Real Field Deployment of a Smart Fiber Optic Surveillance System for Pipeline Integrity Threat Detection: Architectural Issues and Blind Field Test Results

Javier Tejedor, Javier Macias-Guarasa, Member, IEEE, Hugo F. Martins, Juan Pastor-Graells, Sonia Martin-Lopez, Pedro Corredera, Guy De Pauw, Filip De Smet, Willy Postvoll, Carl H. Ahlen, and Miguel Gonzalez-Herraez

Abstract—This paper presents an on-line augmented surveillance system that aims to real time monitoring of activities along a pipeline. The system is deployed in a fully realistic scenario and exposed to real activities carried out in unknown places at unknown times within a given test time interval (so-called blind field tests). We describe the system architecture that includes specific modules to deal with the fact that continuous on-line monitoring needs to be carried out, while addressing the need of limiting the false alarms at reasonable rates. To the best of our knowledge, this is the first published work in which a pipeline integrity threat detection system is deployed in a realistic scenario (using a fiber optic along an active gas pipeline) and is thoroughly and objectively evaluated in realistic blind conditions. The system integrates two operation modes: The machine+activity identification mode identifies the machine that is carrying out a certain activity along the pipeline, and the threat detection mode directly identifies if the activity along the pipeline is a threat or not. The blind field tests are carried out in two different pipeline sections: The first section corresponds to the case where the sensor is close to the sensed area, while the second one places the sensed area about 35 km far from the sensor. Results of the machine+activity identification mode showed an average machine+activity classification rate of 46.6%. For the threat detection mode, 8 out of 10 threats were correctly detected, with only 1 false alarm appearing in a 55.5-hour sensed period.

Index Terms—Distributed fiber sensing, Acoustic sensing, Vibration sensing, Pipeline integrity, phase-sensitive OTDR, Pattern recognition

I. INTRODUCTION

Fiber optic distributed acoustic sensing (DAS) with phase-sensitive optical time-domain reflectometer (ϕ-OTDR) technology has been shown good performance for long perimeter monitorization aiming to detect intruders on the ground [1]–[5] or vibration in general [6]–[14].

Current pipeline integrity prevention systems combine DAS technology with a pattern recognition system (PRS) for continuous monitoring of potential threats to the pipeline integrity [15]–[22].

However, most of the works that employ DAS+PRS have shown significant issues with respect to the pattern classification design and evaluation, as presented in [23]. The main problems are related to the fact that no real classification is conducted nor results are given (e.g., [15], [16], [24]); not enough details on the system description or the experimental procedure are provided (e.g., [4], [16]–[18], [20]); the data generation process is far from being generated in a realistic field environment (e.g., [2], [10], [18], [24]), or this is even simulated (e.g., [7], [9], [10], [18], [25]–[29]).

Some recent works present significant improvements over those previously reported, by adopting more realistic recording environments (e.g., [17], [19]–[21], [30]), and more rigorous experimental procedures (e.g., [19], [20], [30]). However, there are still some issues related to the use of a single measurement position [19], [21], which implies a bias to recognize the position instead of the real event (whose effect was discussed in [22]), or the reduced number of the testing signals, with no additional details regarding the actual recording durations.

There are also companies that describe solutions for pipeline surveillance monitoring [16], although they do not usually provide enough details on their strategies, nor objective data for a proper assessment of their contribution.

The only exceptions that, to the best of our knowledge, aim to real field deployment of a DAS+PRS for pipeline integrity surveillance are found in [17] and [22], [23], [31], [32], but [17] does not provide enough details in its experimental procedure.

To assess the validity of a DAS+PRS that continuously monitors a long pipeline searching for potential threats and aims to real field deployment, a thorough experimental and validation procedure must be designed. This implies an approach that considers two experimental evaluation scenarios:

- An extensive off-line evaluation (that must employ recordings on realistic field data), which allows for an intensive experimental work to decide on the system design strategies and tuning.
- An (as much as possible) extensive on-line evaluation that should be based on tests carried out in real field scenarios, in unknown times, and unknown locations (so-called blind field tests). These blind field tests consist
in carrying out some activities at certain locations along
the surveillance area at certain times within a given time
interval (spanning from a few hours to days).

Regarding the extensive off-line evaluation scenario, in [22]
we presented the first published report on a pipeline integrity
threat detection system that employs DAS+PRS technology,
was evaluated on realistic field data, and whose results are
based on a rigorous experimental setup and an objective
evaluation procedure with standard and clearly defined metrics.
This work was further refined in [23], with an improved pattern
recognition strategy that led to significant performance gains.

Regarding the on-line evaluation, we presented in [32] the
first report on blind field tests of a pipeline integrity threat
detection system, which addressed the experimental and vali-
dation procedure issues with respect to real field deployment.
These field tests were carried out in realistic scenarios that
comprised different locations and soil conditions, and were
managed by Fluxys Belgium S.A., which was also responsible
for hiring the corresponding machinery and equipment to carry
out the required activities.

The pipeline integrity surveillance system consists in a
combination of hardware and software modules. The hardware
side refers to the DAS system used to record the data and the
software side refers to the pattern classification system that
classifies the acoustic data acquired by the sensing system.
Two different operation modes were set up in the pipeline
integrity surveillance system: machine+activity identification
mode, where both the machine and the activity are identified,
and threat detection mode, where just the occurrence of a
threat in the pipeline must be detected.

