identify ELMs by means of software functions that implement the manual procedures of VDA. These functions are completely dependent on waveform amplitudes and noise. Therefore, if the amplitude or noise changes from one discharge to another (or even within a single discharge), the software is modified to be adapted to the signal conditions at any time. In this way, the typical user software may include many IF-THEN-ELSE sentences to manage the recognition of large and small ELMs. This identification procedure is tedious and requires strong human intervention. To avoid continuous editing/compiling of software, automatic location methods should be developed. The term "automatic" refers to the existence of the same software not depending on the ELM size.

*E-mail: sergio.gonzalez@ciemat.es
† See the Appendix of F. Romanelli et al., *Proceedings of the 22nd IAEA Fusion Energy Conference 2008*, Geneva, Switzerland.

This paper develops an automatic method to locate ELMs in plasma signals. The combination of information provided by the D_α emission and the stored diamagnetic energy is used to determine the exact temporal location of every single ELM. Consequently, large ELM databases can be automatically generated with this technique. These databases can be used as input for other automatic methods such as the ELM classification system described in Ref. 1 and other approaches that require large ELM databases with high statistical weight.

A new Universal Multi-Event Locator[2] (UMEL) has recently been developed. It allows the location of events in a wide range of data types, such as waveforms or images. In this paper UMEL is used to locate the specific events in the D_α emission (peaks) and the stored diamagnetic energy (drops) that allow the automatic location of individual ELMs. It should be emphasised that UMEL recognizes large and small ELMs without any change in the source code. Because of its reusability property, exactly the same software is applied to the D_α and the stored diamagnetic energy signals. This means that no separate functions are necessary for each individual waveform, thereby making software maintenance easier.

This paper is structured as follows. Section II introduces Support Vector (SV) Regression (SVR) and UMEL. It explains how it can be applied to the ELM location problem. Section III describes the two main steps required to locate ELMs. Results are presented in Sec. IV. Finally, conclusions and directions of future work are given in Sec. V.

II. UMEL AND SVR

UMEL is a technique to locate relevant events in the signals. UMEL is based on a SVR estimation method. Section II.A provides a brief explanation of the SVR theory, and Sec. II.B presents an overview of the most important aspects of UMEL.

II.A. Support Vector Regression

Let us consider S training samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_S, y_S)$, $(\mathbf{x}_i \in \mathbb{R}^n$ and $y_i = f(\mathbf{x}_i)$, where $f: \mathbb{R}^n \to \mathbb{R})$. The regression function is given by[3]

$$f^*(x) = \sum_{k=1}^{S} \gamma_k^* H(x_k, x) \ . \tag{1}$$

The parameters γ_k^* are determined using the solution of a quadratic optimization problem as follows:

$$\gamma_k^* = \alpha_k^* - \beta_k^* \ , \quad k = 1, \ldots, S \ ,$$

where the parameters are chosen by maximizing the function as follows:

$$Q(\alpha, \beta) = -e \sum_{k=1}^{S} (\alpha_k + \beta_k) + \sum_{k=1}^{S} (\alpha_k - \beta_k) - \frac{1}{2} \sum_{k,l=1}^{S} (\alpha_k - \beta_k)(\alpha_l - \beta_l) H(x_k, x_l)$$

subject to the following constraints:

$$\sum_{k=1}^{S} \alpha_k = \sum_{k=1}^{S} \beta_k \ , \quad 0 \le \alpha_k \le \frac{C}{S} \ ,$$

$$0 \le \beta_k \le \frac{C}{S} \ , \quad k = 1, \ldots, S$$

given the training data $(\mathbf{x}_k, y_k)$, $k = 1, \ldots, S$, an inner product kernel $H(x, x')$, an insensitive zone e, and a regularization parameter C.

II.A.1. Insensitive Zone e

The quality of the approximation produced by a learning machine is measured by a loss function $L(y, f(x))$ or discrepancy between the output produced by the system and the training set for a given point x. Large values of the loss function correspond to poor approximations. The regression formulation for the SV machines uses a special loss function defined in Ref. 4. This loss function is linear with an insensitive zone e (Fig. 1a):

$$L(y, f(x)) = \begin{cases} 0 \ , & \text{if } |y - f(x)| \le e \\ |y - f(x)| - e \ , & \text{otherwise} \ . \end{cases} \tag{2}$$

The area between $fit - e$ and $fit + e$ is called e-tube, and it will have an important role in UMEL as is shown in Sec. II.B.

