IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 3, 2018, accepted September 12, 2018, date of publication September 20, 2018, date of current version October 17, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2871343

Improving Event-Based Non-Intrusive Load Monitoring Using Graph Signal Processing

BOCHAO ZHAO, KANGHANG HE, LINA STANKOVIC, (Senior Member, IEEE), AND VLADIMIR STANKOVIC, (Senior Member, IEEE)
Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XW, U.K.

Corresponding author: Bochao Zhao (b.zhao@strath.ac.uk)

This work was supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC) Projects under the Transforming Energy Demand in Buildings through Digital Innovation (BuildTEDDI) Funding Programme under Grant REFIT EP/K002368.

ABSTRACT Large-scale smart energy metering deployment worldwide and integration of smart meters within the smart grid will enable two-way communication between the consumer and energy network, thus ensuring improved response to demand. Energy disaggregation or non-intrusive load monitoring (NILM), namely disaggregation of the total metered electricity consumption down to individual appliances using purely algorithmic tools, is gaining popularity as an added-value that makes the most of meter data. However, NILM remains a challenging problem since NILM is susceptible to sensor noise, unknown load noise, transient spikes, and fluctuations. In this paper, we tackle this problem using novel graph signal processing (GSP) concepts, applied at both, physical signal level via graph-based filtering and data level, via effective semi-supervised GSP-based feature matching. The proposed GSP-based method is generic and can be used to improve results of various event-based NILM approaches. We demonstrate significant improvement in performance using three state-of-the-art NILM methods, both supervised and unsupervised, and real-world active power consumption readings from the REDD and REFIT[1] data sets, sampled at 1 and 8 s, respectively.

INDEX TERMS Load disaggregation, non-intrusive load monitoring, smart metering, graph signal processing.

I. INTRODUCTION

Integration of smart meters into smart grids, enabled by advanced sensing and communication technologies such as [1] and [2], provides two-way communications between the consumer and energy network to respond in real time to demand [3]. Real-time energy feedback, currently available with widescale smart meter deployments via In Home Displays or web-based, can be made smarter, i.e., more informative and actionable, by non-intrusive load monitoring, as proposed [4]. Non-intrusive load monitoring (NILM) refers to estimating individual appliance energy consumption from the aggregate meter-point measurements (e.g., voltage, current, power) using purely analytical tools. Though first proposed in the 1980's [4], large-scale smart metering deployments worldwide [3] have ignited a renewed interest in NILM, leading to many algorithmic improvements and some commercial products aimed to enrich energy feedback [5]. A systematic review of the literature [6] indicates that NILM feedback may contribute to the reduction of domestic electricity consumption by 0.7%–4.5% on average, compared to the more common aggregate electricity consumption feedback. Additionally, the information obtained from NILM about individual appliance use and consumption is useful to appliance manufacturers, policy makers, for smart home automation, assisted living, and demand response [5].

Driven by these emerging applications, NILM has become a very active research area [7]–[9]. Advanced signal processing and machine learning methods have been proposed for NILM, including support vector machines (SVM) [10], Hidden Markov Models (HMM) and their extensions [11]–[17], sparse coding [18], motif mining [19], artificial neural networks, shallow [20], [21] and deep neural networks [22]–[24].

Despite significant NILM research in recent years (see [7]–[33] and references therein), low-rate NILM,

[1]The REFIT dataset used to generate the results can be accessed via DOI 10.15129/31da3ece-f902-4e95-a093-e0a9536983c4.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/

i.e., NILM of electricity readings captured with sampling rate of $\geq$ 1 second via widespread smart meters [3], is still an open problem. Indeed, current state-of-the-art solutions are susceptible to measurement noise and outliers when dealing with real-world data and do not demonstrate sufficient accuracy [7], [9]. One reason for this is the complex nature of the NILM problem with effective solutions requiring both core physical-level signal processing - to process acquired signals reducing jitter, noise, spurious events [8], [27], [34] - and machine learning-based clustering and classification [7].

In this paper, we propose graph signal processing (GSP) as a tool that brings together low-level *signal* processing and application-driven *data* processing in order to improve the performance of various event-based NILM approaches, suitable for diverse electrical load datasets. GSP is an emerging approach that provides robust means for signal denoising [35], clustering [27], and classification [8], where complex relationships between samples of high dimensional data are represented using graphs.

GSP-based approaches have recently been proposed for tackling the NILM problem, via supervised [8] and unsupervised approaches [27].[2] However, this prior work applied GSP at the data processing stage only, i.e., as a robust classification or clustering tool, without exploiting GSP's properties as effective physical signal filters [36], which can combat NILM sensitivity to measurement noise and the influence of unknown appliances. It is well recognized [7], [8], [27], [34], that without appropriate processing of the physical measured signal, NILM will often not be accurate or successful, regardless of the effectiveness of the employed classification method.

To address the above issues and enhance NILM methods, including [8] and [27], we propose two *universal* algorithms. First, capitalizing on recent advances in GSP filtering (see [35]–[37]), we propose a novel signal processing approach to mitigate sensor noise and sharpen signal edges to improve detection of on/off appliance events, which in turn facilitates more effective feature extraction and classification in NILM. We design two types of GSP filters - one based on total variation regularization [35] and the other based on bilateral filtering [38]. Since graph bilateral filtering results in a smoother output at the cost of occasionally filtering out true events, we develop an algorithm to select, automatically, the best filtering method. Second, relying on robust GSP-based *data* classification [8], [27], [36], [37], we propose a novel NILM result *refining* method, applicable to any NILM algorithm; this method is based on semi-supervised GSP-based feature matching to improve disaggregation results by removing confusion between appliances with similar power levels that are often misclassified by the initial NILM classification engine.

The proposed methods are generic and applicable to a range of NILM approaches, including supervised, semi-supervised and unsupervised NILM. The effectiveness of the proposed methods are demonstrated across three state-of-the-art NILM approaches, based on Decision Trees (DT) [29], supervised GSP [8] and unsupervised GSP NILM [27]. Besides the methods from [8], [27], and [29], we also benchmark performance against two additional NILM methods from the publicly available NILMTK toolbox based on Hidden Markov Model (HMM) and Combinatorial Optimization (CO) [4], [39]. Results are validated using two open access datasets of true power measurements: REDD [40] (US houses) and REFIT [41] (UK houses).

The rest of the paper is organized as follows. In Section II, we briefly review literature on NILM, as well as a discussion of methods proposed for improving NILM performance. In Section III, we present notation details for the entire paper and give an overview of the GSP concepts related to the proposed algorithms. Details of the proposed methods are presented in Section IV. Section V shows experimental results with appropriate validation and benchmarking. Findings of this study are discussed in Section VI.

II. BACKGROUND ON NILM

In this section, an overview of existing NILM approaches is provided followed by a review of core signal processing methods tackling load signal imperfections affecting NILM.

NILM approaches can be state-based or event-based. In state-based NILM approaches, usually based on HMM and its variants, a finite state machine is used to build appliance load distributions models (e,g., [11]–[13], [32], [42]). Four different HMM architectures are proposed in [11] with further improvements based on differential HMM in [12], and using expectation-maximization (EM) to generate accurate appliance state transition models in an unsupervised manner in [13]. A NILM method using particle filtering to estimate Factorial-HMM(FHMM) inference is presented in [14]. In [15], sub-metering measurements are used to build super-state HMMs, and Viterbi algorithm is then exploited for disaggregation. In [16], additional statistical features are considered to improve FHMM accuracy. In [17], a dynamic FHMM-based method is proposed as a new cloud-based online NILM service.

For event-based NILM, on the other hand, windows of *events* are first identified, where an event is defined as a sequence of power measurements, starting with a rising edge due to an appliance being switched on, or a multi-state appliance transiting to a higher power state, and ending with a falling edge, when an appliance is switched off, or a multi-state appliance returning to a lower-power state [43]. The identified events are then used to extract features (e.g., rising/falling power edge magnitude, area, time duration). Finally, classification based on trained models is performed on the extracted features. Examples of event-based NILM include: a hybrid K-means clustering of the training data followed by SVM-based energy disaggregation [10]; Decision Trees (DT)-based algorithm [29], Dynamic Time Warping (DTW) [29], GSP-based NILM [8], [27]; and a method

[2]The Python codes for the NILM approach presented in [27] can be accessed via https:\\github.com\loneharoon\GSP_energy_disaggregator.

of [25] based on maximum likelihood classifier with low sensitivity to noise.

Compared to event-based NILM approaches, state-based NILM approaches commonly require good a-priori information for initialization of appliance state models or a large training dataset for good performance [13], [31], [33]. Event-based NILM approaches, on the other hand, are often easier to implement and deploy due to data reduction via event feature extraction [8], [29]. However, as reported in [8] and [27], due to their reliance on edge detection, event-based NILM methods are susceptible to measurement noise and unknown appliances, and often misclassify appliances with similar operational power range.

In practice, sensor noise, transient spikes and signal fluctuations around an appliance's mean operating power inevitably appear in real-world electrical load meter measurements, as illustrated in Fig. 1. In Fig. 1(a), transient spikes are present due to sensor measurement noise and fast changes in the signal that cannot be captured at low rates. The 'fluctuations' of power level around its mean value during an appliance run can be observed in Fig. 1(b), which often results in the NILM algorithm misclassifying (part of) this appliance run as another appliance that operates at low power levels. In Fig. 1(c), a kettle is switched on/off, which causes 'pre/post state' signal fluctuations, common for appliances with a heating element. Fig. 1(d) shows a 'deep gap' caused by the simultaneous switching on of a PC and fridge-freezer.

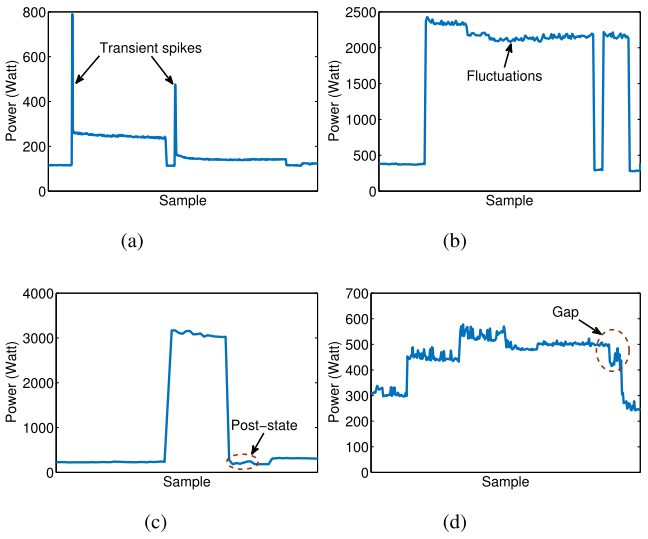


FIGURE 1. Typical "disaggregation" noise in electrical loads measurements, observed in House 17 of the REFIT dataset. (a) Transient spikes. (b) Fluctuations of the signal value around its mean. (c) Post-state noise. (d) Deep gap.

All above examples are considered to be "disaggregation noise" as they negatively affect NILM algorithm performance, which would ideally deal with rectangular-shaped signals with fairly unique amplitude values. For state-based NILM approaches, measured signal processing is based on signal smoothing with the main objective of removing outliers so that state conditions can clearly be identified. Event-based NILM approaches are more sensitive to abrupt state transitions [27], which make edges undetectable. Hence, the key to successful disaggregation lies in efficient smart meter signal processing to ensure that the edges are sharp enough to be effectively detected by edge detection and transients, that cannot be effectively captured with low sampling rates (of the order of 1-60 seconds), are removed. Otherwise, regardless of the effectiveness of the following classification step, the NILM output would not be sufficiently reliable.