With respect to our previous work [32], this paper (1)
presents an augmented system with a new design that com-
pletes the baseline architecture with the required modules and
strategies to face the specific conditions of a blind field test
and discussion are presented in Section IV, and Section V
concludes this paper.

II. PIPELINE INTEGRITY THREAT DETECTION SYSTEM

The pipeline integrity threat detection system integrates
different modules, as shown in Figure 1, being an evolution of
the architecture described in [22] to consider the modifications
required for field operation in the blind field test task (new
modules are shown in bold italic font). These modules are
explained in more detail next.

A. Distributed Acoustic Sensing System

The DAS system is a commercially available φ-OTDR-
based sensor named FINDAS, manufactured and distributed
by FOCUS S.L. [33]. A detailed description of the sensing
principle and experimental setup used in the FINDAS sensor
was presented in [27].

The FINDAS has an (optical) spatial resolution of 5 meters
(readout resolution of one meter) and a typical sensing range
of up to 45 kms, using standard single-mode fiber (SMF). The
fiber scanning rate (pulse repetition rate) was of 1085 Hz, so
that the acoustic sampling frequency is also $f_s = 1085$ Hz.
The optical sampling rate used in the data acquisition system
was of 100 MHz.

B. Threat Location Preselection

Due to the fact that the system is running continuously
and that the sensed positions are in the order of thousands,
it is not possible to record all the acoustic traces along the
fiber (due to restrictions in processing times, storage, and/or
communication throughput). Therefore, a preselection of the
positions which will actually be evaluated in search of possible
threats must be carried out. In this way, only preselected traces
will be processed in the PRS side.

This selection may use a threshold-based strategy from
the energy measurements of the vibrations along the fiber,
as presented in [32]: When the energy of the vibrations
occurring in a given position of the fiber is above a predefined
threshold, an acoustic trace is recorded to indicate a possible
suspicious activity occurring at that point, and the trace is
further processed by the PRS.

The simultaneous detection of multiple activities at different
positions is also possible, by setting a different threshold for
each position. The information of how this energy threshold
profile was estimated is given in Section III-C.

C. Feature Extraction + Normalization

FINDAS is configured to record 20-second length acoustic
traces that will be sent to the feature extraction module, which
adopts a standard overlapped frame analysis approach. Signal
frames to analyze are 1-second long, and the overlap is set
to 95% to achieve a smoother change of the feature vectors.
The Fast Fourier Transform (FFT) size was set to 8192 points.
All these parameters were chosen as the optimal ones in an
extensive experimentation effort carried out in our previous
work [22].

In our current implementation, the feature extraction module
calculates a feature vector composed of the energy values
corresponding to 100 frequency bands for a 100 Hz bandwidth,
using a sensitivity-based normalization (see [22] for more
details). Again, this parameter configuration was selected as it
achieved the best results in our previous work [22].

D. Pattern Classification

The pattern classification module employs a Gaussian Mix-
ture Model (GMM)-based approach to classify each feature
vector into the most likely class (machine+activity pair in
the machine+activity identification mode, and threat/non-threat
in the threat detection mode). This employs the a posteriori
maximum probability criterion to assign the given feature
vector the class with the highest probability given by the
 corresponding GMM (see [22] for more details).
E. Result Smoothing

Given the 20-second length acoustic traces generated by the FINDAS, the frame length of 1 second and the frame overlap of 950 ms, and the fact that the classification process outputs one classification per acoustic frame, there are 415 frames (and therefore 415 classifications) in each acoustic trace.

With this setup (no matter the large frame overlap), it is highly probable that consecutive frames will be assigned different decisions (due to the characteristics of the GMM-based PRS, ‘outlier’ signal behavior, or artificial effects present when the signal was recorded), so that we decided to apply a process to smooth the actual result sequence output by the classifier.

Several strategies can be employed for the smoothing process, which are presented next.

1) Raw Smoothing: This strategy is based on substituting each frame decision (given by the corresponding class label) taking into account the decisions of the surrounding frames given a certain temporal window.

More specifically, given a temporal window of a fixed duration, this strategy replaces the recognition class given to every frame that falls in the current window by the majority class found in this window.

As an example, given the recognition sequence ‘142111211’ which corresponds to a certain temporal window of duration equals to 11 frames, the sixth frame (i.e., the one in the middle) is replaced by the majority class in this window to produce the new recognition sequence ‘14211121222’.

For frames that are at the beginning and end of an acoustic trace, the temporal window size is set to a value so that all the frames are smoothed.

3) Confusion Matrix-based Smoothing: The raw and sliding window smoothing strategies are based on the surrounding frame decisions to re-assign the decision for a given frame, without any additional information on the expected accuracy of the classification process. The rationale for the confusion matrix-based smoothing relies in the fact that performance improvements could be obtained by taking into account the possible classification errors that may exist between two given classes.

To add the information related to the classification errors, a confusion matrix that stores the probability of confusing every class with every other needs to be calculated.

The procedure used to build this confusion matrix is given in Section III.