II.A.2. Support Vectors

Only a subset of the parameters γ_k^* in Eq. (1) is nonzero. The data points x_k associated with the nonzero γ_k^* are called SVs. Therefore, the regression function is actually

$$f^*(x) = \sum_{SV} \gamma_k^* H(x_k, x) \ . \tag{3}$$

II.B. Universal Multi-Event Locator

This section reviews how the information contained in the SVs can be used to perform an automatic method for event location.

The aim of UMEL is the automatic location of any type of event inside any class of signals (for example, images and waveforms). Equation (1) expresses the fact that the SVR estimation of complex data sets can require all samples to obtain a regression model. On the other hand, it should be taken into account that the complexity of a function can be defined in terms of its smoothness

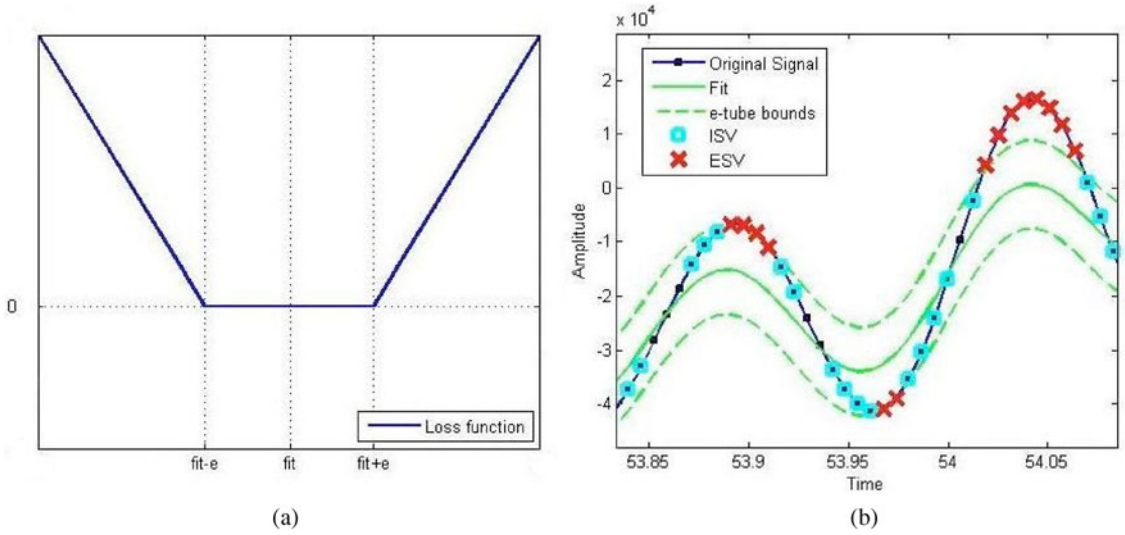

Fig. 1. Loss function for SVR and an example of the two different types of SVs.

since for smoother functions fewer data points are required for an accurate estimation. This is the meaning of Eq. (3). The smoother the function to regress, the fewer SVs are required.

In the UMEL technique, the training samples that lie on or outside the *e*-tube defined by the loss function in Eq. (2) are called External SVs (ESVs) [Eq. (4) and Fig. 1b]. But, some samples within the *e*-tube also become SVs, which are called Internal SVs (ISVs) [Eq. (5) and Fig. 1b]:

$$\mathrm{ESV} \equiv \mathrm{ESV} \subset \mathrm{SV} \ , \quad \forall i \in \mathrm{ESV} \ , \quad |y_i - f(x_i)| > e \tag{4}$$

and

$$\mathrm{ISV} \equiv \mathrm{ISV} \subset \mathrm{SV} \ , \quad \forall i \in \mathrm{ISV} \ , \quad |y_i - f(x_i)| \leq e \ . \tag{5}$$

Although ISVs are necessary samples for the regression estimation, they do not present the degree of relevance that can be assigned to the ESVs. The samples that become ESVs are the most difficult samples to regress (they cannot be fitted inside a smooth *e*-tube), and these samples provide essential additional information in the regression process. The ESVs reveal the occurrence of special patterns inside a signal: peaks, high gradients, or segments with different morphological structure in relation to the bulk of the signal. Typically, several ESVs appear together, and they define limited signal segments. On one hand, singular points show a signature with a strong location in the signal domain. This property is used for the location of ELMs. On the other hand, large sequences of ESVs denote the presence of signal intervals with patterns clearly different from the rest of the signal.

