Abstract — Since the power consumption of different electrical appliances in a household can be recorded by individual smart meters, it becomes possible to start considering in more detail the interactions of the residents with those devices throughout the day. Appliances’ usages should not be considered as independent events, but rather as enablers for activities. Leveraging activity knowledge over time will allow us to design personalized energy efficient measures. We envision the design of future ambient intelligence systems, where the smart home can optimize the energy consumption in regards to the lifestyles of its residents and the smart grid’s needs. In this work, we propose an automated method for determining when an electrical device is triggered by households’ residents solely from its power trace. Knowing when an appliance is in use is required for identifying recurrent patterns that could later be understood as activities.

Index Terms — Time series analysis; Data mining, Information search and retrieval; Clustering; Smart energy; Smart meters; Activity inference; Appliances states; Energy data analytics; Datasets; Algorithms

I. INTRODUCTION

The future smart grid offers the possibility of having fine-grained information and capabilities to monitor its status in real time. Implementing real-time and personalized feedback could amount to a substantial energy reduction in the residential segment [1]. This should be considered with the potential savings during peak time, when high penalties might become a reality in the future. It can also be the cornerstone of future off-the-grid scenarios as micro-generation and battery technologies become more affordable. Focusing on the household scale offers an alternative to aggregating levels in Demand Response Systems. In the context of the smart home, one could foresee trading off users’ lifestyle preferences and comfort with saving measures, while preserving the privacy of the residents, by providing an optimization inside households.

From a technical standpoint, it has yet to be decided how much information should be collected, i.e., the granularity of such data, and which additional sensors should be integrated to provide a better understanding of how energy is consumed. To this end, the access to disaggregated data requires the setup of data collection architectures with prohibitive costs. One practical alternative is single point, non-intrusive sensing of aggregated energy which involves the development of Non-Intrusive Load Monitoring (NILM) algorithms on existing household-level aggregated data to differentiate the devices in use [2]. Given the recent release of a large dataset with appliance-level measurements, abstracting the usage of electrical devices in households by investigating the motives behind them being triggered by a user becomes possible. This involves unraveling information from the collected power measurements and finding out when and how they are used in conjunction.

Until smart appliances become widespread, determining the state of an appliance and in particular, when it is active from when it is idle or in standby mode, can only rely on disaggregated power time series. We investigate how an appliance’s trace properties can be leveraged without side information that could assess the proximity of the residents, nor ground truth data from a journal that documents the activities in the household, to determine when there is interaction with an appliance to carry out a human activity. Setting fixed thresholds based on the analysis of a set of known appliances and building databases of signatures will not scale with the release of new models of appliances, as their characteristics are expected to evolve as devices become more efficient due to technological improvements. Instead, determining these thresholds agnostically of the appliances’ types, models and brands, based on statistical properties of their consumption, would be adaptable for existing and next generation devices.

In order to determine which appliances are utilized conjointly and linked to a human activity, our contribution is to distinguish the active consumption from the baseline and noise in their power traces. Our method could be extended to other types of sensors, where it is necessary to determine useful measurements from baseline noise (such as in the case of inertial sensors).

The remainder of this paper is organized as follows. Section II presents related work. Section III introduces the methodology for the automatic thresholding. Section IV shows the algo-

Hông-Ân Cao
Department of Computer Science
ETH Zurich, Switzerland
Email: hong-an.cao@inf.ethz.ch

Tri Kurniawan Wijaya, Karl Aberer
Department of Computer Science
EPFL, Switzerland
Email: {tri-kurniawan.wijaya, karl.aberer}@epfl.ch

Nuno Nunes
Madeira Interactive Technologies Institute
Funchal, Portugal
Email: njn@uma.pt
Activity recognition is a long-established field of research. Previous work looked at human trajectories, interactions with objects or social activities [3]. However, most approaches neither target energy conservation, nor use the electricity consumption as an input variable for the recognition of activities. Thus, our goal of estimating human interactions with electrical appliances agnostically is most closely related to recent work on demand side management. The ability to accurately predict future energy needs is the cornerstone in proper demand side management, and many research efforts have been devoted to this in the last couple of years [4]. Some of the investigated methods rely heavily on past consumption data to predict future demand, and therefore, we argue that our research can be of added value in this situations, especially given the high granularity of data (one measurement per minute) that can be easily modified to test different predictions periods (e.g., hour, day, week) to evaluate the outcomes of the prediction algorithms in a variety of energy consumption scenarios, including off-the-grid households.