We next review signal processing methods that have been proposed for NILM. Total variation regularization is used prior to the additive Factorial HMM (FHMM)-based NILM algorithm to remove outliers and minimize the influence of rarely used appliances in [12]. Median filtering is commonly used to remove spikes and noise, and smooth the signal [32], [33], [44]–[47] where the window size of the median filter is heuristically chosen based on the signal granularity. For example, median filtering is exploited twice in [45], for the purposes of total noise removal and partly for signal smoothing. Median filtering together with smoothing, is applied in [46] to an unsupervised HMM-based NILM approach. Various smoothing filters for NILM are investigated in [34], including median filter, mean filter, kernel-weighted average filter and the possible combination of multiple filters. Although kernel filter performs the best, median-mean filter is finally chosen due to the high complexity and cost of kernel filters. A neural network architecture performing dimensionality reduction is introduced in [22] as a denoising step. Down-sampling can also be regarded as a signal processing/noise removal method, since the majority of signal fluctuations are filtered out and most appliance state transition edges will be sharpened after down-sampling [8], [13], [27], [48]. In [48], the baseload signal is heuristically estimated and removed before GSP-based disaggregation.

Signal processing and "denoising" can lead to a cleaner signal that would improve subsequent classification. However, signal processing alone cannot solve the issue of similarity of appliance loads, that is, very close operational mean power values of two or more appliances. This can be addressed, for example, via an inference approximation method to refine NILM results, as done in [12], where Additive Factorial approximate MAP is proposed taking advantage of the additive structure of FHMM and the observation of aggregate power; or with a probabilistic search method as in [8], where simulated annealing is added after the primary GSP-based NILM to refine two-state (e.g., ON/OFF) appliance identification by optimizing the difference between the power measurements and the corresponding primary power estimate according to the possible combination of multiple appliances which are switched on simultaneously. In [27], each rising edge, which refers to a switching-on or upward state transition event, is matched with one nearest pair from the cluster with magnitude-wise closest falling edges. Methods of [8] and [27] have drawbacks in that: 1) these algorithms are sensitive to spikes and fluctuations;

2) low-load appliances and appliances with low operational states in multi-state appliances are often misclassified; 3) these algorithms usually perform poorly when disaggregating long-lasting, low-load appliances due to the difficulty of accurate detection and feature matching. This paper aims at addressing these drawbacks and providing a *universal* signal processing solution suitable for all event-based NILM methods.

III. GSP PRELIMINARIES AND NOTATION

In this section we introduce notation, summarized in Table 1, and give a brief overview of the concepts of GSP relevant for the rest of the paper. We closely follow notation from [27]. Bold upper-case and lower-case letters are used to denote matrices and vectors, respectively. The entry in the i-th row and j-th column of matrix $\mathbf{A}$ is denoted by $A_{i,j}$. x_i denotes the i-th element in vector $\mathbf{x}$. $A_{a_1:b_1,a_2:b_2}$, for $a_1 < b_1$ and $a_2 < b_2$, represents the sub-matrix in $\mathbf{A}$, with rows from a_1 to b_1 and columns from a_2 to b_2. Similarly, $\mathbf{x}_{a:b}$, for $a < b$, denotes the vector $[x_a, \ldots, x_b]$. $\mathbf{I}$ is the identity matrix, i.e., $\mathbf{I} = diag(1, \ldots, 1)$. Sets are denoted by upper-case calligraphic letters.

TABLE 1. Notation used in this paper.

Notation	
$\mathcal{V}$	set of N vertices
$\mathbf{A}$	$N \times N$ adjacency matrix
$\mathbf{s}$	$N \times 1$ graph signal
$\mathbf{D}$	$N \times N$ degree matrix
$\mathbf{L}$	$N \times N$ graph Laplacian operator
$\mathbf{P}$	aggregate active power signal
$\Delta\mathbf{P}$	differential power signal
$\mathbf{\Omega}$	window of samples
L_M	window size for median filtering
L_G	window size for graph filtering
T_G	threshold for bilateral graph filter
ρ	scaling factor for Gaussian kernel weighting
α	trade-off factor between fidelity and smoothness terms
L_E	window size for edge sharpening
T_E	threshold for edge sharpening
K	acceptable precision of a cluster
q	threshold for acceptable graph signal estimation
$\mathcal{M}$	set of known appliances in the house

Let $\mathcal{G} = (\mathcal{V}, \mathbf{A})$ be an undirected graph, where $\mathcal{V} = \{v_0, v_1, \ldots, v_{N-1}\}$ is a set of vertices and $\mathbf{A} \in \mathbb{C}^{N\times N}$ denotes a weighted $N \times N$ adjacency matrix [18], that is, $A_{i,j}$ corresponds to the weighted edge from v_i to v_j where the weight depends on the relationship between vertices v_i and v_j.

The Gaussian Kernel weighting function is often used to define the values of $A_{i,j}$, as in [8], [27], [36], and [37]. For example, in [8], the graph is designed to perform supervised classification of signal samples, where each signal sample x_i is indexed by a graph vertex v_i. The adjacency matrix then carries the information about correlation between signal samples and is defined as:

$$A_{i,j} = \exp\left(-\frac{\|x_i - x_j\|_2^2}{\rho^2}\right), \tag{1}$$

where ρ is a scaling factor.

On the other hand, in [35] and [37], the graph is designed to perform denoising of a time-series signal $\mathbf{x}$, where $A_{i,j}$ is set to zero if v_i and v_j are not time consecutive samples, that is:

$$A_{i,j} = \begin{cases} \exp\left(-\dfrac{||x_i - x_j||_2^2}{\rho^2}\right), & \text{for } |i-j| \le 1 \\ 0, & \text{for } |i-j| > 1. \end{cases} \tag{2}$$

A vector $\mathbf{s} \in \mathbb{R}^N$ is then defined as the *graph signal* that maps $\mathcal{V} \to \mathbb{R}$ [36], where each element s_i represents the function value at Node v_i.

Since the majority of natural signals are piecewise smooth, signal global smoothness is often used as a prior for regularization in different inverse problems [49], [50], and is defined as [51]:

$$S_p(\mathbf{s}) = \frac{1}{p}\sum_{i\in\mathcal{V}}\left[\sum_{j\in\mathcal{N}_i} A_{i,j}(s_j - s_i)^2\right]^{\frac{p}{2}}, \tag{3}$$

where $\mathcal{N}_i \subseteq \mathcal{V}$ is the set of vertices that are connected to v_i. For $p = 2$, we have the graph Laplacian quadratic form [35]. The global smoothness of a graph reflects the piecewise smoothness of the signal with respect to the underlying graph structure, i.e., if a graph signal is piecewise smooth, the global smoothness of its underlying graph is generally small.

Let $\mathbf{D}$ be a diagonal matrix with entries on the main diagonal given by $D_{i,i} = \sum_j A_{i,j}$. Let $\mathbf{L}$ denote the graph Laplacian operator [37] defined as: $\mathbf{L} = \mathbf{D} - \mathbf{A}$. Then $S_2(\mathbf{s})$ in (3) can be expressed in terms of the Laplacian matrix as [36]:

$$S_2(\mathbf{s}) = \frac{1}{2}\sum_{i,j} A_{i,j}\left(s_j - s_i\right)^2 = \mathbf{s}^T\mathbf{L}\mathbf{s}. \tag{4}$$

Since $\mathbf{s}^T\mathbf{L}\mathbf{s}$ will be generally small when $\mathbf{s}$ is piecewise smooth with respect to underlying graph structure, a graph total variation minimization $\arg\min_{\mathbf{s}}\left(\mathbf{s}^T\mathbf{L}\mathbf{s}\right)$, is often used to find the smoothest graph signal given training samples, $\mathbf{s}_{1:k}$, $k < N$, and has a closed-form solution [27]:

$$\mathbf{s}^* = \mathbf{L}^{\#}_{k+1:N,k+1:N}(-\mathbf{s}_{1:k})\mathbf{L}^T_{1:k,k+1:N}, \tag{5}$$

where $(.)^{\#}$ denotes the pseudo-inverse matrix.

IV. PROPOSED SIGNAL PROCESSING ALGORITHM FOR ENHANCING NILM

Let P_{t_i} be the household's aggregate active power consumption measurement at time instance t_i, for $i = 1, \ldots, N$. For simplicity, in the following, we assume a constant sampling rate and denote P_{t_i} by P_i. Then, we define the power variation signal between adjacent aggregate power measurements as $\Delta P_{t_i} = \Delta P_i = P_{i+1} - P_i$, for $i = 1, \ldots, N-1$. Similarly, let $P_{m_i} \ge 0$ be the active power consumption measurement for any appliance $m \in \mathcal{M}$ at time instant t_i, where $\mathcal{M}$ is the set of known appliances in the house, and $\Delta P_{m_i} = P_{m_{i+1}} - P_{m_i}$.

To ensure good event detection by NILM, it is essential to reduce the influence of noise while keeping signal edges sharp. Therefore, we design a signal processing

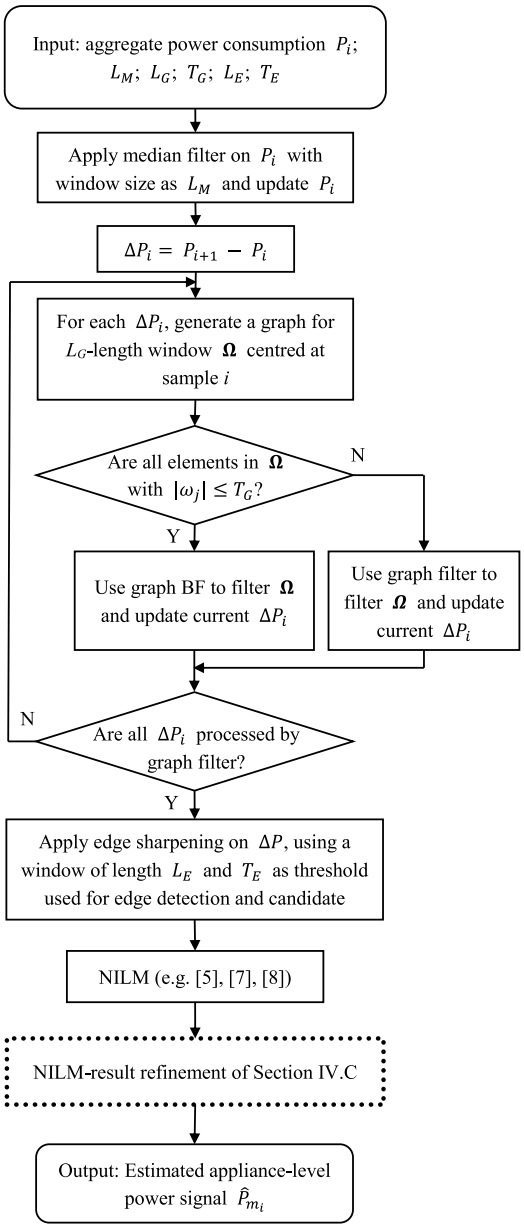


FIGURE 2. Flowchart of the proposed algorithm.

method, illustrated in Fig. 2, that takes advantage of the fact that the power signal should be piecewise smooth, i.e., jitter and spikes, due to sensor malfunction, transients and power noise, will be filtered out.

Firstly, L_M-length median filtering is applied to the aggregate power measurements P_i to remove outliers. L_M must be carefully chosen according to the signal attributes, such as granularity, to ensure that relevant events are not lost. Secondly, graph filtering is applied on ΔP_i to ensure piecewise smoothness of the power signal. Then, edge sharpening is used to merge unclear consecutive edges before NILM. Finally, NILM results for the appliances with similar operating power levels can be optionally refined.