In SVR, there are three parameters that can affect the final model complexity, i.e., the smoothing degree of the regression estimation: the regulation parameter C, the kernel function, and the e value. The regularization parameter C can be estimated (see Ref. 3, p. 448) as

$$C = K_c \max(|\bar{y} + 3\sigma_y|, |\bar{y} - 3\sigma_y|) \ , \tag{6}$$

where

$\bar{y}$ = mean of the training samples

σ_y = standard deviation of the training samples

K_c = constant dependent on the type of signal to fit.

In UMEL, the criterion in the selection of the e value is based on the standard regression formulation: $y = t(x) + \xi$. Typically, it is assumed that the standard deviation of additive noise σ_{noise} is known or can be reliably estimated from the data. Then, the e value should reflect the level of additive noise ξ, that is, $e_{value} \propto \sigma_{noise}$. In particular, and according to Ref. 3, p. 449, the selection criterion used is

$$e_{value} = K_e \sigma_{noise} \sqrt{\frac{\ln(n)}{n}} \ , \tag{7}$$

where

n = number of training samples

$\ln(n)$ = natural logarithm of the number of training samples

K_e = constant dependent on the type of signal to fit.

The kernel function $H(x, x')$ will determine the shape of the regression estimation. This function can be a linear function, a polynomial function, or a radial basis function (rbf), among others. Once the type of kernel is chosen, the kernel parameters (for example, in the case of a polynomial function, the polynomial degree) must be determined. In the present ELM location method, a radial basis kernel function has been used:

$$H(x_k, x) = \exp\left(-\frac{\|x - x_k\|^2}{2\sigma_k^2}\right) , \quad (8)$$

where the σ_k value can be estimated from the input data using the normal reference rule defined in Ref. 5:

$$\sigma_k = K_\sigma \cdot 1.06 \cdot \sigma_y \cdot n^{-1/5} , \quad (9)$$

where

σ_y = standard deviation of the training samples

n = number of training samples

K_σ = constant dependent on the type of signal to fit.

III. ELM LOCATION METHOD

As previously explained, the aim of this work is to develop an automatic off-line method to recognize and locate individual ELMs in JET discharges. It is carried out through the simultaneous location of both a peak in the D_α and a drop in the stored diamagnetic energy signals. It should be mentioned again that "automatic" means to use the same software independently of signal noise and amplitude. UMEL is a proper tool for these purposes. In addition, UMEL provides the extra capability of executing exactly the same software to locate the peaks and the drops in the different signals. The only difference between signals is the selection of the different parameters to accomplish the regression process: the regularization parameter, the *e*-tube width, and the rbf parameter. Estimations to select these values are given in Eqs. (6), (7), and (9), respectively.

The present ELM location method consists of two main phases: locating the global temporal interval with ELMs and detecting individual ELMs:

1. *Locating the global temporal interval with ELMs*: The D_α emission waveform is processed in order to determine the global interval in a discharge with ELMs and to make the individual ELM location process easier. This phase is explained in Sec. III.A.

2. *Detecting individual ELMs*: The D_α emission peaks and the stored diamagnetic energy drops are located. The information is combined to identify each single ELM. This phase is explained in Sec. III.B.

III.A. Locating the Global Temporal Interval with ELMs

The objective of this phase is to delimit the temporal segment of a discharge in which ELMs appear. This preprocessing step is required to save computational time in the regression process. The application of UMEL makes no sense in time slices where no ELM activity is present. Therefore, the search for ELMs is focused only on H-mode segments because ELMs appear only in this confinement mode. So, this phase is actually a gross H-mode locator system that uses the D_α signal as identifier.

Three sequential tasks are performed: normalizing the waveform, reducing the dimensionality, and locating the time interval with ELMs.

III.A.1. Normalizing the Waveform

The D_α waveform is normalized between 0 and 1. The purpose of this normalization is double. On one hand, it allows the use of the same regression parameters defined in Eqs. (6) and (7) for a large number of discharges. On the other hand, it also optimizes the computation of the SVR estimation.