The smoothing process consists in a dynamic programming algorithm [34] that calculates a distance of the given frame decision segment to each of the classes from the confusion matrix. The distance is computed from the minimum amount of substitutions, insertions, and deletions that are necessary to transform the input decision frames to the actual frames of each class. The algorithm next assigns all the frames of the given window the class whose distance to the input decision frames is the lowest given the confusion matrix.

In all the smoothing strategies, temporal window lengths of 1, 5, 10, and 20 seconds were evaluated.

F. Acoustic Trace Decision

After the result smoothing, the system still has 415 frame decisions for a certain acoustic trace. Therefore, to classify
this as a machine+activity pair or threat/non-threat depending on the system mode, a method to combine the 415 individual frame classifications into a single class is needed. We have adopted a majority voting scheme, where the acoustic trace is assigned the class for which more frames are classified.

G. Temporal and Spatial Smoothing

One of the main problems to face in real field deployment of a DAS+PRS is keeping false alarms under reasonable values. The fact that the DAS+PRS is run continuously, the characteristics of the sensing mechanism, and that the vibrations will affect a given length of the optical fiber being sensed, there may be possible threats that:

- Could be detected, but their duration across a given fiber length is not long enough to be considered as potential threats.
- Could be detected as independent of others, but:
  - They may be geographically close enough to each other, so that they should be considered as corresponding to the same physical event.
  - They may be temporally close enough to each other, so that they should be considered as corresponding to the same temporal event.

Specific strategies must be designed to alleviate this problem, for which our proposal contemplates two directions:

- Require the detected activities to have a minimum duration along nearby positions to be actually considered as such. Otherwise, they would be considered spurious events.
- Consider the detected activities to be corresponding to the same one, if they are separated less than a given distance within a given time interval.

H. Threat Level Assignment

In the threat detection mode of the pipeline integrity threat detection system, a method to evaluate the potential risk level of a detected threat was added to the system in order to provide additional details on the threat characteristics to the system operators (in a real world deployment). To do so, we based on the energy of the acoustic signal corresponding to the threat, so that four different risk levels were considered. In a real world deployment, this could be encoded in a color scale, for example: light blue (which represents the lowest energy threats), light green, orange, and red (which represents the highest energy threats and hence the most critical threats) levels. The energy ranges that comprise each risk level are decided dividing the energy range of the detected threats in homogeneous segments.

III. EXPERIMENTAL PROCEDURE

A. Blind Field Tests

Two different rounds of blind field tests were carried out to evaluate the system performance. Table I presents information

1These categories were agreed with the GERG members in the PTT-STOP project.
TABLE I
GENERAL INFORMATION REGARDING TO THE BLIND FIELD TESTS. ‘KMS’ STANDS FOR KILOMETERS.

<table>
<thead>
<tr>
<th>Blind field test ID</th>
<th>Distance from the sensor (kms)</th>
<th>Recording day/s</th>
<th>Recording time</th>
<th>Number of activities</th>
<th>Number of threats</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEC1</td>
<td>0.3</td>
<td>1 day</td>
<td>3 hours</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>SEC2</td>
<td>31.5-36.5</td>
<td>2 days</td>
<td>2 × 24 hours</td>
<td>12</td>
<td>4</td>
</tr>
</tbody>
</table>

TABLE II
ENVIRONMENTAL, DATE, AND DISTANCE DETAILS ON LOCATIONS WHERE TRAINING AND VALIDATION DATA RECORDINGS TOOK PLACE. ‘KMS’ STANDS FOR KILOMETERS.

<table>
<thead>
<tr>
<th>Distance from sensor (kms)</th>
<th>LOC1</th>
<th>LOC2</th>
<th>LOC3</th>
<th>LOC4</th>
<th>LOC5</th>
<th>LOC6</th>
<th>LOC7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.20</td>
<td>22.24</td>
<td>22.49</td>
<td>23.75</td>
<td>27.43</td>
<td>27.53</td>
<td>34.27</td>
<td></td>
</tr>
<tr>
<td>Concrete, wooden &amp; grass in forest</td>
<td>Grass &amp; clay in agricultural field</td>
<td>Grass in agricultural field</td>
<td>Concrete, grass &amp; clay. Next to public street &amp; private house</td>
<td>Wet clay in agricultural field</td>
<td>Clay in agricultural field</td>
<td>Grass in forest</td>
<td></td>
</tr>
<tr>
<td>Sunny/Cloudy/rainy</td>
<td>Sunny/Cloudy</td>
<td>Sunny</td>
<td>Sunny</td>
<td>Rainy</td>
<td>Cloudy</td>
<td>Sunny</td>
<td></td>
</tr>
</tbody>
</table>

TABLE III
TRAINING+VALIDATION DATA: MACHINE+ACTIVITY PAIRS WITH THEIR CORRESPONDING DURATION IN MINUTES AND THREAT/NON-THREAT LABELS (BETWEEN BRACKETS). ALL THE MACHINE+ACTIVITY PAIRS WERE EMPLOYED FOR SYSTEM TRAINING. THE MACHINE+ACTIVITY PAIRS EMPLOYED FOR VALIDATION ARE MARKED WITH ‘*’. ALL THE PAIRS EXCEPT THOSE MARKED WITH ‘*’ WERE RECORDED IN LOC1, AND THE PAIRS MARKED WITH ‘*’ WERE RECORDED IN THE REST OF THE LOCATIONS. ‘(T/NT)’ STANDS FOR (THREAT/NON-THREAT). ‘BIG EXCAVATOR’ IS A 5 TON KUBOTA. ‘MIDDLE EXCAVATOR’ IS A 4 TON KUBOTA. ‘SMALL EXCAVATOR’ IS A 1.5 TON KUBOTA.