A. GRAPH FILTER DESIGN

We perform graph filtering on overlapping sample windows $\mathbf{\Omega}$ of size L_G of the power variation signal ΔP_i obtained after median filtering. Either a graph filter via graph total variation regularization or its variant based on bilateral filtering (BF) is exploited, depending on the magnitude range of the data in the present window. That is, let T_G be a preset magnitude threshold; if all samples, P_j (that is, the output signal values of the median filter), in the current window $\mathbf{\Omega}$, meet the condition that $|P_j| \le T_G$, $j = 1, \ldots, L_G$, then $\mathbf{\Omega}$ will be filtered by a BF-based graph filter introduced in Subsection IV-A.2. Otherwise, the samples in window $\mathbf{\Omega}$ will be filtered by the graph filter presented in Subsection IV-A.1.

1) GRAPH FILTERING VIA TOTAL VARIATION REGULARIZATION

Let $\mathbf{x} = \mathbf{P}$ be an N-length vector of noisy power signal measurements (after median filtering).

Since $\mathbf{x}$ is a time-series signal, we design a graph $\mathcal{G} = (\mathcal{V}, \mathbf{A})$, where adjacency matrix $\mathbf{A}$ is given by (2). Then, the graph filtering-based denoising can be formulated as an optimization problem over all possible graph signals $\mathbf{s}$ [35]:

$$\arg\min_{\mathbf{s}} \frac{1}{2} \left\|\mathbf{s}-\mathbf{x}\right\|_2^2 + \alpha \frac{1}{2} \left\|\mathbf{s}-\mathbf{A}\mathbf{s}\right\|_2^2. \tag{6}$$

The cost function in (6) consists of a quadratic fidelity term to maintain similarity between denoised, output signal and the input signal $\mathbf{x}$, and the quadratic Laplacian smoothness term to guarantee global smoothness of the output graph signal. α in (6) is chosen to tradeoff fidelity and smoothness terms. The smoothness ensures that there are no sudden spikes in the signal that are characteristic of noisy measurement, appliance power value fluctuations around its mean that can cause misclassification, and transients. Thus, we want to find a piecewise smooth signal over the underlying graph, closest to the input signal $\mathbf{x}$.

(6) can be solved by calculating the first derivative of the cost function with respect to the filter input [35]:

$$\begin{aligned}
&\frac{\partial}{\partial x}\left(\frac{1}{2}\left\|\mathbf{s}-\mathbf{x}\right\|_2^2 + \alpha\frac{1}{2}\left\|\mathbf{s}-\mathbf{A}\mathbf{s}\right\|_2^2\right) \\
&\quad = \frac{1}{2}\frac{\partial}{\partial x}\left((\mathbf{s}-\mathbf{x})^*(\mathbf{s}-\mathbf{x}) + \alpha s^*(I-A)^*(I-A)s\right) \\
&\quad = (\mathbf{s}-\mathbf{x}) + \alpha(\mathbf{I}-\mathbf{A})^*(\mathbf{I}-\mathbf{A})\mathbf{s},
\end{aligned} \tag{7}$$

where $(.)^*$ denotes the conjugate transpose. If we set the first derivative in (7) to 0, we solve the minimization in (6) as:

$$\tilde{\mathbf{s}} = (\mathbf{I} + \alpha(\mathbf{I}-\mathbf{A})^*(\mathbf{I}-\mathbf{A}))^{-1}\mathbf{x}. \tag{8}$$

(8) is a closed form solution which presents the exact filtered graph signal. Since the computational complexity of (8) is $O(N^3)$, where N is the number of samples in $\mathbf{x}$, the filtering is not practical when dealing with large scale entries [35]. Hence, in practice, we perform filtering on sliding windows $\mathbf{\Omega}$ of manageable size L_G.

Note that for each input x_i, we generate a graph and perform filtering as described above. An example is shown in Fig. 3, where x_k is the central element in an L_G-length window $\mathbf{\Omega} = \mathbf{x}_{i:j}$, with $j - i = L_G - 1$ and $k = (i + j)/2$. Note that each vertex v_i of $\mathcal{G}$ corresponds to one power measurement x_i in $\mathbf{\Omega}$. Although $\tilde{\mathbf{s}}$ calculated by (8) gives the exact filtered signal for the whole $\mathbf{x}_{i:j}$, we only replace the central element, x_k, by the corresponding filtered solution $\tilde{s}_{\frac{L_G+1}{2}}$. Similarly, the filtered output for x_{k+1} will be generated using a new graph whose entries are $\mathbf{x}_{i+1:j+1}$, and so on.

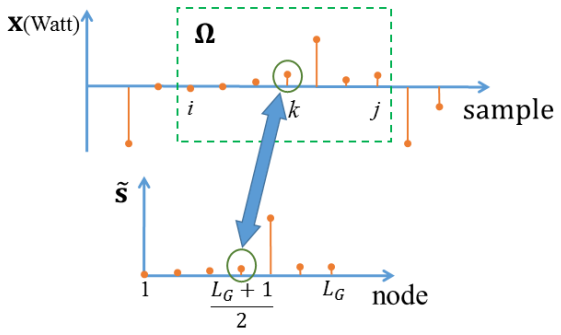


FIGURE 3. Example of graph filtering via total variation regularization, performed on window Ω of power measurements.

2) GRAPH-BASED BILATERAL FILTERING

Assume a graph $\mathcal{G} = (\mathcal{V}, \mathbf{A})$, with an adjacency matrix $\mathbf{A}$ defined by (2), and a diagonal degree matrix $\mathbf{D} \in \mathbb{C}^{N\times N}$ defined in Section III. For an input signal $\mathbf{s}$, the output signal after bilateral filtering is [38]:

$$\tilde{\mathbf{s}} = \mathbf{D}^{-1}\mathbf{A}\mathbf{s}, \tag{9}$$

where $\mathbf{D}^{-1}\mathbf{A}$ denotes the BF operator. Then by inserting the BF operator into (4), we obtain:

$$S_2(\mathbf{s}) = \frac{1}{2}\left\|\mathbf{s} - \mathbf{D}^{-1}\mathbf{A}\mathbf{s}\right\|_2^2. \tag{10}$$

Then (6) can be written as:

$$\arg\min_{\mathbf{s}} \frac{1}{2}\left\|\mathbf{s} - \mathbf{x}\right\|_2^2 + \alpha\frac{1}{2}\left\|\mathbf{s} - \mathbf{D}^{-1}\mathbf{A}\mathbf{s}\right\|_2^2. \tag{11}$$

The first derivative of the cost function in (11) is:

$$\begin{aligned}&\frac{\partial}{\partial x}\left(\frac{1}{2}\left\|\mathbf{s} - \mathbf{x}\right\|_2^2 + \alpha\frac{1}{2}\left\|\mathbf{s} - \mathbf{D}^{-1}\mathbf{A}\mathbf{s}\right\|_2^2\right)\\ &= \frac{1}{2}\frac{\partial}{\partial x}\left((\mathbf{s}-\mathbf{x})^*(\mathbf{s}-\mathbf{x}) + \alpha\mathbf{s}^*(\mathbf{I}-\mathbf{D}^{-1}\mathbf{A})^*(I-\mathbf{D}^{-1}A)\mathbf{s}\right)\\ &= (\mathbf{s}-\mathbf{x}) + \alpha(\mathbf{I}-\mathbf{D}^{-1}\mathbf{A})^*(\mathbf{I}-\mathbf{D}^{-1}\mathbf{A})\mathbf{s},\end{aligned} \tag{12}$$

and after setting the first derivative to zero we get:

$$\tilde{\mathbf{s}}_{BF} = (\mathbf{I} + \alpha(\mathbf{I}-\mathbf{D}^{-1}\mathbf{A})^*(\mathbf{I}-\mathbf{D}^{-1}\mathbf{A}))^{-1}\mathbf{x}. \tag{13}$$

Compared with (8), (13) contains operator weights averaged by neighbouring graph nodes and gives 'smoother' filtering results, useful for flattening small signal fluctuations. BF is commonly used in image denoising to smoothen/denoise each pixel by taking the weighted average of the nearby pixels [38]. However, the graph filter via BF will also split sharp edges into multiple segments, which will affect the edge detection accuracy. Therefore, we only apply BF to the entries in the windows which are magnitude-wise small, and which significantly suffer from signal fluctuations (see Fig. 1(b)).

An example is shown in Fig. 4, as a comparison of the distribution of the "noise", i.e., the difference between the total measured power and the sum of all known loads, (a) before and (b) after BF graph filtering. It is obvious that the difference is significantly reduced, including the standard deviation which is reduced from 236.79 to 211.52.

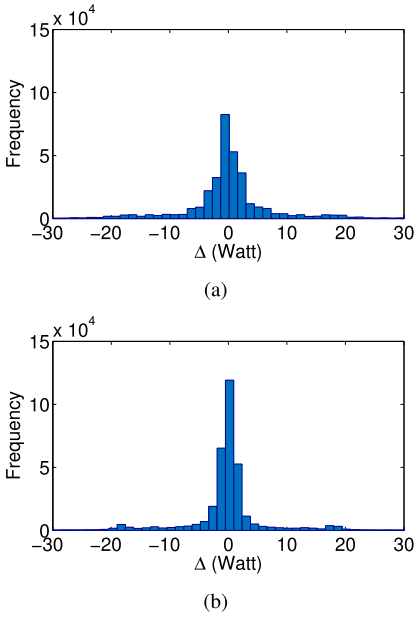


FIGURE 4. Histogram of $\Delta_i = \left|\Delta P_i - \sum_{m\in\mathcal{M}} \Delta P_{mi}\right|$ for House 17 of REFIT dataset. (a) Raw aggregate data (b) denoised data after graph filtering.

B. EDGE SHARPENING

The essential step in identifying events to be classified is edge detection. To ensure successful edge detection, the final signal processing step is applying edge sharpening to the graph-filtered power variation signal. Edge sharpening is used to merge the consecutive rising edges or the consecutive falling edges caused by state changes lasting more than one sample in the time-series power signal. It can be applied to any event-based NILM approach, e.g., [27], [29], as these NILM approaches rely on accurate edge magnitude information during feature extraction.

We perform edge sharpening on sample windows of size L_E samples, i.e., only the consecutive rising edges or falling edges within this window are allowed to be merged (see Fig. 2). T_E is the pre-set magnitude threshold for edge sharpening, i.e., only edges with magnitudes above T_E are merged.

C. SEPARATING SIMILAR LOADS VIA NILM-RESULT REFINEMENT

Some appliances have similar operating power levels, for example, bathroom GFI (1608W) and microwave (1526W) in REDD House 1 (see Fig. 5(a)); and Toaster (902W) and Freezer spikes (909W) in REFIT House 17. Consequently, they are misclassified, causing disaggregation errors [8], [27], [29]. In order to overcome this problem, we propose a post-classification refinement step. The proposed graph-based refining method is generic and applicable to any event-based NILM.

We assume that an event-based NILM algorithm was used to identify all appliance usage events and classify into appliances accordingly. The class associated with Appliance m contains two sets of detected rising and falling power signal edges that designate start and end of the appliance usage events, denoted by $\mathbf{\Pi}_P^m$ and $\mathbf{\Pi}_Q^m$, respectively. The task of the proposed method is then to refine the classification result using the knowledge acquired during the disaggregation process, by removing from $\mathbf{\Pi}_P^m$ and $\mathbf{\Pi}_Q^m$ power edges that do not belong to Appliance m and classifying them as Appliance n, which was labelled during NILM as the appliance with most similar load to Appliance m.

