Engineering and Deploying a Hardware and Software Platform to Collect and Label Non-Intrusive Load Monitoring Datasets

Lucas Pereira
M-ITI/LARSYS and
prsma.com
Funchal, Portugal
lucas.pereira@m-iti.org

Miguel Ribeiro
M-ITI/LARSYS and
Técnico, U. Lisboa
Funchal, Portugal
jose.ribeiro@m-iti.org

Nuno Nunes
M-ITI/LARSYS and
Técnico, U. Lisboa
Funchal, Portugal
njn@m-iti.org

***Abstract*—Current approaches for collecting and labeling Non-Intrusive Load Monitoring (NILM) datasets still rely heavily on a lengthy and error prone manual inspection of the whole dataset. Consequently, it is still difficult to find fully labeled datasets that could help furthering even more the research in this field. In an attempt to overcome this situation, we propose a hardware and software platform to collect and label NILM sensor data in a semi-automatic labeling fashion. Our platform combines aggregate and plug-level smart-meters to measure consumption data, software algorithms to automatically detect changes in the different monitored loads and a graphical user interface where the end-user can supervise the labeling process. In this paper, we describe the different components that comprise our platform. We also present the results of one live deployment that was performed to test the feasibility of our approach. The results of the deployment show that our system was capable of explaining about 82% of the aggregate load, and automatically detect 94% of the power transitions in the plug-level loads.**

***Index Terms*—Hardware-software platform, semi-automatic labeling, intelligent user interface, datasets, non-intrusive load monitoring, event detection.**

I. Introduction

Environmental sustainability is one of the fastest growing areas of activity in several research fields like Ubiquitous Computing (UbiComp) and Human Computer Interaction (HCI). In part, this reflects the observation that pervasive technologies can provide a platform for reflection and intervention with positive social benefits. Thus, many researchers began examining the opportunities to use interactive technologies to promote environmental sustainability and, in particular, sensing and context aware applications for energy monitoring and eco-feedback.

From the early expectations that interactive technologies could make a significant impact in promoting a more sustainable future, the community is currently facing several engineering challenges. Like in any other engineering domain, the transition from research prototypes to large-scale deployment of sensing and interactive infrastructures is creating substantial research challenges that have not been foreseen by the stakeholders.

Within the energy and environment research community energy disaggregation is a promising approach. One particular technology of interest is Non-Intrusive Load Monitoring (NILM) [1], since it enables a cost effective way of providing detailed information about energy consumption from a single sensing point. Research in this field aims at disaggregating and estimating the consumption of individual appliances by means of applying machine learning techniques on top of the aggregated consumption signal. Together with the availability of public datasets this approach is gaining traction as researchers create more systematic evaluation processes that can be used effectively. This is aligned with what happened in other machine learning application domains like, for example, face and motion recognition.

To date NILM research is formally categorized according to two different approaches: i) event-based approaches, which consist of keeping track of every appliance state transition (e.g. TV turning ON or OFF) by means of event detection and classification, assuming that the system was previously trained [1], [2], and ii) event-less approaches, where no previous knowledge of the existing appliances is assumed and the load disaggregation is done by means of techniques like Hidden Markov models or temporal motif mining [3], [4]. Extensive reviews of the existing approaches for single point energy disaggregation can be found in [5], [6]. Here we focus on the challenges posed by the need to create and interact with public datasets for assessing the performance of the different NILM algorithms.

In this paper, we argue that one of the main reasons behind the shortage of fully labeled datasets to evaluate and benchmark both event-based and event-less approaches is the difficulty in labeling the acquired data. On the one hand we cannot rely only on automatic labeling processes, as these methods still do not guarantee enough accuracy. On the other hand, we cannot solely rely on humans to label the data since this is a very time consuming processes that is also very prone to mistakes. Against this background, in this paper, we propose one attempt to overcome this problem, which could be generalized to other datasets from different domains. We

978-3-901882-99-9

describe a proposal for a novel hardware and software platform to collect and label NILM datasets, following a semi-automatic labeling approach.

The remaining of this paper is organized as follows: first we describe the relevant state of the art and thoroughly describe our hardware and software platform. Then we present the results of a pilot study that was done to test the feasibility of the proposed approach. Finally, we conclude and outline future work.

II. State of the Art

A NILM dataset is a collection of electric energy measurements taken from buildings in real world conditions. These usually contain measurements from the whole-house consumption (taken at the mains) and of the individual loads (i.e., ground-truth data). Individual loads are obtained either by measuring each load at the plug-level or measuring the individual circuit where the load is connected.

Like NILM approaches, the currently available datasets are also categorized by event-based or event-less. The major difference between the two categories lies in the fact that the latter does not require the identification the appliances responsible for the different changes in the aggregate power (also known as power events). Consequently, collecting datasets for event-less approaches is more straightforward and less time consuming, which in part explains the higher availability of this kind of datasets [7].

Among the former category, are the Reference Energy Disaggregation Dataset (REDD) [8], the Almanac of Minutely Power dataset (AMPds) [9], the Indian Dataset for Ambient Water and Energy (iAWE) [10] and the "UK-Dale" [11], all providing aggregate and individual appliance consumption data. As for the latter category, only the Building-Level fUlly-labeled dataset for Electricity Disaggregation (BLUED) [12] is currently available.

Semi-automatic labeling of sensor data is not a new topic and it was already attempted in other machine-learning domains including context aware driving [13], image segmentation [14] and video annotations [15]. In the field of energy disaggregation, with the exception of [16], this is the only attempt to label ground-truth data in semi-automatic fashion. In [17] Cao et al. also propose a framework for annotating energy datasets, however their approach differs from ours in two fundamental aspects: i) the users are requested to annotate active states (i.e., periods when appliances are in use) instead of power events, and ii) the annotation process is fully manual.