Previous work used statistical attributes of the data to determine occupancy, we are however assessing activities that incur energy consumption [5]. While NILM has focused on disaggregating loads by supervised learning through ON-OFF events [6], state detection for modeling and maintaining appliances’ signatures [7], [8], spike detection [9] or an analysis of the different appliances patterns [10], determining when an appliance is active, often relies on using a predefined threshold [11]. Activity recognition in households can be assisted through sensor deployment in households [12]–[15] or WiFi signatures [16]. When real-life deployments were not possible, prior work used simulated power traces for investigating human activities in households [17]–[19]. Attempts at using existing publicly released datasets to identify appliances that are used in conjunction and the flexibility of their usage in households have utilized the REDD dataset to support their analysis but have used predefined thresholds for determining when the appliances were on ON or OFF [20].

Our approach attempts to tackle the known limitations of current eco-feedback systems, which focuses on increasing efficiency by raising end-user awareness of how their actions impact the use of energy. Our previous research [21] showed that energy disaggregation strategies, commonly used in eco-feedback systems, are overwhelming for most users, as they lose interest and show relapsing behaviors in their energy conservation actions. From the initial challenge of creating effective low-cost disaggregation strategies we faced the new problem of generating meaningful strategies to re-aggregate consumption data that could effectively lead to long-term sustainable energy conservation practices in domestic environments.

II. RELATED WORK

III. METHODOLOGY

Using only electrical loads (no side information, nor ground truth), it is necessary to evaluate how to differentiate baseline consumption that can be considered as noise, from human-triggered actions. While it would be possible to handpick a threshold to decide when the appliance is powered on and serving a human activity, such a process would be done arbitrarily and would not be generalizable given the multitude of brands and models in consumer electronics and how they change and evolve due to technological advances. To this end, we developed an automated way of deciding when an appliance reaches a power level high enough, such that it can be regarded as being used by a human being. This requires considering each household separately and learning from the specificity of each trace. Such method relates to image thresholding, an essential method for isolating objects or other relevant information in digital images [22].

A. State Estimation

We consider two types of power traces, namely appliance-level data (single appliances), and circuit-level data (aggregated readings recorded by instrumenting circuits at the room level, or obtained from a power strip). We refer to both as appliances from now on. We explain how different power levels are linked to the appliance’s state and its utilization.

Since a human being is not activating the appliances throughout the day, we can distinguish between an idle state (off / stand-by mode, typically low power levels) and an active state (when the residents are powering it on or actively interacting with it). We notice, for example, in the case of a washing machine, that several mechanisms allow running different washing programs and cycles throughout its time of use (soaking, spinning, etc.). In the case of data being collected at the circuit level, we could expect to observe different devices (lights, smaller consumer electronics) being turned on. Each mode of functioning can be related to the internal state of an appliance in the case of single appliances or to different electrical devices being switched on in the case of circuit level data and operating at different power levels [23]. So, we rely on this to suggest that different states in the use of an appliance are linked to different levels of power. Following this idea, we want to observe the relationships between power levels in the distribution of the power measurements of an appliance.

Although we intend to discover activities in a data-driven manner, i.e., without a-priori knowledge, nor human labeling, we have in mind for the time-being high level activities (such as cooking, cleaning, etc.). This means that we do not dwell into the intricacy of the different stages involved in an activity (in the case of cooking: cleaning vegetables, heating ingredients, eating, etc.). Thus, if we consider a power strip in the kitchen and its respective power readings, the transitions in the traces might be due to smaller appliances being powered on (kettle, mixer, etc.). However, since, they are not disaggregated, they cannot be labeled and cannot be directly used. This is why we focus on the overall duration of
the interaction with an appliance, not differentiating between all the stages and sub-activities it might involve, thus we only consider two appliance states, i.e., idle or active.

B. Gaussian Mixture Model

We model the distribution of power levels by approximating it with a Gaussian Mixture Model (GMM) [24]. A GMM is a probabilistic model that assumes that the data points under consideration are generated from a mixture of a finite number of Gaussian distributions. The estimation of the means and covariances that define the Gaussians is obtained by achieving the maximum likelihood of the mixture through the Expectation-Maximization algorithm (also known as EM). We refer the reader to Section 6.8 and Chapter 8 of [25] for a formal definition of the GMM and the EM algorithm respectively.