III.A.2. Reducing the Dimensionality

In the regression estimation process, the more samples to regress, the more computation time is needed. In order to save processing time, the number of samples per waveform should be reduced. However, this reduction should not completely degrade the signal. A proper transformation should provide just a coarser resolution of the original time sequence. To this end a Haar wavelet transform[6] has been used. This transformation is adequate because it simultaneously retains the most relevant signal information in the time and frequency domains.

Wavelet algorithms process data at different resolutions or decomposition levels. Each decomposition level reduces the number of samples by a factor of 2. Therefore, for a waveform with S initial samples, the number of samples becomes $S/2^L$, after using a decomposition level of L. A level $L = 2$ has been used in this paper, which means that waveforms with an initial number of 131 072 samples have been reduced to 32 768 samples without loss of significant information.

III.A.3. Locating the Time Interval with ELMs

UMEL has been used as a gross H-mode locator system. Given a discharge, a SVR is computed with all D_α samples (Fig. 2a). The regression parameters are chosen according to Eqs. (6), (7), and (9) to produce a smooth approach of the waveform, and a set of ESVs is found. The number of ESVs is grouped in slots of 0.1 s (Fig. 2b). This histogram defines the temporal segment that has been more difficult to regress smoothly, and therefore, the regression estimation reveals the signal slice with higher-frequency components (peaks). This temporal slice

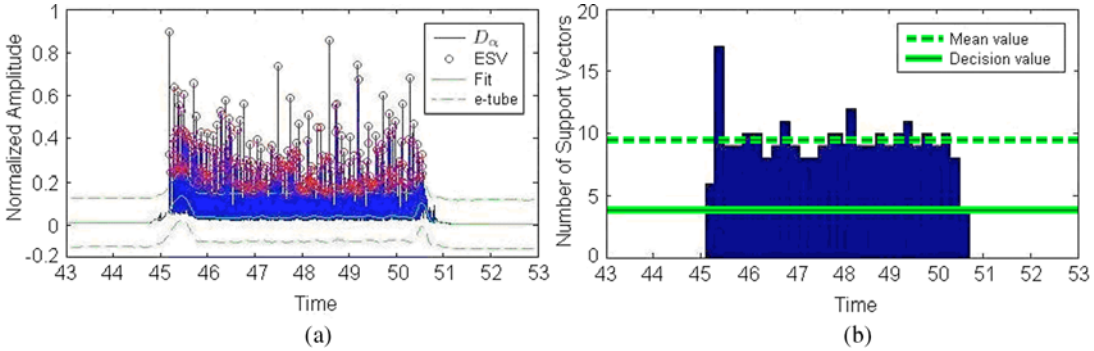

Fig. 2. Detection of the ELM temporal slice.

corresponds to a period of time with ELM activity. To delimit the global temporal interval of ELMs, starting and ending histogram bins are determined. They are defined by the first and last time slots that have more ESVs than a decision value, respectively. This decision value is computed as

$$decision_{value} = K \cdot \frac{\sum_{i=1}^{N} \mathrm{ESV}_i}{N} \quad , \tag{10}$$

where

K = constant depending on the discharge range

N = number of time slots with ELMs (number of bins in Fig. 2b).

An example of the number of ESVs per time slot is shown in Fig. 2b. The dashed line is the mean value of ESVs. The solid line corresponds to the decision value using $K = 0.5$. It is important to note that the value $K = 0.5$ can be used for the detection of the time interval with ELMs in a large range of discharges without any problem. In fact, this value has been used in all the discharges that we have tested in Sec. IV.

To finish, it should be noted that the same global interval is determined with and without the Haar wavelet transform for dimensionality reduction. The difference is the computational time. With an initial waveform of 131 072 samples, UMEL takes 378.9 s in the regression process. However, the time for the wavelet transform at level 5 of decomposition (4096 samples remain) plus the regression estimation is 6.32 s.

III.B. Location Phase

After establishing the global temporal interval with ELMs, the next phase of the method is the temporal location of individual ELMs.

The location capability of this phase relies on the combination of the information provided by the D_α emission and the diamagnetic energy. Every single ELM is located if a diamagnetic energy drop is found very close (within 5 ms) to a D_α peak. Because typically the D_α waveform has a better signal-to-noise ratio, the ELM location process begins searching for the typical peaks in this signal. After determining the time instant of maximum amplitudes in the D_α waveform, simultaneous drops in the diamagnetic energy should appear.

The location phase consists of four steps: D_α peak location, D_α peak combination, diamagnetic energy division, and combination of information.