<table>
<thead>
<tr>
<th>Machine+activity pair (duration) (T/NT)</th>
<th>Machine+activity pair (duration) (T/NT)</th>
<th>Machine+activity pair (duration) (T/NT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big excavator +unloading (4.3) (T)</td>
<td>Rocket drilling +small hitting (23.3) (T)</td>
<td>+big hitting (5.3) (T)</td>
</tr>
<tr>
<td>Big excavator +install wooden plates (47.3) (NT)</td>
<td>Rocket drilling +big drilling (31.0) (T)</td>
<td>+big excavator +metal sheet damming soil (20.7) (T)</td>
</tr>
<tr>
<td>Big excavator +movement (36.0) (NT)</td>
<td>Rocket drilling +big scraping (12.7) (T)</td>
<td>+big excavator +metal sheet damming touching dummy (8.7) (T)</td>
</tr>
<tr>
<td>Big excavator +removing trees (11.7) (T)</td>
<td>+big hitting (28.0) (T)</td>
<td>+big excavator +metal sheet damming hitting dummy (10.0) (T)</td>
</tr>
<tr>
<td>Big excavator +digging pit (30.3) (T)</td>
<td>Directional drilling +unloading (2.0) (NT)</td>
<td>+big excitator +walking (15.3) (NT)</td>
</tr>
<tr>
<td>Big excavator +install dummy (5.7) (NT)</td>
<td>Directional drilling +movement (10.3) (NT)</td>
<td>+big excitator +knocking fence door (7.0) (T)</td>
</tr>
<tr>
<td>Big excavator +insert sand in compact soil (20.0) (NT)</td>
<td>Directional drilling +small scraping (27.3) (T)</td>
<td>+big excitator +calibration (21.3) (NT)</td>
</tr>
<tr>
<td>Big excavator +install concrete plate (1.0) (NT)</td>
<td>Directional drilling +big digging (14.0) (T)</td>
<td>+big excitator +movement (1315.3) * (NT)</td>
</tr>
<tr>
<td>Big excavator +scrapping dummy (16.0) (T)</td>
<td>Directional drilling +big scraping (11.7) (T)</td>
<td>+big excitator +hitting (106) * (T)</td>
</tr>
<tr>
<td>Big excavator +knocking dummy (3.0) (T)</td>
<td>Directional drilling +small hitting (34.7) (T)</td>
<td>+big excitator +scraping (324) * (T)</td>
</tr>
<tr>
<td>Directional drilling +preparations (10.3) (T)</td>
<td>Directional drilling +small drilling (87.7) (T)</td>
<td>+big excitator +movement (610) * (NT)</td>
</tr>
<tr>
<td>Middle excavator +movement (29.7) (NT)</td>
<td>Middle excavator +small scrapping (21.0) (T)</td>
<td>+big excitator +movement (610) * (NT)</td>
</tr>
<tr>
<td>Big excavator +unloading (0.7) (NT)</td>
<td>+knocking (40.0) (T)</td>
<td>+hitting (114) * (T)</td>
</tr>
<tr>
<td>Big excavator +metal sheet damming touching dummy (8.7) (T)</td>
<td>+big excitator +scraping (260) * (T)</td>
<td></td>
</tr>
<tr>
<td>Rocket drilling +big hitting (22.0) (T)</td>
<td>Middle excavator +small scrapping (47.0) (T)</td>
<td>+big excitator +working (388) * (NT)</td>
</tr>
<tr>
<td>Rocket drilling +big scrapping (324) (T)</td>
<td>Plate compactor +big scrapping (324) (T)</td>
<td>+big excitator +working (440) * (NT)</td>
</tr>
<tr>
<td>Rocket drilling +metal sheet damming touching dummy (8.7) (T)</td>
<td>+big excitator +working (388) * (NT)</td>
<td>+big excitator +working (440) * (NT)</td>
</tr>
</tbody>
</table>

First of all, the energy threshold profile that allows FINDAS to select which acoustic traces will be processed by the PRS had to be estimated. The energy threshold values were estimated for each single fiber location (as described in Section II-B). To do so, FINDAS was set to continuously record the energy profile of the fiber segment where the blind field tests were going to be performed. The energy profile recordings were done during 3 consecutive days for the SEC1

hours in the selected day, and SEC2 tests spread continuously during two consecutive days.

C. System Configuration

There are several parameters that are needed to be set along the different modules of the DAS and pipeline integrity threat detection systems.
### TABLE IV

**Testing data:** Machine+activity pairs with their corresponding duration in minutes and threat/non-threat labels (between brackets). SEC1 refers to the first round of blind field test location, and SEC2 refers to the second one. *(T/NT)* stands for (Threat/Non-threat). ‘Big excavator’ is an 18 ton Kubota. ‘Middle excavator’ is a 4 ton Kubota. ‘Small excavator’ is a 1.5 ton Kubota. ‘Damplanken’ is a big excavator.