The different modes of an appliance’s power distribution can be attributed to the different internal states of the appliance or to the sequence of appliances being activated in the case of circuit-level data. Given that most of the appliances operate at low power levels during their idle period, the idle state can be identified as the first set of correlated measurements. Thus, we locate the point that lies in the first valley of the Gaussian mixture (the first Gaussian identified by its mean at \( \mu_1 \) represents the idle status, while starting from the second Gaussian centered at \( \mu_2 \), the appliance is considered in use). We define the bottom of the valley as the minimum of the distribution between the first and the second Gaussians as in Equation 1 for \( p \) being the multimodal distribution modeled by the GMM. In the case where the GMM overfits close small peaks, we merge those peaks and identify \( \mu_1 \) as the largest mean in the set of adjacent peaks.

\[
\arg\min_{\mu_1 \leq x \leq \mu_2} p(x) \tag{1}
\]

We propose GMMthresh as the procedure to determine the best GMM fit for an appliance’s distribution and to locate the threshold between the first two modes of the distribution as can be seen in Algorithm 1.

### IV. Empirical Evaluations

A. Datasets

The Pecan Street dataset (http://www.pecanstreet.org/) originally comprised 239 monitored households mostly located in Texas. Their aggregate power consumption and disaggregated load readings are provided at a rate of once every minute and span from January to May 2014. While 70 different types of appliances are recorded, there are at most 22 actively monitored circuits per household. Appliances with larger ranges of consumption are for example ovens, dishwashers or furnaces.

We leverage the wisdom of the crowd by using expert annotated data from the Pecan Street dataset through our CAFED platform (https://cafed.inf.ethz.ch). This tool allows us to select and display power time series dynamically to users that are familiar with the energy domain and have the required knowledge for discerning when an appliance is active from when it is idle by looking at its power trace. The user can then interact with the platform and highlight portions of the time series where the appliance is active. The expert annotated data are collected through the platform and made available to other researchers in the community. Using this expert crowdsourcing method, over 4500 daily time series have been collected so far and we believe that the framework could be extended to other publicly available datasets [26].

B. Parameter selection

Our algorithm considers one month of data per appliance (to minimize the impact of weather) and to ensure that enough data are available (some appliances might not be used frequently on a weekly basis). The readings’ distribution can be represented by a histogram of the different power measurements, where the modes coincide with Gaussians and the peaks with the Gaussians’ means. We observe for each month that some power level readings amount to thousands of occurrences, while the magnitude of other representatives is in the order of hundreds to a few instances as in Figure 1.

Therefore, the data are scaled to lessen the order of magnitude between the measurements, in particular the lower measurements, since the appliance is expected to be mostly in idle mode. This amplifies all candidate peaks (Gaussians) with regards to the more prominent low power peaks. The scaling of the histogram power distribution consists in selecting for each bin \( i \), the quantity \( n_i \) of power measurements in the bin and to convert it to a logarithmic scale, thus in the order of \( C \times \log(n_i + 1) \), where \( C \) is a constant. The rescaling of the
We use the first week of February to evaluate the thresholds circuits as can be seen in Table I. From the CAFED dataset, monitored appliances combining both single appliances and

C. Evaluation

The evaluation is performed by using January data to determine the threshold for the active state for a set of 8 monitored appliances combining both single appliances and circuits as can be seen in Table I. From the CAFED dataset, we use the first week of February to evaluate the thresholds determined for the selected appliances for 10 households. Additionally, to evaluate the performance of the algorithm over time and show the effect of the input data in determining the threshold, we select one household where the thresholds for the appliances are computed for the first 4 months and use the subsequent first week of the following month as testing data. The available input data for the GMM is shown in Table II.

We compare the performance of the GMM thresholding to two arbitrary thresholds, i.e., 0 W, which can be used in the case where the baseline is zero and 50 W, which can be considered as an educated guess for detecting most of the major appliances [11] and taking into account the standby-power of most consumer electronics devices [27], [28].

We score the different parametrizations by using common information retrieval scores as follows. The precision as defined in Equation 4 measures the fraction of data points that were actually annotated as active against all data points that the algorithm determined to be active. The recall as in Equation 5 measures the proportion of data points that the algorithm determined to be active in comparison with the actual number of available active points. Its limitation relies in the fact that a perfect recall score can be achieved by deciding that all data points should be considered as active. This is why, another common score is the $F_1$ score as in Equation 6, which combines both previous measurements and balances their effect. Additionally, we define a score $s_H$ as in Equation 8 based on the Hamming distance as defined in Equation 7.

$$precision = \frac{TP}{TP+FP} \quad (4)$$

$$recall = \frac{TP}{TP+FN} \quad (5)$$

$$F_1 \ score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (6)$$

$$d_H(a, b) = \sum_{i=0}^{n} a(i) \oplus b(i) \quad (7)$$

$$s_H = \frac{1}{N} \sum_{j=1}^{N} d_H(a_j, b) \quad (8)$$

The evaluation is performed by determining the thresholds in January and evaluating them against the annotated ground truth of the first seven days of February. We however distinguish two cases in the handling of the annotated ground