III. Hardware and Software Platform

The underlying platform for our NILM research is based on low cost technologies. On the hardware side, a bespoke smart-meter is combined with a commercially available plug-level energy monitoring system to collect aggregated and individual appliance consumption data. On the software side, a web-based application was developed to support the semi-automatic labeling of the individual appliances data. In the following we discuss the implementation details of our system.

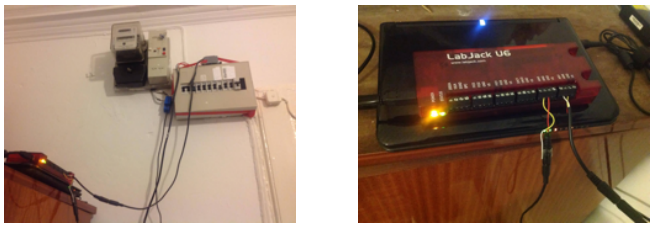

Fig. 1. Whole-house monitoring hardware installed in the main breaker box: general overview (left), DAQ system (right).

A. Whole-House Data Collection Setup

To collect data about the whole-house and individual circuits consumption we created our own energy monitoring setup by extending the multi-house energy monitoring and eco-feedback platform described in [7]. The setup consists of a multi-channel data acquisition board (LabJack U6[1]), one processing unit (Toshiba NB300), and a combination of current transformers (CT) and voltage transformers (VT). The number of channels to sample simultaneously and the desired sampling rate varies according to the number of current and voltage sensors. For example, the U6 DAQ takes 14 analog input channels with 16-bit resolution and a maximum sampling rate of 50 kHz, which is enough to measure the whole house consumption (current and voltage) and 12 individual circuits (current) with a maximum sampling rate of 3.2 kHz. Figure 1 shows an example of how the system looks like once installed in the main breaker box of a house.

In this setup, the sampled current and voltage waveforms are stored in the EMD-DF file format [18] in one-hour files. The EMD-DF is a common file format and API to create and maintain energy disaggregation datasets. The proposed file format is an extension of the Waveform Audio File Format (WAVE) and enables the storage of different power demand metrics (e.g. apparent, real and reactive power) along with individual appliance activity (power events), groups of appliance activities (e.g. combine all the clothes-washer, dryer and iron individual activities to form the laundry user activity) and other relevant metadata in a single compact file.

In order to minimize the effects of synchronization issues that may occur due to the differences in the internal clocks[2] of the data acquisition devices, we decided to perform a hardware-timed data acquisition (i.e., the DAQ hardware is responsible for acquiring the requested number of samples) and a software-timed storage of the data (i.e., a new file is always created exactly one hour after the previous one, independently of the number of samples that has been stored).

Finally, since the EMD-DF format only supports real values between -1 and 1, the current and voltage measurements are scaled before being stored. To state more precisely, each channel is scaled according to the maximum expected value (plus a confidence interval) that varies according to the sen-

[1]LabJack U6 DAQ, http://www.labjack.com/U6

[2]LabJack Forums, https://goo.gl/GTMp9Y

sor characteristics. An example of the scaling procedure in provided in sub-section IV-A.

B. Plug-Level Data Collection Setup

For the appliance-level data collection we used the Plugwise[3] system (this was also used in [19], [20]). This plug level system is a commercially available, distributed sub-metering platform that consists of three main components namely the circle (also called module), the stick and the *source* software.

Each circle is connected between the appliance being measured and the power outlet. The stick connects (using the ZigBee[4] protocol) the deployed circles to a computer that processes and displays the consumption of each individual appliance using the *source* software. This software aggregates the individual appliance measurements (e.g., by hour, day or month) and generates multiple consumption reports to the end-users.

This is however limited when we consider the level of granularity required to label a dataset for event-based NILM, e.g., knowing if an appliance is ON or OFF and when the transitions happen. Consequently, we developed our own monitoring software by extending an open-source Plugwise python[5] package that enables direct access to the raw measurements of each module.

The original version of this package maintains a hard-coded list with the mac-address of the modules to be monitored and sequentially scans each of them in 10 seconds intervals. When a given module does not reply within six attempts (each failed attempt takes about one second) an exception in thrown and the system moves to the next plug in the list. Finally, the obtained measurements (timestamp and power) for each plug are continuously stored in individual comma-separated-value (CSV) files, created daily at 12 AM.

This solution presents some considerable caveats when collecting NILM datasets, for instance: i) in the best scenario (when all the modules are accessible), individual appliance measurements are only available once every 10 seconds, and ii) the sampling interval increases by six seconds for each module that happens to be out of range or offline. Against this background, we modified the original scripts to make the data collection process more suitable for NILM datasets. The following changes were made:

- The 10-second interval between scans was set to zero, meaning that a new scanning round-trip begins right after the previous one is concluded. Our tests have shown that it takes about one second to scan 10 plugs (1 Hz) and two seconds for 19 plugs (0.5 Hz), when all the plugs are online;
- We added the possibility of enabling and disabling modules in runtime. For example, whenever a module is not plugged to a socket (e.g., a vacuum cleaner is normally stored in a closet) it can be disabled, thus avoiding unnecessary scans to disconnected devices and the consequent delays when scanning the remaining modules;
- We added the possibility of renaming modules in runtime. For example, if an appliance is replaced during the deployment it is possible to reflect that change in the dataset;
- We developed a daemon that reads the raw CSV files and pushes the measurements to a database. This daemon was set to run once per minute (1/60 Hz), but if needed it is possible to have it running at least every second (1 Hz).

[3]Plugwise, https://www.plugwise.com/

[4]Zigbee, http://www.zigbee.org

[5]Plugwise Python, https://github.com/SevenW/Plugwise-2-py

C. Semi-Automatic Labeling Application

The core of our platform is the semi-automatic labeling application that is divided in two major tasks: i) the automatic event detection, and ii) the human supervision. The automatic event detection relies on an event detection algorithm, that attempts to locate the power events in the ground-truth data. These events are then presented to the end-user in a graphical user interface where the human supervision task happens. In the following we describe the event detection algorithm and the semi-automatic labeling application.