<table>
<thead>
<tr>
<th>Blind field test ID</th>
<th>Machine+activity pair (duration) (T/NT)</th>
<th>Machine+activity pair (duration) (T/NT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEC1</td>
<td>People+unloading rocket tools (20) (NT)</td>
<td>People+detect pipeline with hand shovels (20) (NT)</td>
</tr>
<tr>
<td></td>
<td>Big excavator+movement (7) (NT)</td>
<td>People+unloading of excavator and tools (21) (NT)</td>
</tr>
<tr>
<td></td>
<td>Big excavator+digging pit (5) (T)</td>
<td>Big excavator+hitting (1) (T)</td>
</tr>
<tr>
<td></td>
<td>Rocket drilling+small drilling (24) (T)</td>
<td>Rocket drilling+big drilling (14) (T)</td>
</tr>
<tr>
<td></td>
<td>Plate compactor+working (1) (NT)</td>
<td>Big excavator+hitting pit (19) (NT)</td>
</tr>
<tr>
<td></td>
<td>Prepare damplanken (21) (NT)</td>
<td>Damplanken+working (12) (T)</td>
</tr>
<tr>
<td></td>
<td>Small excavator+movement (5) (NT)</td>
<td>Small excavator+digging and closing pit (8) (T)</td>
</tr>
<tr>
<td>SEC2</td>
<td>Middle excavator+movement (10) (NT)</td>
<td>Truck+movement (5) (NT)</td>
</tr>
<tr>
<td></td>
<td>People+digging with shovels (11) (NT)</td>
<td>Middle excavator+digging pit (7) (T)</td>
</tr>
<tr>
<td></td>
<td>Middle excavator+hitting (1) (T)</td>
<td>Drilling+preparations (6) (NT)</td>
</tr>
<tr>
<td></td>
<td>Rocket drilling+small drilling (15) (T)</td>
<td>Rocket drilling+big drilling (17) (T)</td>
</tr>
<tr>
<td></td>
<td>Small excavator+closing pit (9) (NT)</td>
<td>Small excavator+movement (2) (NT)</td>
</tr>
<tr>
<td></td>
<td>Plate compactor+working and damming (10) (NT)</td>
<td>Drill compactor+drilling on soil and steel (6) (NT)</td>
</tr>
<tr>
<td></td>
<td>Middle excavator and truck+movement (2) (NT)</td>
<td></td>
</tr>
</tbody>
</table>

Tests and 5 consecutive days for the SEC2 tests. From these recordings, the average energy was calculated for each fiber position, and this was established as the background noise level profile.

To establish the actual energy threshold profile, the energy range at the selected position must be taken into account, considering the background noise level, and also the expected energy levels for activities at the corresponding positions (that heavily depend on the distance to the interrogator). For example, in the SEC1 tests, which took place at an average of 300 meters far from the FINDAS, the threshold energy was set to 10 times the average background noise energy level measured for each fiber location point. For the SEC2 tests, which took place at an average of 34 kms far from the FINDAS, the threshold energy was simply set to the average noise energy level. These threshold energy values were selected by varying the multiplicative factor with respect to the average background noise energy level from 1 to 40 times in the corresponding location, and choosing the factor so that all the acoustic traces of the training data have an energy above the energy threshold (no testing data were used in the threshold estimation procedures).

Regarding the pattern classification module, a single-component GMM was trained for each class.²

In the confusion matrix-based smoothing (described in Section II-E3), the confusion matrix was built using a leave-one-out cross-validation (CV) approach from all the locations except LOC1 in Table II. To do so, four locations were used for system training, one location was used to build the confusion matrix, and the other location was used for system evaluation.

In the temporal and spatial smoothing module (see Section II-G), the acoustic traces corresponding to activities that spread less than 80 seconds along 40 meters were considered spurious. In the same way, consecutive threat decisions that are separated less than 80 seconds and 40 meters from the previous threat were grouped as corresponding to the same threat. These values were tuned on validation data across different time-space configurations. To do so, the time values were swept from 10 to 130 seconds, and the space values were swept from 0 to 60 meters. Again, no testing data were used in the parameter tuning procedures.

For the threat detection mode of the system, the class (threat or non-threat) with the highest probability (generated by the GMM-based pattern classification module) is selected as the corresponding to the evaluated acoustic trace, so that there is no need to establish a threshold value to compare with.

### D. Evaluation Metrics

Classification accuracy has been the main metric to evaluate the system performance both for the machine+activity identification and threat detection modes. This is computed as the ratio between the number of correctly classified testing machines or activities, and the total number of evaluated activities. A machine+activity pair is considered to be correctly detected in case the machine or the activity output by the system coincides with that of the ground truth and is within the temporal limits of the activity ground-truth time span. In the same way, a threat is correctly detected in case the system coincides with that of the ground truth temporal limits.