Step 1: D_α Peak Location

The first step of the ELM location requires the location of peaks in the D_α signal. It must be emphasized that the identification of peaks as the points above a certain threshold is not enough because ELMs have different amplitudes. The main advantage of using UMEL resides in the fact that UMEL looks for samples that do not fit a smooth regression, independently of amplitudes.

UMEL is applied to the D_α temporal slice determined in the phase described in Sec. III.A. Again, the regression parameters are estimated according to Eqs. (6), (7), and (9) (Fig. 3, step 1). The ESVs appear in the most difficult areas to fit smoothly because high gradients are present. Although the regression estimation can seem a straight line in Fig. 3, it is not true. The regression process carried out by UMEL adapts the fit to the low-frequency shape of the D_α waveform. All D_α samples outside the *e*-tube become ESVs, and they define the peak location.

Step 2: D_α Peak Combination

The second step concentrates all the ESVs of each peak into a single one. The selected point is the D_α sample with the highest amplitude (Fig. 3, step 2) and whose time instant is denoted by t_D. This combination is required to represent each individual peak by a single sample.

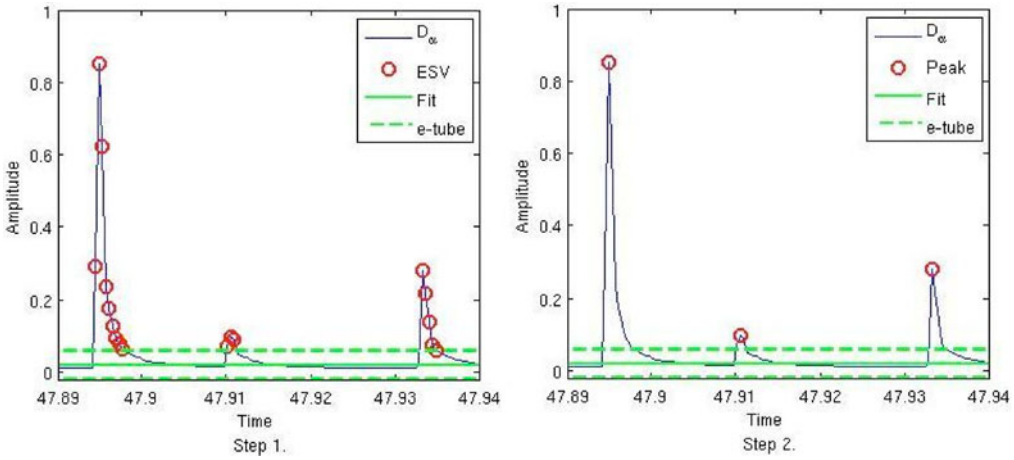

Fig. 3. Location method steps 1 and 2. This corresponds to shot 75742 of JET. The recognition of peaks independent of amplitudes should be emphasized. All the regression parameters (regularization parameter, *e*-tube width, and kernel parameter) are obtained from each signal itself.

Step 3: Diamagnetic Energy Division

Once the t_D time instants have been determined, the stored diamagnetic energy signal is divided into small temporal segments around the time of each D_α peak (Fig. 4, step 3). These temporal slices must be short enough to recognize drops within 5 ms. The possibility of considering the interval $[t_D - 5, t_D + 5]$ (the times are in milliseconds) is not a good selection mainly with regard to the right limit. It is necessary to ensure the identification of the drop in the right side without possible confusion from signal noise. To this end, it has been empirically