1) Event Detection Algorithm: Our event detection algorithm is a modified version of the expert heuristic event detector presented in [21]. It works with one sliding window (detection window) of length L, that is used to scan the power signal looking for changes in the signal (DP) that are above a pre-defined power threshold (PT) in absolute value. The value of L depends on the number of samples to be averaged before and after the sample of interest, i.e., $L = L_{pre} + L_{post} + 1$ where L_{pre} and L_{post} are the number of samples before and after the sample of interest, respectively.

The algorithm works as follows: for each power sample, the amount of power change DP is calculated by subtracting the average power before, from the average power after that sample. Then, in the second step, samples with a DP above the predefined threshold PT (in absolute value) are flagged as possible power events. Finally, in the third step, the flagged power events that are separated by at least TS seconds are confirmed as events, whereas the others are discarded.

Overall, this algorithm has four parameters, L_{pre}, L_{post}, PT and TS. These can be changed individually for each appliance per its characteristics, thus increasing the event detector accuracy.

2) Semi-Automatic Labeling Interface: In the semi-automatic labeling user interface, the power measurements and the power events are plotted together in separate series (see figure 2) offering to the user the possibility of supervising the labeling process. The power samples are represented in a area series, whereas the power events are represented in column series.

The semi-automatic labeling follows a two-step process, namely: i) loading and plotting the power measurements, and ii) the event supervision. Next, we describe the two steps in mode detail.

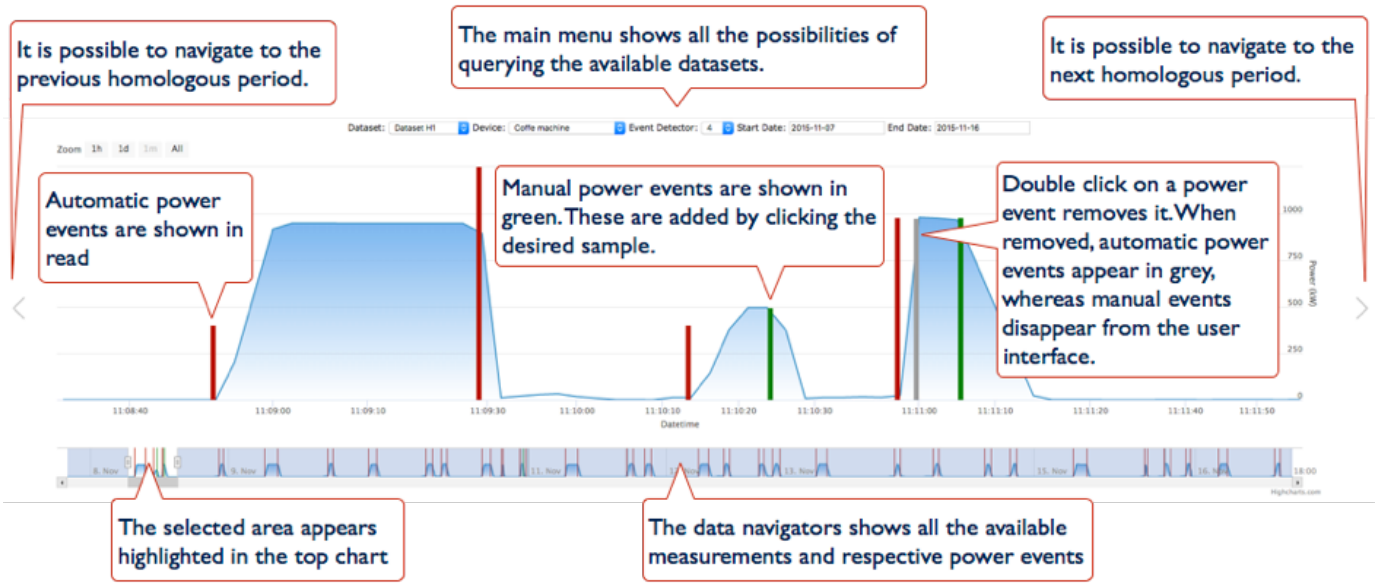

Fig. 2. User interface of the semi-automatic labeling application. Power samples are represented by an area series, whereas the power events are represented as column series.

Loading power measurements: A system like this one generates a large amount of data. For example, with two-second read intervals one day of measurements represents 43.2k samples. Furthermore, since most of those samples refer to steady-state periods (i.e., continuous periods when no relevant power changes are observed), much of the sampled data has similar values, and therefore are not necessary in the user interface. Consequently, in order to keep the amount of data to be transferred and presented to the user to a minimum, we implemented a steady-state detection algorithm (SSD) to automatically remove such periods. This allows the data to be transferred and rendered in the semi-automatic labeling application much faster, and allows a more fluid user interaction due to the reduced number of samples to be computed.

Our SSD algorithm works by dividing the power signal in W windows of length L, and removing the samples from the windows with a standard deviation bellow a pre-defined threshold (SDT). Whenever a window has a standard deviation above the pre-defined threshold, the samples in the previous and the next window are kept to make sure that steady state areas are correctly represented in the visualization. Lastly, the data from the fist and last windows are always kept, to guarantee that the time span represented in the visualization is the same as in the original data.

Overall, this algorithm contains two parameters, L and SDT that were experimentally set to one minute (30 samples) and 10% of the most represented power value above 50 Watts, respectively. In the case of appliances that consume less than 50 Watts, the SDT was set to 10% of the difference between the maximum and minimum observed power values.

In figure 3 we show the results of applying the SSD algorithm to one day of consumption from a dishwasher. On the top, we show the original consumption data with a total of 30.004k samples. In the center, we depict the original data zoomed to the area where the dishwasher is active. Finally, in the bottom we show the data returned by the SSD algorithm, with a total of 374 samples, which is less than 2% of the original data.