Additionally, for the threat detection mode, the Threat Detection Rate (TDR), which corresponds to the percentage of threat testing activities that are classified as threat, usually referred to as true positives, and the False Alarm Rate (FAR), which corresponds to the percentage of non-threat testing activities that are classified as threats, usually referred to as false positives, were also calculated.

### IV. Experimental Results

#### A. Performance of the Result Smoothing Module

As described in Section II-E, three different strategies were examined to improve the system performance at frame level. However, during the blind field tests, one of these strategies must be chosen in the system pipeline.

²Training multi-dimensional GMMs was also tested, but the system performance was lower, probably due to data scarcity for certain machine+activity pairs.
Validation data were employed to select the optimal strategy. To do so, and since these data were recorded in 6 different locations (all except LOC1 in Table II), we applied a CV approach in a location basis (more details can be found in Section IV.D of [22]). This CV slightly varies depending on the smoothing strategy. For the raw smoothing and sliding window smoothing strategies, 5 locations were used for GMM training and the other location was employed as test to form 6 different folds. The different window sizes were examined for each fold, and the average performance from the results obtained in each fold and window size was computed. For the confusion matrix-based smoothing strategy, 4 locations were used for GMM training, one location was used to build the confusion matrix, and the other for testing. This derives in 30 folds. As in the other strategies, the average performance is obtained by averaging the performance for each fold.

The smoothing results are presented in Table V for the machine+activity identification and threat detection modes, compared with the ones obtained without smoothing. These results show that longer window sizes are better than shorter window sizes. Since more information (decisions of more frames) is considered for the decision, and all the frames that comprise an acoustic trace belong to a same activity, better performance is expected.

The best result is obtained by the raw smoothing strategy with the 20-second length window for almost all the metrics. For the machine+activity classification accuracy (MAC accuracy in Table V), the raw smoothing and sliding window smoothing strategies obtain similar performance, being the former slightly better than the latter. On the other hand, the confusion matrix-based smoothing strategy obtains the worst results, probably due to data scarcity issues.

For the threat classification accuracy (TC accuracy in Table V), the confusion matrix-based smoothing strategy obtains a slightly better overall performance. However, in this system mode, the TDR and the FAR are more adequate to decide which strategy is better, so that we can see that the raw smoothing strategy obtains better TDR. Since detecting as many threats as possible is crucial when building this kind of systems, we have selected the raw smoothing strategy with window size equals to 20 seconds for the real field deployment tests.

B. Time-Location Based Analysis

Preliminary analysis was carried out to get an initial idea on the feasibility of the pipeline integrity surveillance system for detecting suspicious activities in the surveillance zones corresponding to the two rounds of blind field tests. This analysis consists in a time-location sensed energy representation of the surveilled zone that aims to be monitored in each blind field test. Figure 2 shows the time-location energy representation for the SEC1 blind field tests, where the horizontal axis shows the time, and the vertical axis shows the distance at which a given energy value (in a thermal color scale from blue to red) was measured. We do not show the full time-location energy representation for the SEC2 blind field tests, as it was mostly empty, given that the test was only done during 4 hours of the day 2 of the tests. Instead, we show Figure 3, which is a zoomed in version of the time-location energy representation of the SEC2 blind field tests, centered in the time-location area where the test activities were carried out (around 34.4 km from the sensor and between 9:00 and 13:00 of the second recording day).

These figures show that all activities were detected by the FINDAS in both rounds of blind field tests, and they turned out to be at the correct times when the activities were being carried out. This means that the energy-based threshold estimation procedure was highly accurate. In the blind field test location near the FINDAS (Figure 2), the time-location energy representation reveals clearer information, due to the higher SNR in the data acquisition. In the case of the SEC2 test, where the sensed location is 34 kms far from the sensor, much lower energy values are reported, which derives in a much more difficult activity detection. However, the figure shows that the system was able to detect all the activities.

C. Blind Field Test System Results

After verifying that all the activities were correctly detected by the FINDAS, the system performance for the blind field tests in terms of quantitative rates regarding to the classification itself are presented in Table VI.

These results show that, as expected, SEC1 obtains better performance than SEC2, due to the favorable acoustic signal conditions. It can also be seen that almost all the threats are correctly detected, which suggests that the system is suitable for real field deployment.

With respect to the identification of the machine+activity pair, the performance drops. On the one hand, this is due to the very high number of classes involved in the system (45 in our case). Nevertheless, the average MAC accuracy is 46.6%, much higher than the pure chance rate (11% given that there are 10 different machines and 31 different activities), which is promising given the extreme difficulty of the task. This performance for the machine+activity classification task is comparable with the 54.92% presented in our previous study for offline tests described in [23], in which a much simpler experimental setup was used (4 different machines and 8 different activities, with a 12.5% pure chance rate).