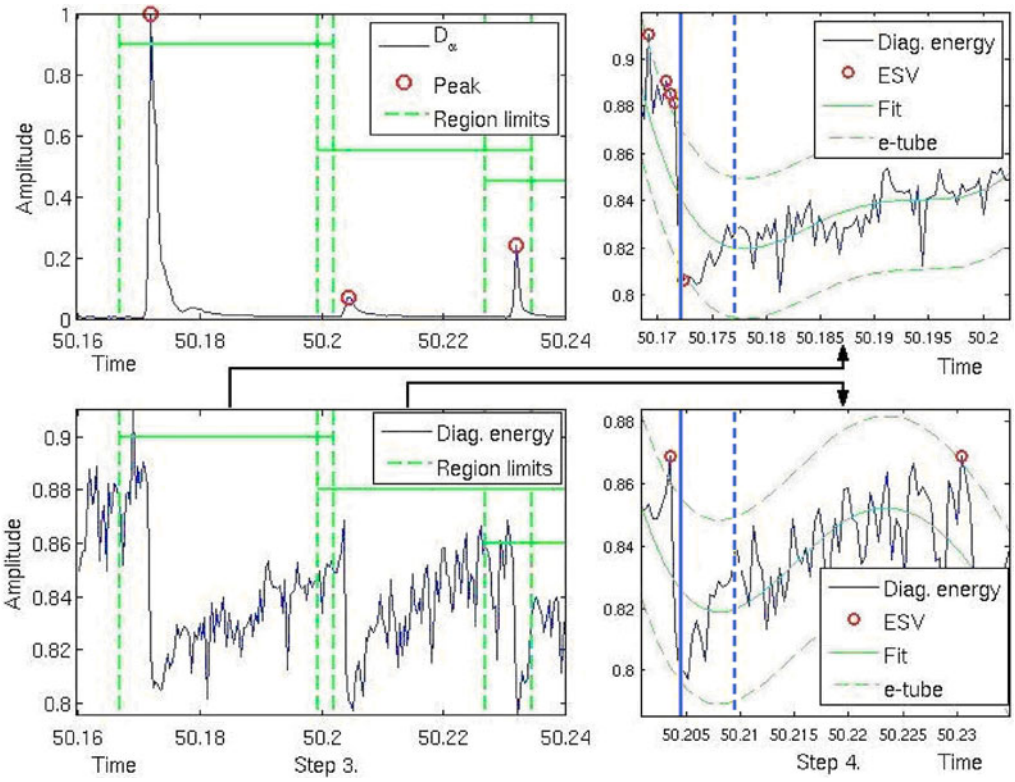

Fig. 4. Location method steps 3 and 4. This corresponds to shot 75742 of JET.

determined that 30 ms is enough for a clear recognition of drops. Therefore, the interval to look for the diamagnetic energy drop is $[t_D - 5, t_D + 30]$.

Step 4: Combination of Information

The last step of the location phase requires identifying the drop in the above interval. Again, the selection of a certain threshold to define a drop is not a practical criterion because of the presence of changing amplitudes. Once again, UMEL is used as an event locator to discover the samples in the diamagnetic energy waveform that do not fit smoothly. Thus, a SVR is performed over each interval $[t_D - 5, t_D + 30]$ (Fig. 4, step 4). If ESVs do not appear, the method determines that an ELM is not present, and therefore the peak in the D_α waveform has been generated by a different physical mechanism.

As in the D_α case, it is necessary to identify the energy drop by a unique sample. The temporal coordinate of this sample will be the temporal location assigned to the ELM. As shown in Fig. 4, UMEL determines the samples outside the *e*-tube. The one having the maximum amplitude inside the above interval determines the temporal location of the ELM. The accuracy of the prediction is the sampling period of the signal.

IV. A CASE STUDY

We have applied the ELM location method to a database of more than 1200 JET discharges in the range 73337 to 78156. A total number of 226 751 ELMs have been recognized and located in their respective discharges.

Table I shows the values of all parameters used in Sec. III. These values have been chosen to maximize the performance of the method within the considered discharge range, and it should be noted that the values could change for different discharges. However, they seem to be robust enough taking into account the big set of discharges that have been considered.

Because of the lack of a large ELM database to test the performance of the ELM location method, we have performed a manual search of individual ELMs in 20 JET discharges in the above range. This extremely tedious search has allowed obtaining a success rate of 95% in the recognition of ELMs with the automatic method based on UMEL.

Figure 5 and Table II show the number of ELMs located and the ELM period. The most common period is between 0.02 and 0.03 s with 61 373 ELMS.

TABLE I

Parameter Values for ELM Location

Parameter	Value
Preprocessing phase	
D_α wavelet descomposition level	5
K value for ELM temporal slice detection	0.5
UMEL K_c value for ELM temporal slice detection	1
UMEL K_e value for ELM temporal slice detection	10
UMEL K_σ value for ELM temporal slice detection	20
Location phase	
Step 1	
D_α wavelet decomposition level	2
UMEL K_c value for D_α peak location	1
UMEL K_e value for D_α peak location	8
UMEL K_σ value for D_α peak location	50
Step 3	
Size of diamagnetic energy temporal segments (s)	[D_α peak − 0.005, D_α peak + 0.03]
Step 4	
UMEL K_c value for diamagnetic energy event location	100
UMEL K_e value for diamagnetic energy event location	1
UMEL rbf kernel parameter σ_k for diamagnetic energy event location	100 000
D_α and diamagnetic energy peak maximum distance (s)	0.005