As it can be observed, the application of a very simple SSD algorithm considerably reduces the overhead in data transmission, rendering and navigation, while keeping the original data in the areas where the appliances are active. We should remark that this last point is of crucial importance in our case, since we want the labels to be as close as possible

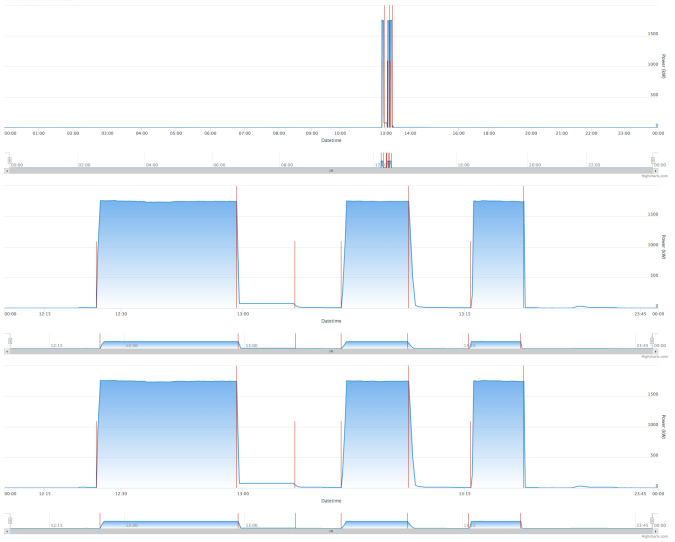

Fig. 3. Example of the SSD algorithm application: original data (top), original data zoomed (center), output data (bottom).

to the actual time of occurrence.

Event supervision: During the supervision process, users can delete the automatically generated labels by double-clicking on the respective column in the chart. Users can also add and remove labels in runtime by clicking the desired power measurement or double-clicking the created labels.

Deleting a label that was created by the user completely removes it from the chart. Removing a system labels will only change its color, hence offering the user the possibility of reactivating that label. All the generated labels (system and user) are kept in a database independently of being active or not in the user interface. Furthermore, the number of times that a label is created and / or removed is also recorded in the database, for future data analysis.

Finally, we added additional mechanisms to make the navigation in the data smoother: i) we allow the data to be filtered by time intervals, ii) we allow users to automatically move to the next our previous analogous time interval, iii) we provide a data navigator, where the users can get a sense of the available data, and iv) we added the possibility of automatically setting different zoom levels, hence allowing faster navigation to the desired data points. Figure 2 shows an annotated screenshot of the semi-automatic labeling application.

IV. Pilot Deployment

To understand the feasibility of our proposed platform we conducted a 10-day long pilot study in one household with four residents (two adults and two children). Next we describe the deployment setup and the evaluation methodology.

A. Deployment Setup

The monitored house is an apartment from the early 2000s and is comprised of seven divisions. The collected measurements include aggregated (whole-house) consumption and the individual consumption of 17 appliances, as listed in table 1 (columns 1 to 3).

The whole-house monitor (computer, DAQ and sensors) and the plugwise system (computer with the USB stick) were installed in the living room where the main breaker box is located. The monitored appliances were distributed across 6 rooms: kitchen (7 appliances), living room (3), laundry room (3), bedroom I (1), bedroom II (2) and bedroom III (1). Some devices like lamps, hard-wired appliances (e.g., ceiling lights) and miscellaneous chargers (e.g., cell phones and tablets) were not individually monitored, either because they cannot be monitored using our system (ceiling lights) or they don't draw enough power (less than 30 Watts) to justify fully labeling the dataset with their power changes (the miscellaneous chargers).

The current waveforms were sensed using a 30 A to 1 V current transformer with a peak instantaneous voltage of about 1.47 V. Therefore, to avoid clipping when storing in the EMD-DF format, the data was scaled by a factor of 1.5 V (1. 47 V $\pm$ 5 % margin). The voltage waveforms were sensed using a 230 V to 0.5 V voltage transformer with a peak voltage of 0.7 V. As such, no scaling was necessary. As for the ground-truth data, our system registered a new power measurement roughly every two seconds (0.5 Hz).

B. Evaluation Methodology

After deploying our system, we conducted an evaluation that looked at the efficiency of the hardware and software platform to explain the aggregate consumption, and also how to fine tune the event detector parameters to each individual load. In the following sub-sections we describe the evaluation methodology.

1) Individual Appliance Monitoring: First, we evaluated the efficiency of the system in monitoring the consumption of the individual appliances. To this end, we look at the number of samples obtained for each appliance, the time between consecutive measurements and the percentage of energy explained (i.e., the % of aggregate energy for which there is appliance-level consumption data).

2) Event Detection: Regarding the event detection, our goal was to find the best parameter combinations, i.e., the best combination of L_{pre}, L_{post}, PT , and TS, for each individual appliance. This is done following a threefold approach: i) manually label the power events of each appliance using the semi-automatic labeling interface in order to collect ground-truth data, ii) execute a parameter sweep of the event detection algorithm against each appliance to get the different event detection results, and iii) evaluate the performance of the parameter sweep results against the manually labeled power events. Next, each step is described in detail.

Manual labeling of power events. To collect ground-truth of the power events we first had to manually label the power events using the interface described in the previous section. We should remark that we could have used the semi-automatic labeling to provide the ground-truth labels, but we wanted to make sure that all the labels were only from manual inspection of the data, otherwise the position of the labels could be biased by the output of the different models.

The number of power events for each appliance are listed in table I (power events column). As it can be observed, there are no events for the freezer and the clothes washer. Regarding the freezer, we noticed that it was always ON (which was surprising, since such appliances usually have a duty cycle). As for the clothes washer, it was not possible to manually label the data due to the considerably high number of transitions. As such, we have decided to leave this device out of the semi-automatic labeling experiment.