Firstly, for two edges classified as belonging to Appliance m, i.e., caused by Appliance m being switched on and off, $\Delta P_i \in \mathbf{\Pi}_P^m$ and $\Delta P_j \in \mathbf{\Pi}_Q^m$, we define a *feature vector* as:

$$l_{i,j}^m = \sqrt{\left(\Delta t_{i,j}\right)^2 + \left(|\Delta P_i| + |\Delta P_j|\right)^2}, \tag{14}$$

where $\Delta t_{i,j}$ is the time duration between the detected rising edge i and falling edge j. An example of $l_{i,j}^m$ is shown in Fig. 5, where two appliances were switched on/off – microwave (switched on three times around 12:55, 13:00 and 13:04) and bathroom GFI (switched on around 14:08).

Fig. 6 (top) shows disaggregated edges for Appliance m, containing rising edges (broken-line arrows) and falling edges (solid-line arrows). For each rising edge, all falling edges detected between that rising edge and the next rising edge (temporally) are regarded as candidates for the matched falling edge (dashed blocks in Fig. 6 (top)). A rising edge and a falling edge pair are a *match* if they represent the start and the end of the same appliance usage event.

For each rising edge $\Delta P_i \in \mathbf{\Pi}_P^m$, we generate a graph, $\mathcal{G}_i = (\mathcal{V}_i, \mathbf{A}_i)$, where each vertex $v_{i,j}, j > 1$ (shown as a green solid circle in Fig. 6 (bottom, left)) in the graph corresponds to a candidate falling edge with a signal value $l_{i,j-1}^m$ defined in (14). The reference vector $\overline{l_{i,j}^m}$ denoted by vertex $v_{i,1}$, on the other hand, corresponds to the average $l_{i,j}^m$ for all i and j of the Appliance m class. Note that the graph adjacency matrix is defined as (1), where $x_1 = \overline{l_{i,j}^m}$ and $x_n = l_{i,n-1}^m$ for $n > 1$.

For the constructed graph $\mathcal{G}_i$, we perform a graph total variation minimization as in (3), where $k = 1$ and $\mathbf{s}_{1:k}$ in (5), corresponds the reference vector and is set to 1, leading to the solution $\mathbf{s}^*$. Using a threshold $0 < q \leq 1$, if there exists a solution greater than q, then the matched falling edge for a rising edge ΔP_i will be set to $\Delta P_{\tilde{j}}$ with $s_{\tilde{j}}^* \geq s_j^*, \ \forall\, j$.

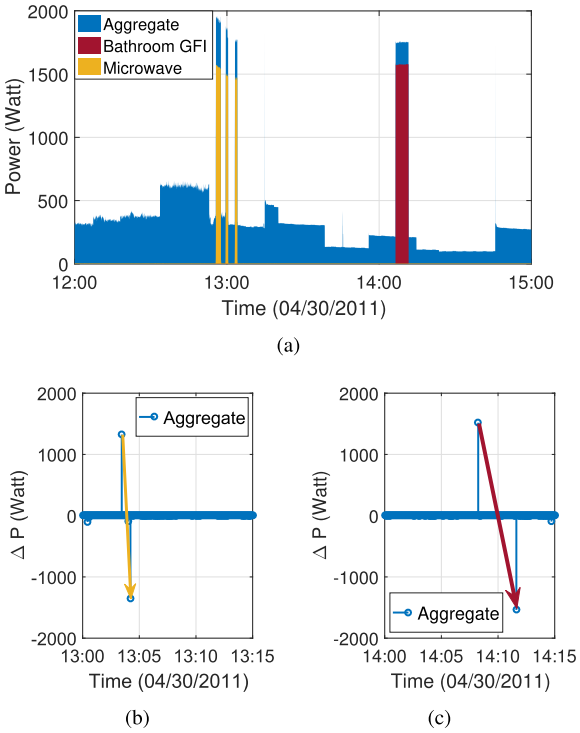


FIGURE 5. An example of vector $l_{i,j}^m$ for House 1 from the REDD dataset. (a) Aggregate power and individual load measurements for Bathroom GFI and microwave. (b) Power variation signal (ΔP_i) with a feature vector $l_{i,j}^m$ belonging to microwave (shown in yellow). (c) Power variation signal (ΔP_i) with a feature vector $l_{i,j}^m$ belonging to bathroom GFI (shown in red).

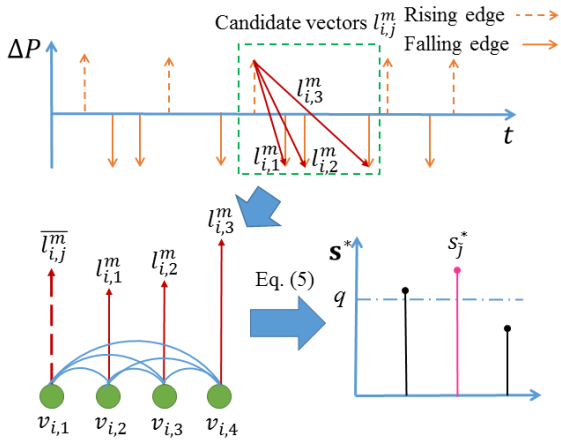


FIGURE 6. Graph generation. For a given rising edge, three vectors are selected as candidate falling edges (top). Graph with nodes corresponding to the reference vector and all candidate vectors (bottom, left). Graph total variation smoothness result $\mathbf{s}^*$, with the matched falling edge corresponding to $s_{\tilde{j}}^*$ where $\tilde{j} = 2$ (bottom, right).

Fig. 6 (bottom, right) gives the result calculated by (5), where the selected edge is shown in red. For the case when $s_j^* \leq q$, the rising edge ΔP_i and all candidates ΔP_j will be classified as Appliance n.

The procedure is independently repeated for all rising edges $\Delta P_i \in \mathbf{\Pi}_P^m$ in the window. Note that in view of efficiency and complexity, the proposed vector-feature

matching is only designed for refinement of a subset of loads likely to be misclassified, that is, for appliance classes with similar loads.

V. EXPERIMENTAL VALIDATION

In this section, our experimental results are reported. Specifically, the proposed signal processing methods are tested in conjunction with three low-complexity, state-of-the-art event-based NILM approaches: 1. Decision Tree (DT)-based NILM of [29], 2. supervised GSP-based (SGSP) NILM [8], and 3. unsupervised GSP-based (UGSP) NILM [27]. Additionally, we benchmark our proposed methods with the baseload removal (BR) pre-processing method of [48], also in conjunction with NILM algorithms of [8], [27], and [29]. We also compare aforementioned results with the state-of-the-art state-based NILM algorithms, i.e., FHMM-based and CO-based methods from the NILMTK [39] with median filtering for pre-processing.

We use active power readings from two open-access datasets - the US REDD dataset [40], at 1Hz resolution, and the UK REFIT raw dataset [41], collected every 6 to 8 seconds, both supported by the NILMTK. The REDD dataset is widely used for the evaluation of various NILM approaches in the literature, such as [29], [32], [33], and [47], and contains only few unknown loads. On the other hand, the REFIT dataset is a more realistic recent test-bed in the sense that it records smart meter aggregate power measurements from UK households continuously over a period of over two years while the occupants were uninterruptedly performing their daily routines. REFIT dataset is more challenging (see Table 6), as the houses contain noise due to numerous unknown appliances and some measurement noise. We selected two houses from each dataset with varying levels of unknown appliance influence.

For all results presented, experiments were carried out over a full month for both REDD and REFIT houses; specifically the period 18/04/2011 - 24/05/2011 for REDD House 1, 17/04/2011 - 22/05/2011 for REDD House 2, 01/10/2014 - 31/10/2014 for REFIT Houses 2 and 17. Note that the entire available dataset is used for two REDD houses. For REFIT houses, we use the same periods as in [8] and [27] for the purposes of benchmarking.

The abbreviations of domestic loads, using the labels provided in the datasets, that we disaggregate are as follows: B for Bathroom GFI; DW for Dishwasher; F for Fridge; KO for Kitchen outlet; L for Light; M for Microwave; O for Oven; WD for Washer dryer; S for Stove; FFZ for Fridge-freezer; K for Kettle; T for Toaster; WM for washing machine and FZ for Freezer.

In all results tables and figures, 'P' denotes the proposed algorithm, as illustrated in Fig. 2. For example, P-UGSP is the UGSP NILM method of [27] used in conjunction with the proposed method, where the NILM block of Fig. 2 is UGSP NILM [27].

Parameter values for the algorithms are chosen as discussed next (see Table 2). The window size for the median filter is defined as $L_M = 2\lfloor 15f \rfloor + 1$, where f is the dataset's sampling frequency in Hz. L_M corresponds to an odd window size of roughly 30 seconds to avoid false flattening of short operational states and false segmentation of edges. Thus, L_M is set to 31 and 5 for REDD and REFIT datasets, respectively. We heuristically fix L_G to 11 in graph filtering to trade off complexity and performance, exploiting, in this way, correlation between sample x_i with samples $\mathbf{x}_{i-5:i+5}$. We set $T_G = 50$W for all datasets, which implies that all detected power changes below 50W will be filtered using graph BF instead of graph filter. This significantly avoids over-smoothing and improves performance. ρ and α for graph filtering via BF, are heuristically fixed to 30 and 1, respectively. On the other hand, for graph filtering via total variation regularization, ρ and α are adaptively generated based on the input signal values. That is, for a current window $\mathbf{\Omega}$, $\rho = \max(\mathbf{\Omega}) * 0.3$ and $\alpha = \max(\mathbf{\Omega})/20000$. T_E, an edge magnitude threshold for edge sharpening, is set to 100W for all datasets to minimize the influence of measurement noise and magnitude-wise low loads which usually have rapid state transitions. L_E is empirically fixed as 5 to reduce complexity and maximize edge sharpening performance. In the proposed NILM-result refining step, ρ is set to 5 to achieve high classification precision. K and q are the same as in [27] for evaluating cluster quality for all appliances and reducing falsely clustered edges, respectively.

TABLE 2. Parameter setting used in the experiments.

Parameter	
L_M for REDD houses	31
L_M for REFIT houses	5
L_G	11
T_G	50 Watt
ρ in graph filtering via BF	30
α in graph filtering via total variation regularization	1
ρ in graph filtering via total variation regularization	max($\mathbf{\Omega}$)*0.3
α in graph filtering via BF	max($\mathbf{\Omega}$)/20000
T_E	100 Watt
L_E	5
ρ in NILM-resrefining	5
K	10%
q	0.98

The results are evaluated by comparing the output with the submetered measurements, as is common practice. The evaluation metrics used in this paper are *F-Measure* (F_M), Acc_m, *TER* and *DEM*, as described next.

1) F-MEASURE

The evaluation metrics, adapted from [11] and [27], are *Precision* (*PR*), *Recall* (*RE*) and *F-Measure* (F_M), and defined as:

$$PR = ATP/(ATP + FP) \tag{15}$$

$$RE = ATP/(ATP + ITP + FN) \tag{16}$$

$$F_M = 2 \cdot (PR \cdot RE)/(PR + RE), \tag{17}$$

where accurate *true positive* (*ATP*) denotes the correct claim the detected appliance was used and the corresponding events

are correctly named; *inaccurate true positive* (*ITP*), on the other hand, denotes the correct claim the detected appliance was used but the corresponding events are falsely named; *false positive* (*FP*) represents an incorrect claim that the appliance was not used; and *false negative* (*FN*) indicates that the appliance operational events were not detected. *PR* represents the event detection accuracy where high *PR* reflects low *FP*, and *RE* represents the events detection strength and clustering accuracy where lower *FN* and *ITP* result in a higher *RE*. F_M balances *PR* and *RE*. To compare results with the state-based NILM methods such as FHMM, confusion matrices are built for ON/OFF state transitions as in [11].