<table>
<thead>
<tr>
<th>Appliance</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>bathroom1</td>
<td>Circuit</td>
</tr>
<tr>
<td>clotheswasher1</td>
<td>Single Appliance</td>
</tr>
<tr>
<td>dishwasher1</td>
<td>Single Appliance</td>
</tr>
<tr>
<td>kitchen1</td>
<td>Circuit</td>
</tr>
<tr>
<td>light_plugs1</td>
<td>Circuit</td>
</tr>
<tr>
<td>livingroom1</td>
<td>Circuit</td>
</tr>
<tr>
<td>microwave1</td>
<td>Single Appliance</td>
</tr>
<tr>
<td>oven1</td>
<td>Single Appliance</td>
</tr>
</tbody>
</table>

**TABLE I: Selected appliances and their categories**
TABLE II: Appliances per household

<table>
<thead>
<tr>
<th>household_id</th>
<th>bathroom1</th>
<th>clotheswasher1</th>
<th>dishwasher1</th>
<th>kitchen1</th>
<th>light_plugs1</th>
<th>livingroom1</th>
<th>microwave1</th>
<th>oven1</th>
</tr>
</thead>
<tbody>
<tr>
<td>6910</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>1632</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5568</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2974</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>9922</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>9737</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>7982</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>8142</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>8197</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>8669</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

TABLE III: GMM parametrization: selected configuration 15 GMM, no binning (higher is better for the precision, the recall and \(F_1\) score and lower is better for s\(_H\))

| gmm | bin | score | bathroom1 | clotheswasher1 | dishwasher1 | kitchen1 | light_plugs1 | livingroom1 | microwave1 | oven1 | avg | std |
|-----|-----|-------|-----------|----------------|-------------|----------|--------------|-------------|------------|-------|-----|-----|-----|
| 10  | 1   | 1.000 | 1.000     | 0.981         | 1.000       | 0.992    | 1.000        | 1.000       | 1.000      | 1.000 | 0.977 | 0.007 |
| 10  | 5   | 1.000 | 0.994     | 1.000         | 0.981       | 0.999    | 0.841        | 1.000       | 1.000      | 0.977 | 0.055 |
| 10  | 10  | 1.000 | 0.999     | 0.954         | 0.944       | 0.999    | 0.855        | 0.902       | 1.000      | 0.957 | 0.055 |
| 15  | 1   | 1.000 | 0.999     | 0.999         | 0.981       | 1.000    | 0.992        | 1.000       | 0.900      | 1.000 | 0.901 | 0.029 |
| 15  | 5   | 1.000 | 0.994     | 0.981         | 0.999       | 0.841    | 0.999        | 1.000       | 1.000      | 0.976 | 0.055 |
| 15  | 10  | 0.804 | 0.883     | 0.953         | 0.944       | 0.999    | 0.855        | 0.902       | 1.000      | 0.918 | 0.066 |

truth data. In the fetching process of dispatching curves to be annotated by our contributors, we enforce majority voting, i.e. each curve should be annotated by 3 users and for each data point, the most frequent annotation is chosen (2 are necessary in this case). In the case where 2 annotations per data point are obtained, annotators could diverge on some annotated points. This is why in the latter, we evaluate the precision, recall and \(F_1\) score on points where the annotations concord, while the Hamming score consists on a weighted average of the individual annotations provided by each annotator as in Equation 8.

D. Results

As can be seen in Table III, we compute the average scores per appliance and per household as defined in subsection IV-C. Then, we combine the scores obtained for all appliances in each household by averaging them to evaluate the model’s predictive power. The best approximation for the power distribution should be such that its modes are fitted by the Gaussians determined by GMMthresh. This means that the best scores should be achieved, i.e., higher precision, recall and \(F_1\) score and lower Hamming score \(s_H\). Two parameters are evaluated: the number of Gaussians in the model and the effect of the binning (or rounding) of the power measurements.

We can see from Table III that a configuration allowing to search for more Gaussians fits the power distribution more closely. The rounding effect is to reduce the effect of neighboring modes, allowing to reduce their overfit. However, aggregating measurements also reduces the accuracy of the thresholding when modes are adjacent, especially in the case of the largest tested bin size (10 W). This is particularly noticeable for appliances whose states operate at a more fine-grained power scale. Overall, the best configuration that minimizes the Hamming score (the least differences between the binary output from the GMM and the annotated data) and maximizes the \(F_1\) score consists in modeling 15 Gaussians and not binning the data.