On the other hand, a detailed analysis of the errors in the machine+activity identification mode reveals interesting findings: For SEC1, a performance loss of 26.7% is due to the fact that there was no trained model for the performed activity. The other 20% drop corresponds to machine+activity pairs that have been incorrectly classified by the system. For SEC2, a performance loss of 30.7% is again due to the unavailability of a trained model for the performed activity, being the other 30.7% performance drop due to actual classification errors. The reason why there are some activities carried out in the blind field tests that do not have a corresponding model trained, is that the GERG members wanted to evaluate what would be the system response in such cases (i.e., facing an unknown activity). The training and validation data sets (used for system training and building) were recorded before the blind field tests were carried out. Therefore, the non-previously trained activities used in the blind tests could
In both SEC1 and SEC2 blind field tests, the classification errors come from different sources: (1) the machines carried out the activities during certain times that did not fulfill the temporal restriction in the temporal and spatial smooth module, (2) confusions due to similarity in activities such as hitting or scrapping (scrapping sometimes includes some hitting), and (3) overlapping of different activities in a single one. In addition, the different (unknown) testing locations along the pipeline imply different acoustic conditions (soil, weather, etc.) in the acquired acoustic traces, and this variability affects to a great extent the system performance, as was already discussed in our previous publications (c.f. Section IV.C of [22]). Our strategy partially alleviates the issue related to acoustic variability in unknown environments, by first obtaining training data from as many different locations, days, and times as possible (so that the acoustic models can learn and generalize such variability); then, by applying specific normalization approaches to the acoustic traces; and finally, by employing smoothing procedures to the obtained results.

Since this is an on-line system that continuously monitors potential threats for the pipeline, real time response is crucial. All the system modules except the Pattern Classification module have a computational cost that does not depend on the number of models. For the latter module, more models increase the response time as the acoustic trace has to be compared with all the trained models. Experiments were run

---

**TABLE V**

RESULTS OF THE DIFFERENT SMOOTHING STRATEGIES FOR MACHINE+ACTIVITY IDENTIFICATION AND THREAT DETECTION MODES WITH THE BEST RESULT IN BOLD FONT. ‘MAC’ STANDS FOR MACHINE+ACTIVITY CLASSIFICATION, ‘TC’ FOR THREAT CLASSIFICATION, ‘TDR’ FOR THREAT DETECTION RATE, AND ‘FAR’ FOR FALSE ALARM RATE.

<table>
<thead>
<tr>
<th>Smoothing type</th>
<th>Window length (seconds)</th>
<th>MAC accuracy</th>
<th>TC accuracy</th>
<th>TDR</th>
<th>FAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>-</td>
<td>45.2% ± 0.11%</td>
<td>64.3% ± 0.11%</td>
<td>80.7% ± 0.09%</td>
<td>40.3% ± 0.11%</td>
</tr>
<tr>
<td>Raw</td>
<td>1</td>
<td>50.6% ± 0.11%</td>
<td>65.4% ± 0.11%</td>
<td>84.5% ± 0.08%</td>
<td>40.3% ± 0.11%</td>
</tr>
<tr>
<td>Raw</td>
<td>5</td>
<td>56.9% ± 0.11%</td>
<td>67.1% ± 0.11%</td>
<td>88.9% ± 0.07%</td>
<td>39.2% ± 0.11%</td>
</tr>
<tr>
<td>Raw</td>
<td>10</td>
<td>59.6% ± 0.11%</td>
<td>68.2% ± 0.11%</td>
<td>91.0% ± 0.07%</td>
<td>38.4% ± 0.11%</td>
</tr>
<tr>
<td>Raw</td>
<td>20</td>
<td>61.9% ± 0.11%</td>
<td>69.2% ± 0.11%</td>
<td>92.0% ± 0.06%</td>
<td>37.5% ± 0.11%</td>
</tr>
<tr>
<td>Sliding window</td>
<td>1</td>
<td>50.4% ± 0.11%</td>
<td>65.4% ± 0.11%</td>
<td>84.4% ± 0.08%</td>
<td>40.2% ± 0.11%</td>
</tr>
<tr>
<td>Sliding window</td>
<td>5</td>
<td>56.8% ± 0.11%</td>
<td>67.2% ± 0.11%</td>
<td>88.9% ± 0.07%</td>
<td>39.1% ± 0.11%</td>
</tr>
<tr>
<td>Sliding window</td>
<td>10</td>
<td>61.3% ± 0.11%</td>
<td>68.8% ± 0.11%</td>
<td>91.3% ± 0.06%</td>
<td>37.7% ± 0.11%</td>
</tr>
<tr>
<td>Sliding window</td>
<td>20</td>
<td>49.9% ± 0.11%</td>
<td>65.9% ± 0.11%</td>
<td>81.2% ± 0.09%</td>
<td>39.1% ± 0.11%</td>
</tr>
<tr>
<td>Confusion matrix</td>
<td>1</td>
<td>55.4% ± 0.11%</td>
<td>68.0% ± 0.11%</td>
<td>87.2% ± 0.08%</td>
<td>37.7% ± 0.11%</td>
</tr>
<tr>
<td>Confusion matrix</td>
<td>5</td>
<td>57.7% ± 0.11%</td>
<td>68.7% ± 0.10%</td>
<td>89.1% ± 0.07%</td>
<td>37.3% ± 0.11%</td>
</tr>
<tr>
<td>Confusion matrix</td>
<td>10</td>
<td>59.6% ± 0.11%</td>
<td>69.7% ± 0.10%</td>
<td>90.4% ± 0.07%</td>
<td>36.3% ± 0.11%</td>
</tr>
</tbody>
</table>

---

**TABLE VI**

BLIND FIELD TEST SYSTEM RESULTS. ‘MAC’ STANDS FOR MACHINE+ACTIVITY CLASSIFICATION, ‘ACC.’ FOR ACCURACY, ‘TC’ FOR THREAT CLASSIFICATION, ‘TDR’ FOR THREAT DETECTION RATE, AND ‘FAR’ FOR FALSE ALARM RATE.