2) DISAGGREGATION ACCURACY

The disaggregation accuracy metric for Appliance m, is defined as:

$$Acc_m = 1 - \frac{\sum_{i=1}^{N} |\hat{P}_{m_i} - P_{m_i}|}{2\sum_{i=1}^{N} P_{m_i}}, \tag{18}$$

where N is the number of samples, P_{m_i} and $\hat{P}_{m_i}$ refer to the measured power of Appliance m at time instant i and its estimated value after disaggregation, respectively. This evaluation metric, demonstrating for each appliance, the error between actual power consumption and its estimate, is used in [22], [33], [40], and [47].

3) ERROR IN ESTIMATING TOTAL POWER CONSUMPTION

In addition to the above two measures, we introduce a metric, complimentary to disaggregation accuracy, to further explain performance of the proposed methods on NILM algorithms. The error rate of total power consumption (*TER*) measure, using the same notation as above, is defined as:

$$TER = \left| \sum_{i=1}^{N} P_{m_i} - \sum_{i=1}^{N} \hat{P}_{m_i} \right| / \sum_{i=1}^{N} P_{m_i}, \tag{19}$$

4) DISAGGREGATION ERROR MEASURE

We also introduce the Disaggregation Error Measure (DEM) to assess overall performance with respect to the noise measure (NM) of [15] to understand the performance of the proposed algorithms in correlation with the 'noisiness' of the dataset:

$$DEM = \frac{\sum_{i=1}^{N} \left| P_i - \hat{P}_{base} - \sum_{m=1}^{\mathcal{M}} \hat{P}_{m_i} \right|}{\sum_{i=1}^{N} P_i}, \tag{20}$$

where $\hat{P}_{base}$ refers to the estimated base-load.

In Subsection V-A, we present results with and without the optional NILM-result refinement block, described in Subsection IV-C. Then, in Subsection V-B, we present the F_M and *Acc* results for the proposed method shown in Fig. 2 with aforementioned benchmark algorithms. Finally, in Subsection V-C, we discuss results obtained with the TER and DEM metrics to shed further insight on the effect of the proposed methods and their impact on the NILM algorithm performance.

A. DISAGGREGATION RESULTS USING THE PROPOSED NILM-RESULT REFINEMENT APPROACH

In this section, we highlight the performance of our proposed NILM-result refinement approach of Subsection IV-C for separating loads that are similar power-wise. From Table 3, we can see that F_M and Acc_m improve by 0.13 and 0.07, respectively, on average, for all three disaggregation methods for similar appliance loads from REDD houses.

For REFIT houses, an average improvement of 0.08 for both F_M and Acc_m is observed. Significant improvement in disaggregation of Bathroom GFI in REDD House 1 and Dishwasher in REFIT House 2 can also be observed. The proposed NILM-result refinement method further refines the clusters containing state transition events due to Bathroom GFI and Microwave, but labelled as Bathroom GFI. Similarly, low-state transition events of Dishwasher from REFIT House 2 are further separated from the clusters of events labelled as Fridge-freezer during the proposed refinement step. REFIT House 17 contains an unknown appliance whose power level is similar to Kettle, leading to high *ITP* for Kettle. With the proposed GSP-based feature matching refinement, some events of Kettle are isolated and correctly labelled

TABLE 3. Performance of the proposed NILM-result refinement for multiple appliances. N refers to disaggregation with the proposed methods before NILM-result refinement in Fig. 2. P refers to disaggregation with all steps of Fig. 2, including NILM-result refinement (discussed in Section IVC).

House			REDD House 1		REDD House 2		REFIT House 2		REFIT House 17		
Appliance			B	DW	KO1	M	FFZ	DW	WM	K	FFZ
UGSP	F_M	N	0.04	0.39	0.72	0.64	0.28	0.38	0.74	0.81	0.48
		P	0.41	0.52	0.82	0.69	0.42	0.79	0.76	0.84	0.5
	Acc_m	N	0.15	0.42	0.77	0.76	0.58	0.4	0.5	0.68	0.61
		P	0.44	0.66	0.89	0.82	0.77	0.42	0.53	0.79	0.66
SGSP	F_M	N	0.53	0.54	0.81	0.73	0.39	0.68	0.74	0.96	0.79
		P	0.58	0.63	0.9	0.83	0.59	0.73	0.77	0.96	0.82
	Acc_m	N	0.51	0.71	0.87	0.84	0.81	0.51	0.57	0.78	0.67
		P	0.53	0.72	0.86	0.85	0.8	0.67	0.61	0.8	0.7
DT	F_M	N	0.36	0.44	0.73	0.51	0.33	0.73	0.79	0.94	0.83
		P	0.44	0.57	0.85	0.78	0.54	0.73	0.78	0.95	0.82
	Acc_m	N	0.31	0.56	0.86	0.71	0.7	0.35	0.48	0.72	0.56
		P	0.39	0.58	0.86	0.76	0.73	0.61	0.52	0.77	0.67

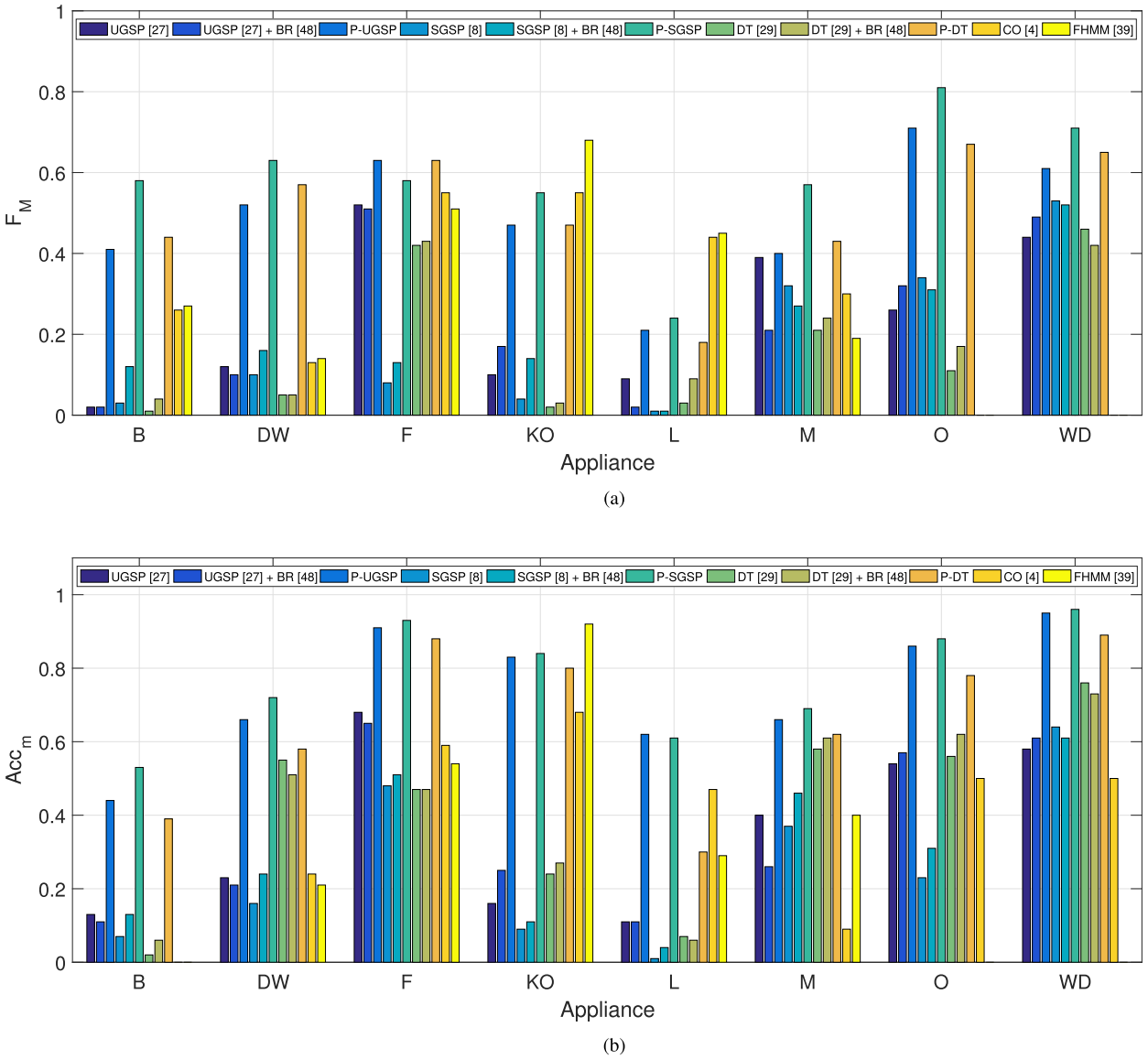


FIGURE 7. Performance of proposed method (P) with benchmarks UGSP [27], UGSP [27] with BR [48], SGSP [8], SGSP [8] with BR [48], DT [29], DT [29] with BR [48], CO and FHMM for REDD House 1.

as *ATP* which resulted in corresponding increase in both F_M and Acc_m.

From this point in the paper, proposed method (P) refers to the scheme with the NILM-result refinement, i.e., with all steps of Fig. 2.

B. F_M AND ACC_M APPLIANCE-LEVEL RESULTS

The performance of the proposed and benchmark methods for House 1 from REDD dataset is shown in Fig. 7. The performance improvement due to the proposed algorithm is noticeable for all eight appliances for all three event-based NILM approaches, with average F_M improvement of 0.25 for P-UGSP over [27], 0.4 for P-SGSP over [8] and 0.34 for P-DT compared to [29]. Similarly, an average Acc_m improvement of 0.39 for P-UGSP, 0.51 for P-SGSP and 0.25 for P-DT is observed compared to their respective benchmarks. Worth noting is the oven (O), a significantly large load, which could hardly be disaggregated without the proposed methods with $F_M \leq 0.4$, but with the proposed method can achieve F_M classification accuracy between 0.7 and 0.8. This is because the state transition duration in the case of oven is often longer than the sampling period (1sec), affecting edge detection, that is, instead of detecting a single rising/falling edge due to the oven being switched on/off, multiple edges of smaller amplitude will be detected causing subsequent classification errors. However, the proposed GSP-filtering based method merges the segmented edges into a distinct edge transition. This is shown in Fig. 12 (a).

Similar results are obtained for REDD House 2, as can be seen from Fig. 8. In this case, the average Acc_m results among

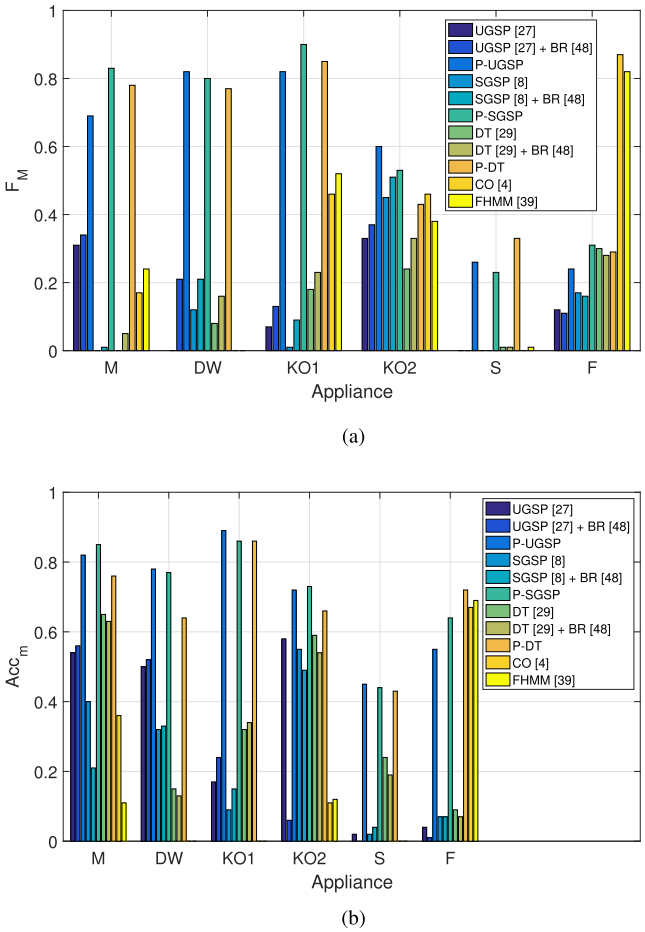


FIGURE 8. Performance of proposed method (P) with benchmarks UGSP [27], UGSP [27] with BR [48], SGSP [8], SGSP [8] with BR [48], DT [29], DT [29] with BR [48], CO and FHMM for REDD House 2.