Parameter sweep of the event detection algorithm. A parameter sweep is a controlled variation of a number of parameters in a particular algorithm (i.e., structural changes) and provides insights into how the different parameters affect the final results In this case, considering that the measurements are separated by at least two seconds, we have decided to fix the *pre* and *post* window lengths to one sample, and change only the PT and TS parameters. The power threshold (PT) was set to vary between 10%, 25%, 50% and 75% of the most represented sample in each appliance (i.e., the mode). Finally, the minimum elapsed time between events (TS), was set to vary between 2, 5, 10, 15, 30 and 60 seconds.

Ultimately, this parameter sweep returned 24 different event detection models, each of which was applied to 15 appliances

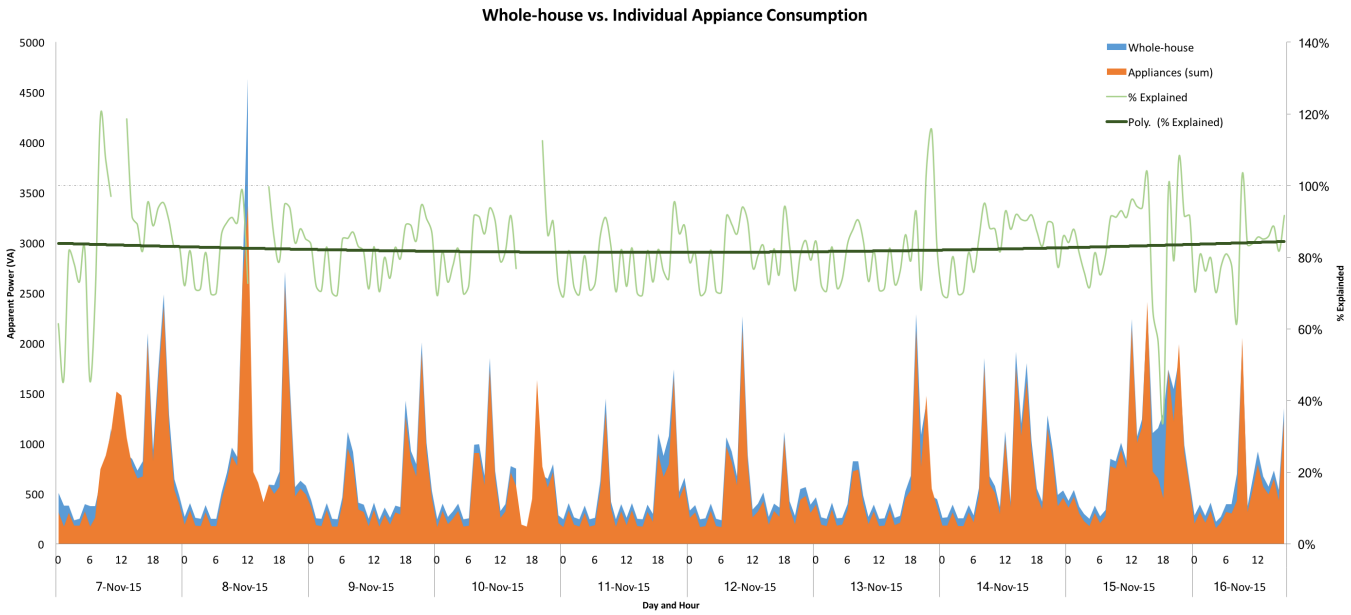

Fig. 4. Aggregate consumption vs. the sum of the individual appliances summarized by date and hour.

(excluding the freezer and the clothes washer).

Model evaluation against the ground-truth data. In order to evaluate the different models, we first had to compute the contingency matrix for each event detection test (i.e., count the number of True Positives -TP-, False Positives -FP-, False Negatives -FN- and True Negatives -TN-). To do this, we considered TP to be those events that were detected at most two seconds before or after the ground-truth timestamp. The FP positives are obtained by subtracting the number of TP from the total number of detected events. The FN are obtained by subtracting the number of TP from the number of ground-truth events. Finally, the TN are computed by subtracting the sum of TP, FP and FN from the total number of power samples in the respective appliance. We then computed the following performance metrics from the resulting contingency matrix: Precision (P), Recall (R) and F_1.

V. Results and Discussion

In this section, we present and discuss the results of the pilot deployment.

A. Individual Appliance Monitoring

In order to evaluate the performance of the system as a tool to collect datasets, we first look at the number of acquired samples and compare that value with what was expected in theory.

Overall, the deployment lasted 9 days and 18 hours, meaning that with an acquisition rate of 0.5 Hz, the system should have collected 421242 samples per appliance. As for the actual collected sample, the average was 356705 (n = 17, SD = 952). This is about 92% of the expected samples, and the low standard deviation indicates that this value is consistent across all the 17 appliances. Additionally, we also looked at the time interval between consecutive measurements, and the results shown that about 93% of the samples are separated by exactly two seconds, 2.6% separated by three seconds, 1.6% by one second and 1.10% by four seconds.

To further understand the potential of our platform we also examined how much of the demand was captured by the sub-metering platform. To this end we computed the aggregate consumption for the duration of dataset (from the raw current and voltage waveforms) and compared that with the sum of the aggregate consumption of the 17 monitored loads.

In figure 4 we show a plot of the aggregated and the sum of the individual appliances consumption grouped by date and hour. We also plot the percentage of energy explained (%EE), showing that in average the ground-truth data is able to explain 82% of the aggregate consumption. Likewise, it is also possible to observe that there are occasions when the %EE is above 100%. This happens because during some periods the whole-house data acquisition hardware stopped working. As such, when summarizing the data (e.g., by hour) there were occasions when the sum of the individual appliances consumption is higher than the aggregated demand, which is then reflected by an %EE above 100%.