<table>
<thead>
<tr>
<th>Blind field test ID</th>
<th>MAC acc.</th>
<th>TC acc.</th>
<th>TDR</th>
<th>FAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEC1</td>
<td>53.3%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>SEC2</td>
<td>38.5%</td>
<td>76.9%</td>
<td>50.0%</td>
<td>25.0%</td>
</tr>
<tr>
<td>Average</td>
<td>46.6%</td>
<td>88.9%</td>
<td>80.0%</td>
<td>10.0%</td>
</tr>
</tbody>
</table>

---

not (and must not) be included in the system training. The activities for which there is no trained model were: People+unloading rocket tools, People+detect pipeline with hand shovels, People+unloading of excavator and tools, Big excavator+filling pit, Prepare damplanken, Small excavator+digging and closing pit, Truck+movement, People+digging with shovels, Drilling+preparations, Small excavator+closing pit, and Drill compactor+drilling on soil and steel. The performance drop observed for activities with unseen data in the training stage is obviously expected, and proves the need of addressing a data acquisition effort as wide as possible, as stated in [35].
on an Intel Quad Q9550 2.83GHz processor and 4GB RAM. Table VII shows the computational time of the system modules. It can be seen that the system is able to run on real-time, since the time response is less than a second in any mode. The table also shows that each model added to the system increases the computational time in 10 ms. This means that, after the 20-second length acoustic trace has been recorded by the FINDAS, the decision assigned to the suspicious acoustic trace is given in 0.69 seconds in the machine+activity identification mode, and 0.26 seconds in the threat detection mode.

The information about the threat level assignment discussed in Section II-H was not used in the system result evaluation (as we were only interested in assessing the threat detection accuracy). The threat level calculation is meant to be exploited in future research.

As a final comment, the blind field test evaluation procedures will face, in most cases, the problem of the statistical significance of the results. In cases such as the one described in this paper, the number of blind field tests that can be performed will be much lower than the tests conducted during the off-line database evaluation process. Therefore, the reported results in the off-line evaluation will have a higher statistical significance than those from the blind field tests. This must be taken into account to put the blind field test results in an adequate context, and also to drive the design of a broad enough blind field test campaign.

V. CONCLUSIONS AND FUTURE WORK

We have presented an evolution of the systems presented in [22] and [32] that is able to continuously monitor potential threats to the pipeline integrity in real field blind conditions. On-line pipeline integrity monitoring systems have to resort to a series of blind field tests to clearly assess their performance when facing at real field deployment. For this, two different rounds of blind field tests were conducted in different days and locations, which convey different acoustic signal properties each. The blind field test locations are placed near the sensing system (300 meters far) and far the sensing system (35 kilometers far). Results show good performance in terms of threat detection, since 8 out of 10 threats were correctly detected, and just 1 false alarm in 55.5 hours was generated. In terms of machine+activity classification, the average performance is 46.6%, which is still well above chance (11%, given the 10 potential machines and 31 potential activities the system is able to detect). Results also show degradation as long as the distance to the sensing system increases, both in terms of machine+activity classification and threat detection rates.

Future work should address the signal degradation issue when the distance to the sensor grows. In addition, other feature extraction techniques (e.g., those based on wavelet transforms that take advantage of time and frequency domains simultaneously) and pattern classification techniques (e.g., combination of GMM and neural networks) must also be examined.

ACKNOWLEDGMENTS

Some authors were supported by funding from the European Research Council through Starting Grant UFINE (grant number #307441), Water JPI, the WaterWorks2014 Cofunded Call, the European Commission (Horizon 2020) through project H2020-MSCA-ITN-2016/722509 - FINESSE, the Spanish Ministry of Economy and Competitivy, the Spanish “Plan Nacional de I+D+i” through projects TEC2013-45265-R, TEC2015-71127-C2-2-R, TIN2013-47630-C2-1-R, and TIN2016-75982-C2-1-R, and the regional program SINFOTONCM: S2013/MIT-2790 funded by the “Comunidad de Madrid”. HFM acknowledges funding through the FP7 ITN ICONE program, grant number #608099 funded by the European Commission. JPG acknowledges funding from the Spanish Ministry of Economy and Competitivity through an FPI contract. SML acknowledges funding from the Spanish Ministry of Science and Innovation through a “Ramón y Cajal” contract.

REFERENCES

Table VII

<table>
<thead>
<tr>
<th>System mode</th>
<th>Feature Extraction + Normalization</th>
<th>Pattern Classification</th>
<th>Rest of Modules</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine-Activity identification</td>
<td>140</td>
<td></td>
<td>450 (45 models)</td>
<td>100</td>
</tr>
<tr>
<td>threat detection</td>
<td>140</td>
<td></td>
<td>20 (2 models)</td>
<td>100</td>
</tr>
</tbody>
</table>