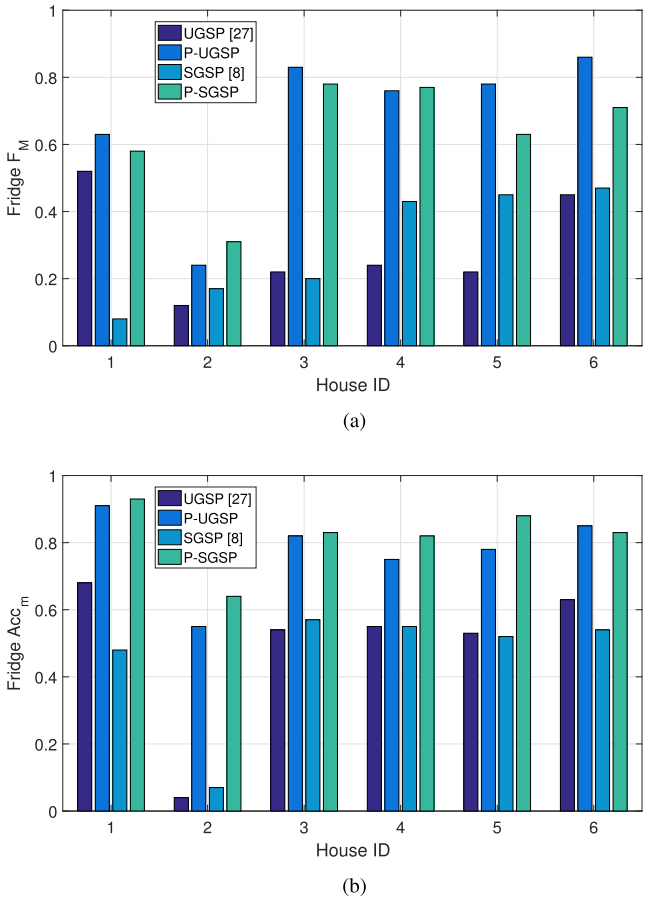


FIGURE 9. Performance of the proposed method with benchmarks UGSP [27], SGSP [8] for fridge in each house from REDD dataset.

all appliances with proposed method are 0.7 for P-UGSP, 0.73 for P-SGSP, and 0.68 for P-DT, which are significantly better than the benchmarks and competitive with NILM methods based on FHMM and its extensions [33], bearing in mind that the period of testing in [33] is unreported.

Furthermore, from Fig. 7 and 8, it can be seen that removal of the baseload (the BR method in [48]) does not show performance improvement compared to the results without BR, and may lead to worse performance for some appliances such as, Washer Dryer (WD) for disaggregation based on SGSP and DT, and Microwave (M) for disaggregation based on UGSP. This can be explained as follows: with data granularity of 1 sec, the appliance operational state transition edges can be captured by the BR method as consecutive smaller edges in power measurements; since BR identifies small power changes under 50 Watts as fluctuations of the baseload, some small discrete edges will be removed, leading to worse classification results. The performance of state-based NILM methods such as FHMM-based methods is possibly affected as low-value states will be removed while small edges are removed.

Similar results were obtained for other REDD houses, shown in Figs. 9 and 10, for four regularly used appliances present in all 6 REDD houses. We particularly demonstrate appliance-level NILM results for fridge in Fig. 9 as fridge is a major consumer present in all houses. Significant improvements can be observed, 0.34 for F_M and 0.32 for Acc_m on average across houses. House-level results presented in Fig. 10 show F_M and Acc improved by 0.35 and 0.54, respectively, with the proposed methods. Note that the accuracy obtained with the proposed methods varies only slightly across the houses. However, a much larger variation in performance is observed in the absence of the proposed method demonstrating robustness and consistency of the proposed approach to varying noise levels across houses.

The results in Fig. 10 are comparable to those of [15], [25], and [47], taking into account different sampling rates, composition of subset of appliances for house-level accuracy calculation, and number of days/samples (sometimes unknown) used for training/testing.

The impact of the proposed algorithm on multiple NILM methods for REFIT House 2 is shown in Fig. 11. House 2 is a typical house in the REFIT dataset with two multi-state appliances. The F_M performance improvement of UGSP with and without the proposed method is significant for most appliances, except Fridge-freezer (FFZ) and Washing

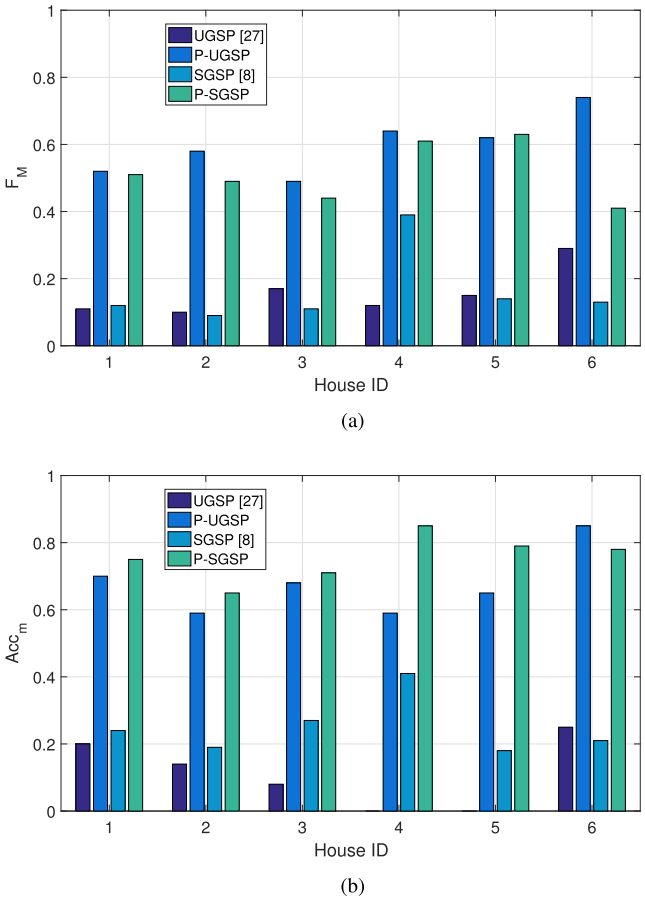


FIGURE 10. Performance of the proposed method with benchmarks UGSP [27], SGSP [8] for the four most common appliances (i.e., B, DW, F, WD) in each house from REDD dataset.

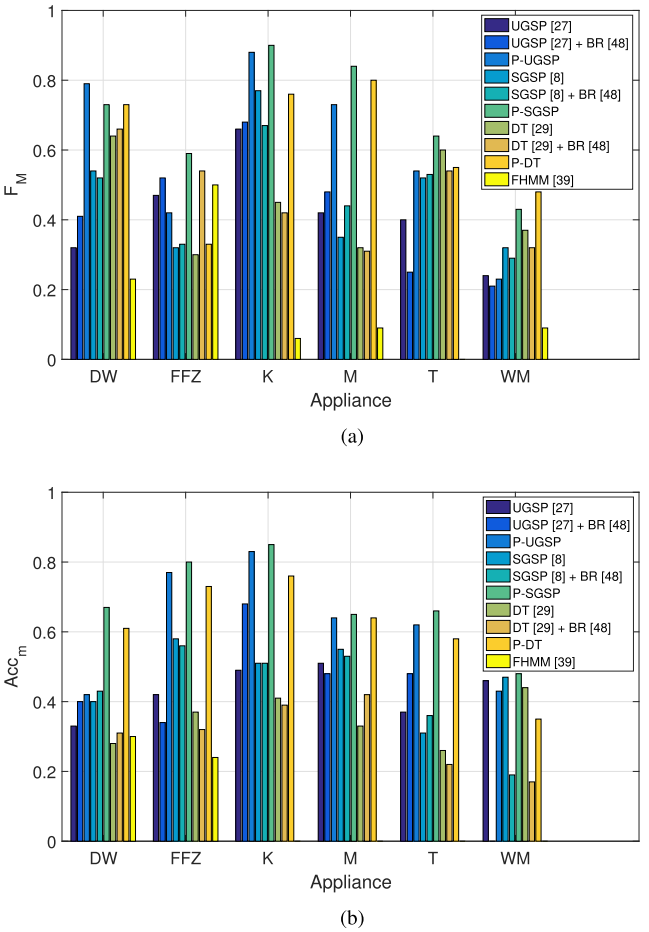


FIGURE 11. Performance of proposed method (P) with benchmarks UGSP [27], UGSP [27] with BR [48], SGSP [8], SGSP [8] with BR [48], DT [29], DT [29] with BR [48] and FHMM for REFIT House 2.

Machine (WM). Note that although the F_M of FFZ reduced a little by 0.05, caused by reduction of correctly identified state transition edges, many falsely identified edges are avoided, resulting in 0.35 improvement in Acc_m and reduction in power consumption estimation error (see Subsection V-C). The WM in REFIT House 2 is the only appliance for which the results have not improved in Acc_m, for two reasons - (a) a low number of WM events during the testing period of one month (16 runs in total); (b) large fluctuation of WM power load, resulting in poor performance for all tested methods.

Note that the largest improvement is observed for the Dishwasher (DW) because in addition to the state transition edges being sharper for DW, multiple operational states of DW are well shaped after the proposed processing, enabling low-load events to be captured and distinguished from the cluster of Fridge-freezer. This is illustrated in Fig. 12(c), which also shows an example of the raw and processed signal of Fridge-freezer; sharp spikes are removed, reducing the edge magnitude range and improving this way the edge detection accuracy and classification precision.

There is a significant improvement in the SGSP algorithm after applying the proposed methods, where F_M improves by 0.22 on average for all appliances. Similarly, the performance improvement between DT with and without the proposed method is noticeable for all appliances, besides Toaster, in REFIT House 2. Apart from DW in REFIT House 2, no significant improvement is observed for the case of multi-state appliances because the edges are not sufficiently denoised and sharpened as in the case of the REDD houses. This can be explained by the higher sampling rate of 1Hz for the REDD dataset, where there exists more state changes longer than 1 second which are split into multiple small segments; hence, the proposed method leads to higher gains. As for the REDD houses, baseload removal seems counterproductive for REFIT House 2 for most appliances. This is due to the consecutive low state signatures being removed during BR [48].

The FHMM-based NILM approach does not perform as well as the three event-based NILM approaches on REFIT House 2 except for Fridge, since FHMM is good at identifying appliance operation cycles of fridges.