Ultimately, considering the fact that most of of the existing datasets normally present considerable gaps in the ground-truth data [22], we believe that this is a very promising result. Nevertheless, our system suffers from the fact that using only plug-level meters it is not possible to collect ground-truth data for any loads that are not connected to a wall-socket (e.g., ceiling lights and appliances with dedicated circuits), which in this particular deployment represent about 20% of the household consumption.

TABLE I

LIST OF MONITORED APPLIANCES AND EVALUATION RESULTS: POWER (W) REFERS TO THE MOST REPRESENTED POWER MEASUREMENT OF THE APPLIANCE, AND EVENTS TO THE NUMBER OF POWER EVENTS. "E" IS THE NUMBER OF TP DETECTED IN THE POSITION THAT WAS LABELED MANUALLY.

Appliance	Power (W)	Events	Detection Results							
			TP	FP	FN	P	R	F_1	E (%)	BDP (PT%; ET)
Clothes washer	150	*	-	-	-	-	-	-	-	-
Coffee machine	950	60	57	7	3	0.89	0.95	0.92	57 (100%)	(10%, 15)
Dishwasher	50	50	46	1	4	0.98	0.92	0.95	46 (100%)	(50%; 30), (50%, 60)
Freezer	50	0	-	-	-	-	-	-	-	-
Kettle	1800	62	62	0	0	1	1	1	62 (100%)	(50 %; 30)
Laptop 1	24	6	5	1	1	0.83	0.83	0.83	5 (100%)	(25%; all)
Laptop 2	50	22	7	64	15	0.1	0.32	0.15	7 (100%)	(75%; 60)
Microwave	1350	152	123	22	29	0.85	0.81	0.83	121 (98%)	(25%; 10)
Oven	1600	20	20	0	0	1	1	1	20	(10%; 30), (10%; 60)
PlayStation 4	100	22	21	4	1	0.84	0.96	0.89	21	(25%; 60)
Refrigerator	100	514	514	0	0	1	1	1	514	(25%; 30), (25%; 60)
Stove	1750	818	777	24	41	0.97	0.95	0.96	773	(10%; 15)
TV 1	50	6	6	0	0	1	1	1	6	(25%; 15), (25%, 30), (25%, 60)
TV 2 + Box	100	60	58	1	2	0.99	0.97	0.98	58	(10%; 60)
TV 3 + Box	50	22	22	0	0	1	1	1	22	(10%; 30), (10%; 60)
TV 4 + Box + Home cinema	300	46	12	39	34	0.24	0.26	0.25	12	(50%; 15)
Water heater	2800	336	336	0	0	1	1	1	336	(10%; 60)

B. Event Detection

In semi-automatic labeling systems, the ultimate goal is to minimize the need for user intervention. In other words, minimize the number of FN (i.e., events that must be added by the user) while also minimizing the number of FP (i.e., events that must be removed from the interface). In terms of Precision and Recall, this means that the best event detectors are those that are able to maximize both metrics, i.e., F_1 is close to P and R. In table I we show the best models that we obtained for each appliance, according to the F_1 metric.

An overall look at the data shows that selecting the best parameters for each appliance it is possible to automatically detect 94% of the existing power events (2066 out of 2196). This number of automatically detected events comes at a cost of 163 false positives that should be removed from the interface, and 130 false negatives that still have to be manually added by the end-user. Ultimately, if we consider only the act of providing the correct labels to the data, these results show that the need for user intervention is reduced by about 86% (i.e., 293 clicks instead of 2196).

A more in-depth analysis also reveals that for some appliances (kettle, oven, refrigerator, TV 1, TV 3 and the water heater) the need for end-user intervention is reduced by 100% (i.e., $FN = FP = 0$). On the other hand, this analysis also reveals that in two particular cases the system performs very poorly. Namely, the *laptop 2* and the *television 4 + box + home cinema*.

One possible explanation for such poor results is the considerable noise that is added by these appliances, as shown in figures 5 and 6. Ultimately, for the selected detection algorithm, this noise produces a higher number of absolute step changes (PS) above the pre-defined threshold (PT), which are erroneously considered power events (i.e., the number of FP increases considerably, thus deteriorating Precision).

To further understand this issue, in figure 7 we plot the Precision and Recall values for these two appliances. As it can be observed, in the case of the *laptop 2* it is possible to find models with Recall values above 0.86 (1, 2, 7, 8). However, this decrease in FN comes at a cost of a prohibitive number of FP, and consequent deterioration of Precision.

As for the *television 4 + box + home cinema*, the results indicate that even the more liberal detectors (i.e., favours models with less FN independently of the number of FP) are not able to find more than 60% of the power events, which penalizes both Precision and Recall.

VI. CONCLUSIONS AND FUTURE WORK

Energy monitoring and especially Non-Intrusive Load Monitoring is a very active field of research with continuous interest over many years now, and the lack of proper datasets is believed to be one of the main reasons behind the almost absence of formal evaluations of the existing approaches. In this paper, we have presented and evaluated an approach that we believe can help produce better datasets by attempting to provide labels to the data in a semi-automatic fashion, thus alleviating researchers from the burden of doing this process manually.