The outcome of the algorithm can be seen in Figure 2 in the case of dishwasher1 (single appliance) and of livingroom1
(a) Overlay of a histogram and its GMM approximation for livingroom1 (circuit-level data). Underlying Gaussians in dashed lines. (b) Overlay of a daily power trace and binary decision (idle below / active above the threshold line) for livingroom1

(c) Overlay of a histogram and its GMM approximation for dishwasher1 (single appliance data). Underlying Gaussians in dashed lines. (d) Overlay of a daily power trace, and binary decision (idle below / active above the threshold line) for dishwasher1

Fig. 2: Outcome of the GMM for livingroom1 (circuit-level data) and dishwasher1 (single appliance). In (b) and (d), power below the threshold is considered to be in the idle state, and in the active state otherwise.

(circuit / room). In both cases, as can be seen in the respective test (annotated) time series in Figures 2d and 2b, 50 W would not highlight the smaller power measurements (the ramping up and the ramping down of the device) in the case of dishwasher1, while in the case of the livingroom1, the baseline is above 50 W. If the baseline level is close to the arbitrarily chosen threshold (for testing purposes it was set to 50 W), the decision for livingroom1 would be to classify it erroneously as being active throughout the day.

We compare the performance of GMMthresh in terms of the $F_1$ score and Hamming score $s_H$ of the selected model against the usage of 50 W and 0 W as thresholds. Figure 3 shows that GMMthresh performs steadily well for all appliances and consistently outperforming the 0 W threshold. It outperforms the 50 W threshold in all cases, except for kitchen1 and light_plugs1. From Table III, the other scores' performance similarity is linked to the fact that the determined thresholds lie generally below 20 W as can also be seen across households in Figure 6a. dishwasher1 is however better detected by the GMM and the 50 W thresholds as the determined thresholds are more spread than in the case of clotheswasher1 as can be seen in Figure 6a. microwave1 and oven1 show the worst performance for the 0 W threshold as low power measurements (< 10 W) are erroneously detected as showcasing human activity.

In the case of circuit-level data, we have seen that when the baseline is above 50 W as in Figure 2b, the appliance is considered active during the whole day. The baseline can be attributed to consumer electronics for entertainment in the case of livingroom1 that remain in standby mode and are thus not voluntarily powered on to be used by the residents. The predictive power per household combines the scores for all appliances belonging to each household. As can be seen in Figure 4, when combining the previous observations, the GMMthresh performs better overall. While all households are single-family homes, the performance varies across the households due to the set of appliances available and the residents’ lifestyles as can be seen in Figure 6a.

We expect that some appliances are used less frequently than others (for example oven1). Since the determination of the threshold through GMMthresh depends on the input data, we show the scores combined from the thresholds computed monthly for January through April for household 6910 in Figures 5. Throughout those 4 months, the GMM maintains its prediction power close to the 0 W and above the 50 W thresholds and outperforms both static thresholding methods in the case of livingroom1. As can be seen from Figure 6b, the determined thresholds do not vary significantly for appliances that are used regularly (such as bathroom1 or kitchen1). dishwasher1 and light_plugs1 show the most variance. Since the method depends on historical data, it is to be expected that it requires enough data to estimate the power distribution of an appliance.
V. CONCLUSIONS AND FUTURE WORK

In this work, we introduced an automated way of determining when an appliance is activated by a human being by filtering out baseline noise from the readings and by looking at the distribution of the power measurements with consistently high accuracy. Our methods performed better than the generally accepted best guess thresholds and achieve an F1 score score of about 0.9 for all appliances that were evaluated.

Having now obtained binary vectors of data, we intend to consider daily time windows and infer patterns of appliances being used conjointly and derive temporal rules. In a real-life deployment, to mitigate the fact that the thresholds depend on the available data (the appliances have to be used by the households residents), the accuracy could be improved by developing an online version of the algorithm with a decay factor for forgetting past thresholds and balancing with newly evaluated thresholds.

We believe that our approach for automatically detecting changes between the active and idle states of appliances could lead to important and practical applications that move beyond traditional eco-feedback systems and anticipate distributed micro-generation scenarios leading to important changes in energy sustainability and ultimately the utility business. To this end, we anticipate to provide a combination of (i) actionable
recommendations for energy conservation including those that take advantage of the availability of renewable sources and new battery technologies, (ii) suggesting novel approaches for in-house automation that could leverage smart appliances and grid supply / demand balance.

VI. ACKNOWLEDGMENTS

The authors would like to thank Régis Blanc for his support and invaluable help.

REFERENCES