We present the performance of the proposed method for House 17 from REFIT dataset in Fig. 13. In REFIT House 17, Washing Machine (WM) regularly contains significant power variations, resulting in a high number of *FP*s for magnitude-wise similar appliances, including Fridge-Freezer (FFZ), Freezer (FZ) and Toaster (T). Performance

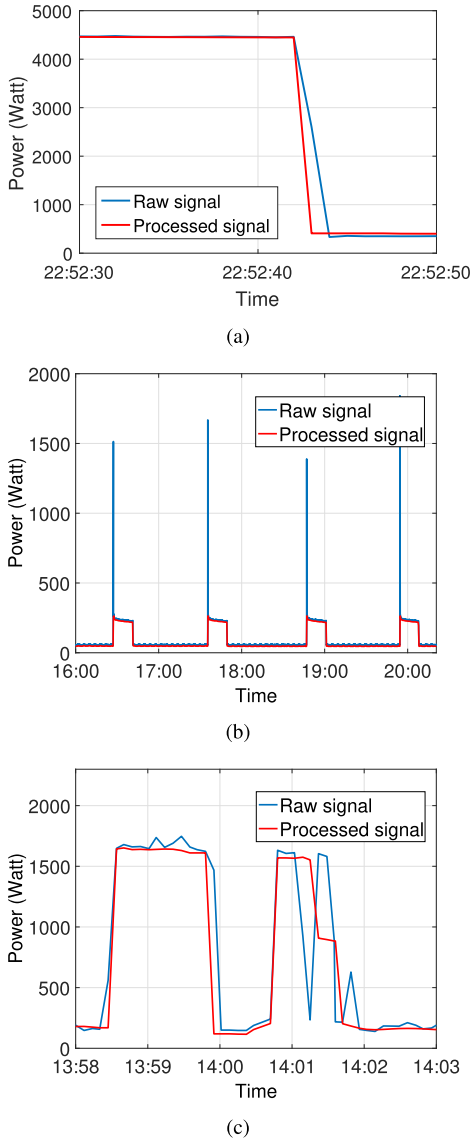


FIGURE 12. Typical appliance operation in the aggregate power consumption data before (shown in blue) and after processing via the proposed method (shown in red). (a) Oven REDD House 1; (b) Fridge-freezer REFIT House 2; (c) Dishwasher REFIT House 2.

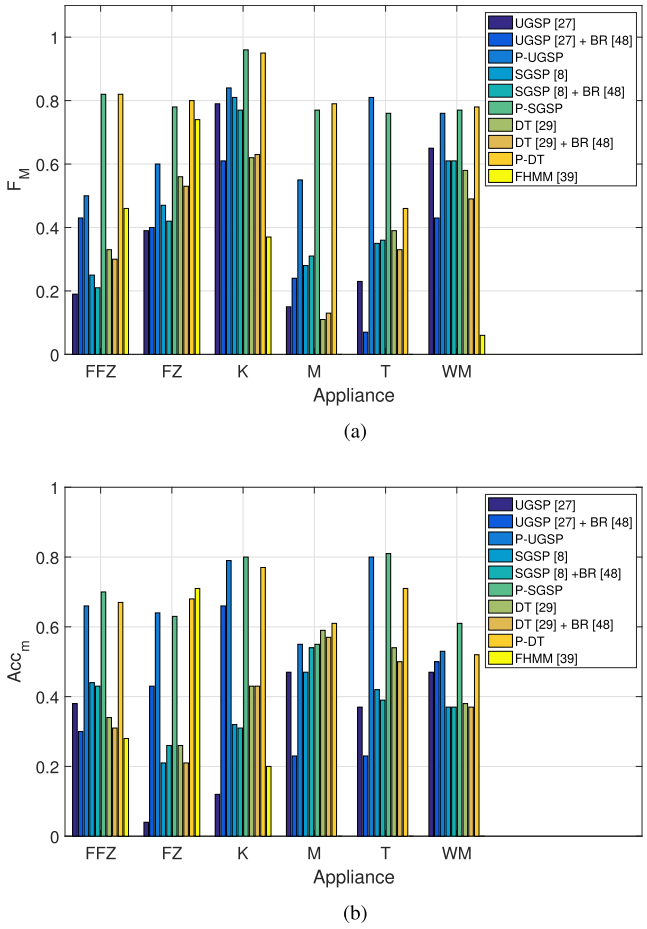


FIGURE 13. Performance of proposed method (P) with benchmarks UGSP [27], UGSP [27] with BR [48], SGSP [8], SGSP [8] with BR [48], DT [29], DT [29] with BR [48], CO and FHMM for REFIT House 17.

improvement of FFZ, FZ and T is noticeable in Fig. 13, since fluctuations of WM power signal are effectively reduced by the proposed method. Note that some very poor Acc_m results of FHMM were omitted. The performance of our proposed methods is comparable with the results in [17], which uses the cleaned version of REFIT data, where measurement errors were removed from the raw dataset as per [41]. Our results therefore show that with the proposed signal processing of the raw measured signal, which would be available directly from the meter, comparable results can be obtained if a cleaned, measurement-error-free signal was used instead.

C. FURTHER INSIGHTS WITH ADDITIONAL METRICS

Table 4 shows that the normalised error between actual and total power consumption, *TER* as defined in (19), for all

TABLE 4. Normalised total power consumption estimation error (*TER*) per appliance of proposed method (P) benchmarked against UGSP [27], UGSP [27] with BR [48], SGSP [8], SGSP [8] with BR [48], DT [29], DT [29] with BR [48] and FHMM for REFIT House 2.

Appliance	DW	FFZ	K	M	T	WM
UGSP [27]	0.73	0.1	0.34	0.43	0.83	0.17
UGSP [27] + BR [48]	0.08	0.53	0.06	0.17	0.81	0.4
P-UGSP	0.66	0.31	0.05	0.34	0.82	0.09
SGSP [8]	0.47	0.19	0.27	0.52	0.9	0.32
SGSP [8] + BR [48]	0.13	0.42	0.11	0.16	0.88	0.52
P-SGSP	0.33	0.27	0.04	0.28	0.73	0.1
DT [29]	0.64	0.47	0.31	0.66	0.78	0.28
DT [29] + BR [48]	0.35	0.67	0.25	0.59	0.71	0.33
P-DT	0.49	0.31	0.15	0.63	0.42	0.22
FHMM	0.75	0.38	0.41	13.63	7.53	3

three NILM methods, is reduced with the proposed methods for most appliances. Our proposed method not only removes outliers and spikes, but also reshapes edges for better feature matching results. The relatively large estimation errors of FHMM-based method for Microwave, Toaster and Washing machine reflect over-estimation, which matches the corresponding poor performance of F_M and Acc_m in Fig. 11. Note that although BR significantly reduces the error for

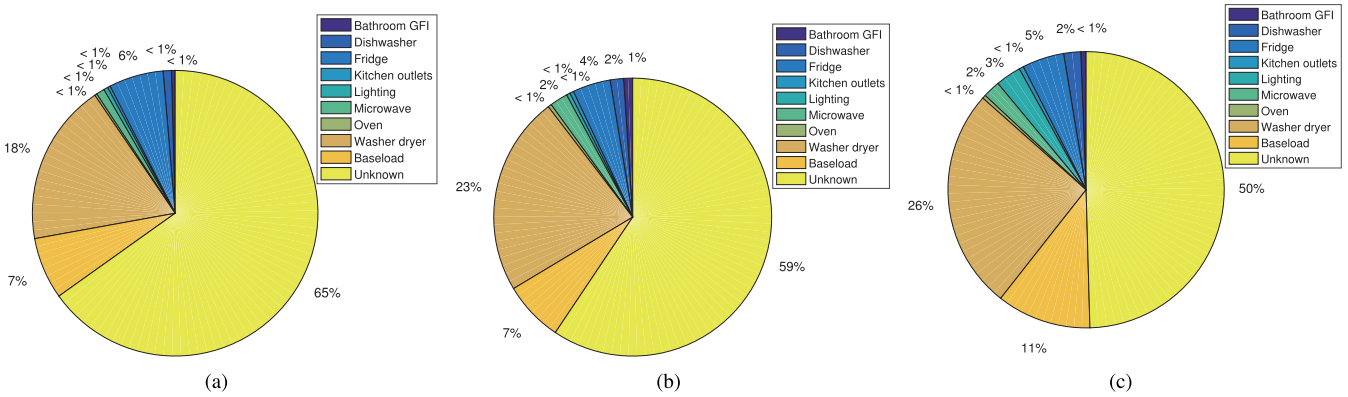


FIGURE 14. Disaggregation results, presented as percentage load contribution per appliance relative to the aggregate load, for House 1 from REDD datasets, where (a) results after disaggregation with UGSP [27], (b) results after disaggregation with P-UGSP, and (c) ground truth.

Dishwasher, it is at the expense of increased error for Fridge-freezer and Washing machine, since these two appliances are often confused with Dishwasher. Overall, as observed in the previous section, BR does not show performance improvement for the majority of appliances.

Similar results are obtained for the REDD houses. Indeed, for REDD House 1, the difference between the disaggregated energy consumption and measured energy consumption per appliance is small as illustrated in Fig. 14 and Table 5 for UGSP in [27]. The disaggregation results using pure UGSP [27], shown in Fig. 14 (a), are inline with the argument that the algorithm suffers from high sensitivity to disaggregation noise. With our proposed method, the total disaggregated energy is increased by 6%.

TABLE 5. Normalised total power consumption estimation error (*TER*) of UGSP [27] and P-UGSP for REDD House 1.

Appliance	B	DW	F	KO	L	M	O	WD	BL
[27]	0.32	0.53	0.26	0.41	0.84	0.53	0.37	0.29	0.37
P	0.7	0.22	0.1	0.15	0.86	0.03	0.01	0.11	0.37

Similar to Fig. 7, it can be observed from Table 5 that for the majority of appliances, applying the proposed method to the event-based NILM approach from [27] reduces *TER*. The *TER* increase for Bathroom GFI is due to a few microwave events misclassified as the events of Bathroom GFI, resulting in overestimation and high *TER*. With the benchmark, fewer events can be matched as operational state edge pairs, leading to lower estimated total consumption and lower *TER*. Both F_M and Acc_m results of Bathroom GFI in Fig. 7 indicate more accurate classification with the proposed method.

Table 6 compares four selected houses, two from REDD and two from REFIT dataset, in terms of disaggregation noise using the noise measure (NM) of [15]. It can be seen that both REFIT houses have much higher noise level compared to the REDD houses, making the disaggregation process very challenging. In addition, comparison of overall disaggregation performance for the selected houses is shown. Note that the overall F_M values in Table 6 are calculated using (15), (16) and (17), by summing *TP*, *FP* and *FN* values for all appliances. From Table 6, disaggregation performance for REDD houses is generally better than that of REFIT houses, with higher F_M and lower DEM. Note that though REDD House 2 is 'noisier' than REDD House 1 (based on NM inline with DEM values), the overall F_M performance for House 2 is better than that for House 1. This is because REDD House 1 contains more appliances than House 2, making the disaggregation problem more complex. For both REFIT houses, multiple disaggregation methods give DEM values which are 0.17 higher than those of REDD houses on average. Although NM values for REFIT houses are almost twice of those for REDD houses, corresponding overall F_M results are only 0.1 lower on average. One reason for this is fluctuation in baseload in REFIT houses, which increases NM, but does not significantly affect event-based NILM methods as only state transitions are considered.

TABLE 6. F-measure performance for four selected houses.