Our initial results clearly advocate in favor of our a solution, in particular if we consider the high percentage of sub-metered energy (only about 7% of the samples were lost) and aggregated energy explained (over 80%). This said, future work should look at incorporating new sensors, and/or sensing techniques, such that it is possible to increase the percentage of energy explained. These solutions, should be able to capture the consumption of three types of loads that were only seldom considered in our approach. These are, hardwired loads like

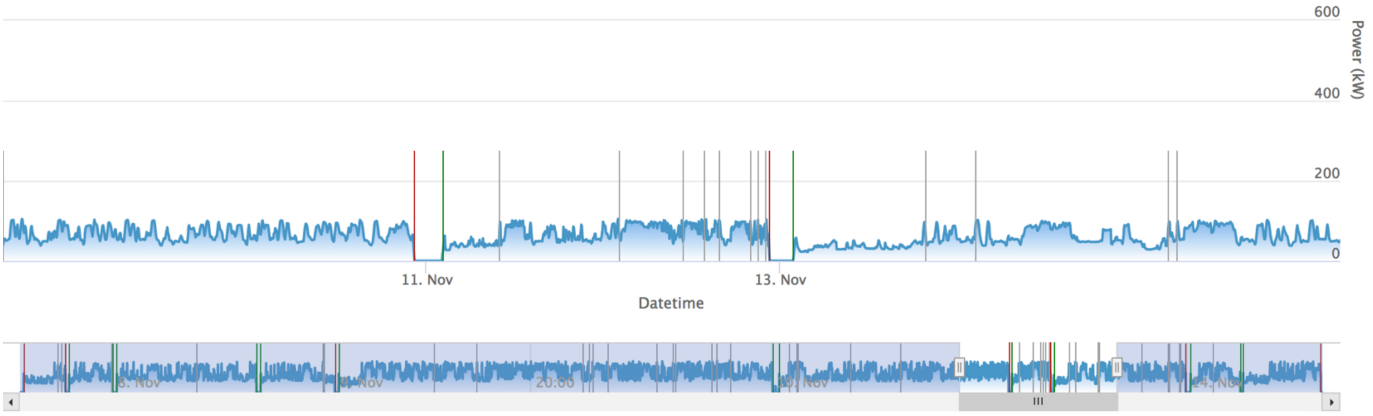

Fig. 5. Weekly consumption of the laptop # 2, showing how the the noise in the signal increases the number of false detections.

Fig. 6. Weekly consumption of the TV # 4 with the TV Box and Home Cinema, showing how the the noise in the signal increases the number of false detections.

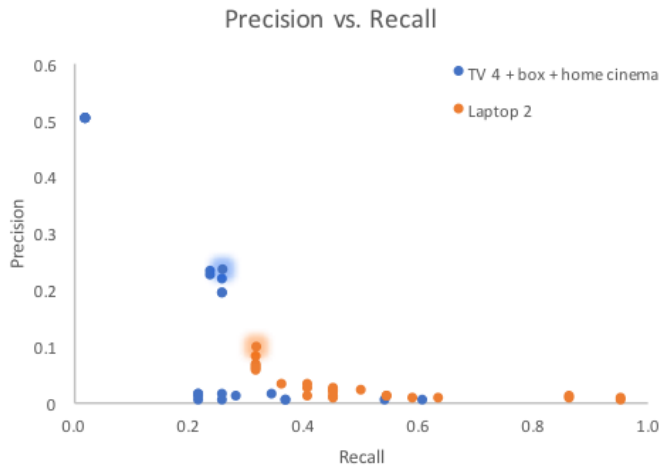

Fig. 7. Precision and recall for the laptop 2 and the TV 4. The models selected using the F_1 metric are highlighted with a glow effect.

ceiling lights, fast switching loads like washing machines, and loads without a fixed socket (e.g., the vacuum cleaner).

Regarding the semi-automatic labeling process, our results clearly indicate that when the sub-metered data is available it is possible to automatically find and label most of the transitions in the dataset with minimal user interference. Still, future work should look at incorporating additional algorithms, in particular targeting appliances that generate more noise. Furthermore, future work should also target different datasets, such that is possible to find which event detection models are more suitable for each appliance type.

Additionally, in future work we also aim at understanding how different users use the semi-automatic labeling application to label the same data, such that it is possible to quantify and better understand the levels of labeling ambiguity.

One possible improvement for this application would be the introduction of a second mode of interaction in which the end-users confirm the correct labels, instead of removing the wrong ones. It would therefore, be important to understand in which mode the end-users feel more comfortable. For example, confirming true positives implies a number of clicks at least equal to the number of events, which can be problematic for appliances that draw many power events like the refrigerator.

Another very important aspect that was not discussed in this paper is the direct mapping between the ground-truth labels

and the aggregate consumption data. For example, we are currently not aware of how the differences in the sampling rate of the two data streams will affect this mapping. Consequently, it would be of crucial importance to address this topic in future works.

Finally, future work should also look at ways to engage participants to the labeling process. One possibility is to rely on crowd-sourcing and citizen science for the supervision tasks, by adding the notions of experts and non-expert users. The former can be anyone willing to participate in the semi-automatic labelling process, whereas the latter should be someone that already has deep knowledge of the underlying research problems, as these would be responsible for assuring the quality of the labelling process.

Downloads For replication and re-utilization purposes, the source-code, pilot data, and the event detection results are publicly available in the following URL: http://aveiro.m-iti.org/feel

Acknowledgements

This research was supported by the FCT doctoral grant SFRH/DB/77856/2011 and project UID/EEA/50009/2013.

References

[1] G. Hart, "Prototype Nonintrusive Appliance Load Monitor," MIT Energy Laboratory Technical Report, and Electric Power Research Institute Technical Report, Tech. Rep., Sep. 1985.

[2] M. Berges, E. Goldman, H. Matthews, L. Soibelman, and K. Anderson, "User-Centered Nonintrusive Electricity Load Monitoring for Residential Buildings," *Journal of Computing in Civil Engineering*, vol. 25, no. 6, pp. 471–480, 2011. [Online]. Available: http://ascelibrary.org/doi/abs/10.1061/%28ASCE%29CP.1943-5487.0000108

[3] Z. Kolter and T. Jaakkola, "Approximate Inference in Additive Factorial HMMs with Application to Energy Disaggregation," in *JMLR: W&CP 22*, vol. 22, La Palma, Canary Islands, Spain, 2012, pp. 1472–1482. [Online]. Available: http://people.csail.mit.edu/kolter/lib/exe/fetch.php?media=pubs:kolter-aistats12.pdf

[4] H. Shao, M. Marwah, and N. Ramakrishnan, "A Temporal Motif Mining Approach to Unsupervised Energy Disaggregation: Applications to Residential and Commercial Buildings," in *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, ser. AAAI'13. Bellevue, Washington: AAAI Press, 2013, pp. 1327–1333. [Online]. Available: http://dl.acm.org/citation.cfm?id=2891460.2891645

[5] M. Zeifman and K. Roth, "Nonintrusive appliance load monitoring: Review and outlook," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 1, pp. 76–84, 2011.