V. CONCLUSIONS AND FUTURE WORK

We have introduced an automatic method to locate and identify ELMs. This method uses the combination of information of the D_α emission and the stored diamagnetic energy to determine the exact temporal location of every single ELM. It uses UMEL as the event locator (a universal event locator based on the information retrieved by a SVR) to discover peaks in the D_α emission independently of the amplitudes. UMEL is applied again around the time instant of every D_α peak to discover a corresponding drop in the stored diamagnetic energy. If this drop is found, the method recognizes the presence of an ELM.

The method has been applied to a large set of JET discharges creating a database of ELMs with their temporal location. This database will allow other methods and techniques that study the ELM behavior to have available a dataset of high statistical weight. Automatic ELM classification systems, such as the one introduced in Ref. 1, can be applied to a large ELM database to classify the types of ELMs: type I, type III, compound, and so on.

The fact of using just the same software (UMEL) to process all the signals (D_α and stored diamagnetic

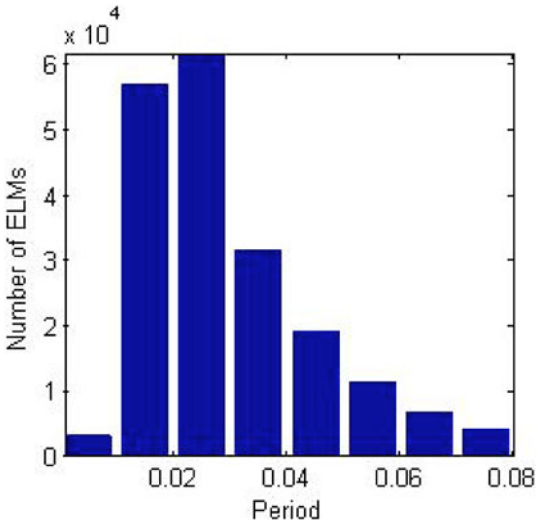

Fig. 5. Distribution of the temporal intervals of the located ELMs.

TABLE II

Number and Temporal Interval of the Located ELMs

Period (s)	Number of ELMs	Period (s)	Number of ELMs
<0.01	3 200	0.01 to 0.02	56 550
0.02 to 0.03	61 373	0.03 to 0.04	30 100
0.04 to 0.05	19 809	0.05 to 0.06	11 409
0.06 to 0.07	68 00	0.07 to 0.08	4 000

energy) is crucial in our automatic method. A unique software function to perform the regression estimations has been necessary. The method does not need to locate the D_α peaks or to locate the drops in the energy waveform with different software; i.e., the method is independent of the structural form of the signals. Even more, the software is also the same in spite of different amplitudes of a signal from shot to shot. The "universal" character of our technique is the essential feature to take advantage of in the automatic location of ELMs.

However, automatic validation methods should be developed in order to fully test the results generated by the location method based on UMEL.

ACKNOWLEDGMENTS

This work was partially funded by the Spanish Ministry of Science and Innovation under Project ENE2008-02894/FTN.

This work, supported by EURATOM, was carried out within the framework of the European Fusion Development Agreement.

REFERENCES

1. N. DURO et al., "Automated Recognition System for ELM Classification in JET," *Fusion Eng. Des.*, **84**, 712 (2009).

2. J. VEGA, A. MURARI, S. GONZALEZ, and JET-EFDA CONTRIBUTORS, "A Universal Support Vector Machines Based Method for Automatic Event Location in Waveforms and Video-Movies: Applications to Massive Nuclear Fusion Databases," *Rev. Sci. Instrum.*, **81**, 023505 (2010).

3. V. CHERKASSKY and F. MULIER, *Learning from Data*, 2nd ed., John Wiley and Sons (2007).

4. V. N. VAPNIK, *Statistical Learning Theory*, John Wiley and Sons (1998).

5. W. L. MARTINEZ and A. R. MARTINEZ, *Computational Statistics Handbook with Matlab*, Chapman and Hall/CRC Press (2002).

6. S. MALLAT, *A Wavelet Tour of Signal Processing*, 2nd ed., Academic Press, New York (2001).