[6] A. Zoha, A. Gluhak, M. A. Imran, and S. Rajasegarar, "Non-Intrusive Load Monitoring Approaches for Disaggregated Energy Sensing: A Survey," *Sensors*, vol. 12, no. 12, pp. 16 838–16 866, Dec. 2012. [Online]. Available: http://www.mdpi.com/1424-8220/12/12/16838

[7] L. Pereira, "Hardware and Software Platforms to Deploy and Evaluate Non-Intrusive Load Monitoring Systems," PhD, Universidade da Madeira, Funchal, Portugal, 2016.

[8] Z. Kolter and J. Matthew, "REDD: A public data set for energy disaggregation research," in *1st KDD Workshop on Data Mining Applications in Sustainability (SustKDD)*, San Diego, CA, USA, 2011. [Online]. Available: http://redd.csail.mit.edu/kolter-kddsust11.pdf

[9] S. Makonin, B. Ellert, I. V. Bajić, and F. Popowich, "Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014," *Scientific Data*, vol. 3, p. 160037, Jun. 2016. [Online]. Available: http://www.nature.com/articles/sdata201637

[10] N. Batra, M. Gulati, A. Singh, and M. B. Srivastava, "It's Different: Insights into Home Energy Consumption in India," in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, ser. BuildSys'13. New York, NY, USA: ACM, 2013, pp. 3:1–3:8. [Online]. Available: http://doi.acm.org/10.1145/2528282.2528293

[11] J. Kelly and W. Knottenbelt, "The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes," *Scientific Data*, vol. 2, p. 150007, Mar. 2015. [Online]. Available: http://www.nature.com/articles/sdata20157

[12] K. Anderson, A. Ocneanu, D. Benitez, D. Carlson, A. Rowe, and M. Berges, "BLUED: A Fully Labeled Public Dataset for Event-Based Non-Intrusive Load Monitoring Research," in *2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD)*, Beijing, China, Aug. 2012. [Online]. Available: http://www.marioberges.com/SustKDD12/

[13] K. Torkkola, C. Schreiner, M. Gardner, and K. Zhang, "Development of a Semi-Automatic Data Annotation Tool for Driving Data," in *IEEE Intelligent Transportation Systems Conference, 2006. ITSC '06*, Sep. 2006, pp. 642–646.

[14] S. Bianco, G. Ciocca, P. Napoletano, and R. Schettini, "An interactive tool for manual, semi-automatic and automatic video annotation," *Computer Vision and Image Understanding*, vol. 131, pp. 88–99, Feb. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314214001544

[15] J. Niño-Castañeda, A. Frías-Velázquez, N. B. Bo, M. Slembrouck, J. Guan, G. Debard, B. Vanrumste, T. Tuytelaars, and W. Philips, "Scalable Semi-Automatic Annotation for Multi-Camera Person Tracking," *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, vol. 25, no. 5, pp. 2259–2274, May 2016.

[16] L. Pereira and N. J. Nunes, "Semi-automatic labeling for public non-intrusive load monitoring datasets," in *2015 Sustainable Internet and ICT for Sustainability (SustainIT)*, Apr. 2015, pp. 1–4.

[17] H. A. Cao, T. K. Wijaya, K. Aberer, and N. Nunes, "A collaborative framework for annotating energy datasets," in *2015 IEEE International Conference on Big Data (Big Data)*, Oct. 2015, pp. 2716–2725.

[18] L. Pereira, N. Nunes, and M. Bergés, "SURF and SURF-PI: A File Format and API for Non-intrusive Load Monitoring Public Datasets," in *Proceedings of the 5th International Conference on Future Energy Systems*, ser. e-Energy '14. New York, NY, USA: ACM, 2014, pp. 225–226. [Online]. Available: http://doi.acm.org/10.1145/2602044.2602078

[19] C. Beckel, K. Kleiminger, R. Cicchetti, T. Staake, and S. Santini, "The ECO Data Set and the Performance of Non-Intrusive Load Monitoring Algorithms," in *1st ACM International Conference on Embedded Systems for Energy-Efficient Buildings*. Memphis, TN. USA: ACM, Nov. 2014.

[20] A. Monacchi, D. Egarter, W. Elmenreich, S. D'Alessandro, and A. M. Tonello, "GREEND: An Energy Consumption Dataset of Households in Italy and Austria," in *In proceedings of the 5th IEEE International Conference on Smart Grid Communications*. Venice, Italy: IEEE, May 2014, arXiv: 1405.3100. [Online]. Available: http://arxiv.org/abs/1405.3100

[21] P. Meehan, C. McArdle, and S. Daniels, "An Efficient, Scalable Time-Frequency Method for Tracking Energy Usage of Domestic Appliances Using a Two-Step Classification Algorithm," *Energies*, vol. 7, no. 11, pp. 7041–7066, Oct. 2014. [Online]. Available: http://www.mdpi.com/1996-1073/7/11/7041

[22] N. Batra, J. Kelly, and O. Parson, "NILMTK: An Open Source Toolkit for Non-intrusive Load Monitoring," in *e-Energy '14: Proceedings of the 5th International Conference on Future Energy Systems*. Cambridge, UK: ACM, Jun. 2014.