Dataset		REDD		REFIT	
House		1	2	2	17
NM		0.24	0.35	0.68	0.58
DEM	P-UGSP	0.37	0.54	0.67	0.66
	P-SGSP	0.44	0.57	0.66	0.61
	P-DT	0.46	0.56	0.69	0.64
F_M	P-UGSP	0.67	0.72	0.59	0.53
	P-SGSP	0.69	0.73	0.61	0.62
	P-DT	0.7	0.68	0.59	0.6

VI. CONCLUSION AND FUTURE WORK

This paper addresses the challenging problem of mitigating the effect of measurement noise and unknown loads on load disaggregation (NILM) performance. Two signal processing methods based on GSP are proposed in conjunction with existing NILM approaches to improve any low-rate supervised and unsupervised event-based NILM classification and estimation accuracy. In particular, we propose a graph-based filtering approach to clean the power signal before

classification. The main motivation comes from the fact that event-based low-rate NILM approaches require clean power consumption measurements containing sharp and accurate state transition events. Besides the proposed graph-based filtering, we proposed a post-classification refinement method, which improves NILM by mitigating the effect of misclassification of loads with similar operational range.

We demonstrate the improvement in NILM performance with the proposed methods when applied to three distinct event-based NILM methods and across two real-world datasets with multiple houses with different levels of actual measured noise. The effect of sampling rate on graph filtering and edge sharpening is discussed and we show that the proposed method can significantly improve performance for smart meter data gathered at sampling rates of 1Hz and lower, allowing appliances which could not be detected by event-based NILM previously, possible.

Future work could include: investigation of adaptive parameter selection derived from measurements and robustness enhancement by applying the proposed algorithm to other energy usage measurements (e.g., reactive power) with various sampling rates.

REFERENCES

[1] E. Fadel *et al.*, "Spectrum-aware bio-inspired routing in cognitive radio sensor networks for smart grid applications," *Comput. Commun.*, vol. 101, pp. 106–120, Mar. 2017.

[2] M. Faheem and V. C. Gungor, "Capacity and spectrum-aware communication framework for wireless sensor network-based smart grid applications," *Comput. Standards Interfaces*, vol. 5, pp. 48–58, Aug. 2017.

[3] "Smart metering equipment technical specifications version 2," Dept. Energy Climate Change, London, U.K., Tech. Rep., 2013. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/68898/smart_meters_equipment_technical_spec_version_2.pdf

[4] G. W. Hart, "Nonintrusive appliance load data acquisition method," Energy Lab., MIT, Cambridge, MA, USA, Tech. Rep., Sep. 1984.

[5] K. C. Armel, A. Gupta, G. Shrimali, and A. Albert, "Is disaggregation the holy grail of energy efficiency? The case of electricity," *Energy Policy*, vol. 52, no. 1, pp. 213–234, Jan. 2013.

[6] J. Kelly and W. Knottenbelt, "Does disaggregated electricity feedback reduce domestic electricity consumption? A systematic review of the literature," in *Proc. 3rd Int. NILM Workshop*, May 2016, pp. 1–5.

[7] 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. 16838–16866, Dec. 2012.

[8] K. He, L. Stankovic, J. Liao, and V. Stankovic, "Non-intrusive load disaggregation using graph signal processing," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 1739–1747, May 2018.

[9] R. Bonfigli, S. Squartini, M. Fagiani, and F. Piazza, "Unsupervised algorithms for non-intrusive load monitoring: An up-to-date overview," in *Proc. IEEE 15th Int. Conf. Environ. Elect. Eng. (EEEIC)*, Roma, Italy, Jun. 2015, pp. 1175–1180.

[10] H. Altrabalsi, V. Stankovic, J. Liao, and L. Stankovic, "Low-complexity energy disaggregation using appliance load modelling," *AIMS Energy*, vol. 4, no. 1, pp. 884–905, Jan. 2016.

[11] H. Kim, M. Marwah, M. Arlitt, G. Lyon, and J. Han, "Unsupervised disaggregation of low frequency power measurements," in *Proc. 11th SIAM Int. Conf. Data Mining*, Mesa, AZ, USA, Apr. 2011, pp. 747–758.

[12] J. Z. Kolter and T. Jaakkola, "Approximate inference in additive factorial HMMs with application to energy disaggregation," in *Proc. 15th Int. Conf. Artif. Intell. Stat.*, vol. 22, 2012, pp. 1472–1482.

[13] O. Parson, S. Ghosh, M. Weal, and A. Rogers, "Non-intrusive load monitoring using prior models of general appliance types," in *Proc. 26th Conf. Artif. Intell. (AAAI)*, Toronto, CA, USA, Jul. 2012, pp. 1–7.

[14] D. Egarter, V. P. Bhuvana, and W. Elmenreich, "PALDi: Online load disaggregation via particle filtering," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 2, pp. 467–477, Feb. 2015.

[15] S. Makonin, F. Popowich, I. V. Bajić, B. Gill, and L. Bartram, "Exploiting HMM sparsity to perform online real-time nonintrusive load monitoring," *IEEE Trans. Smart Grid*, vol. 7, no. 6, pp. 2575–2585, Nov. 2016.

[16] M. Zhong, N. Goddard, and C. Sutton, "Latent Bayesian melding for integrating individual and population models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3618–3626.

[17] M. A. Mengistu, A. A. Girmay, C. Camarda, A. Acquaviva, and E. Patti, "A cloud-based on-line disaggregation algorithm for home appliance loads," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2018.2826844.

[18] J. Kolter, S. Batra, and A. Y. Ng, "Energy disaggregation via discriminative sparse coding," in *Proc. 24th Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, Dec. 2010, pp. 1153–1161.

[19] H. Shao, M. Marwah, and N. Ramakrishnan, "A temporal motif mining approach to unsupervised energy disaggregation," in *Proc. 1st Int Workshop Non-Intrusive Load Monit.*, Pittsburgh, PA, USA, May 2012, pp. 1–2.

[20] A. G. Ruzzelli, C. Nicolas, A. Schoofs, and G. M. P. O'Hare, "Real-time recognition and profiling of appliances through a single electricity sensor," in *Proc. IEEE SECON*, Boston, MA, USA, Jun. 2010.

[21] S. Biansoongnern and B. Plangklang, "Nonintrusive load monitoring (NILM) using an artificial neural network in embedded system with low sampling rate," in *Proc. 13th Int. Conf. Elect. Eng./Electron., Comput., Telecommun. Inf. Technol. (ECTI-CON)*, Chiang Mai, Thailand, Jun./Jul. 2016, pp. 1–4.

[22] J. Kelly and W. Knottenbelt, "Neural NILM: Deep neural networks applied to energy disaggregation," in *Proc. 2nd ACM Int. Conf. Embedded Syst. Energy-Efficient Built Environ.*, Nov. 2015, pp. 55–64.

[23] C. Zhang, M. Zhong, Z. Wang, N. Goddard, and C. Sutton, "Sequence-to-point learning with neural networks for nonintrusive load monitoring," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI)*, New Orleans, LA, USA, Feb. 2018, pp. 1–8.

[24] J. Kim, T.-T.-H. Le, and H. Kim, "Nonintrusive load monitoring based on advanced deep learning and novel signature," *Comput. Intell. Neurosci.*, vol. 2017, Oct. 2017, Art. no. 4216281.

[25] N. Henao, K. Agbossou, S. Kelouwani, Y. Dubé, and M. Fournier, "Approach in nonintrusive type I load monitoring using subtractive clustering," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 812–821, Mar. 2017.

[26] L. Henning and M. Bergés, "The neural energy decoder: Energy disaggregation by combining binary subcomponents," in *Proc. 3rd Int. NILM Workshop*, Mar. 2016, pp. 1–5.

[27] B. Zhao, L. Stankovic, and V. Stankovic, "On a training-less solution for non-intrusive appliance load monitoring using graph signal processing," *IEEE Access*, vol. 4, pp. 1784–1799, Apr. 2016.

[28] M. Zeifman and K. Roth, "Nonintrusive appliance load monitoring: Review and outlook," *IEEE Trans. Consum. Electron.*, vol. 57, no. 1, pp. 76–84, Feb. 2011.

[29] J. Liao, G. Elafoudi, L. Stankovic, and V. Stankovic, "Non-intrusive appliance load monitoring using low-resolution smart meter data," in *Proc. IEEE Int. Conf. Smart Grid Commun.*, Venice, Italy, Nov. 2014, pp. 535–540.

[30] G. Zhang, G. Wang, H. Farhangi, and A. Palizban, "Residential electric load disaggregation for low-frequency utility applications," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Denver, CO, USA, Jul. 2015, pp. 1–5.

[31] F. Paradiso, F. Paganelli, D. Giuli, and S. Capobianco, "Context-based energy disaggregation in smart homes," *Future Internet*, vol. 8, no. 1, p. 4, Jan. 2016.

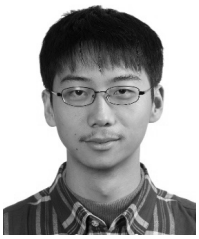
[32] D. Egarter and W. Elmenreich, "Autonomous load disaggregation approach based on active power measurements," in *Proc. IEEE Workshop Pervasive Energy Services*, St. Louis, MO, USA, Mar. 2015, pp. 293–298.

[33] M. Aiad and P. H. Lee, "Non-intrusive load disaggregation with adaptive estimations of devices main power effects and two-way interactions," *Energy Buildings*, vol. 130, pp. 131–139, Oct. 2016.

[34] M. Weiss, A. Helfenstein, F. Mattern, and T. Staake, "Leveraging smart meter data to recognize home appliances," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, Lugano, Switzerland, vol. 22, Mar. 2012, pp. 190–197.

[35] S. Chen, A. Sandryhaila, J. M. F. Moura, and J. Kovacevic, "Signal denoising on graphs via graph filtering," in *Proc. 2rd IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Atlanta, GA, USA, Dec. 2014, pp. 872–876.

[36] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.
[37] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1644–1656, Apr. 2013.
[38] A. Gadde, S. K. Narang, and A. Ortega, "Bilateral filter: Graph spectral interpretation and extensions," in *Proc. 20th IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2013, pp. 1222–1226.
[39] N. Batra *et al.*, "NILMTK: An open source toolkit for non-intrusive load monitoring," in *Proc. ACM 5th Int. Conf. Future Energy Syst.*, Jun. 2014, pp. 265–276.
[40] J. Z. Kolter and M. J. Johnson, "REDD: A public data set for energy disaggregation research," in *Proc. Workshop Data Mining Appl. Sustainab. (SIGKDD)*, San Diego, CA, USA, Jan. 2011, pp. 59–62.
[41] D. Murray, L. Stankovic, and V. Stankovic, "An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study," *Nature Sci. Data*, vol. 4, p. 160122, Jan. 2017.
[42] A. Cominola, M. Giuliani, D. Piga, A. Castelletti, and A. E. Rizzoli, "A hybrid signature-based iterative disaggregation algorithm for non-intrusive load monitoring," *Appl. Energy*, vol. 185, pp. 331–344, Jan. 2017.
[43] M. Dong, P. C. M. Meira, W. Xu, and W. Freitas, "An event window based load monitoring technique for smart meters," *IEEE Trans. Smart Grid*, vol. 3, no. 2, pp. 787–796, Jun. 2012.
[44] M. Figueiredo, B. Ribeiro, and A. de Almeida, "Electrical signal source separation via nonnegative tensor factorization using on site measurements in a smart home," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 2, pp. 364–373, Feb. 2014.
[45] C. C. Yang, C. S. Soh, and V. V. Yap, "A systematic approach to ON-OFF event detection and clustering analysis of non-intrusive appliance load monitoring," *Frontiers Energy*, vol. 9, pp. 231–237, Jun. 2015.
[46] S. Pattem, "Unsupervised disaggregation for non-intrusive load monitoring," in *Proc. 11th Int. Conf. Mach. Learn. Appl. (ICMLA)*, vol. 2, Dec. 2012, pp. 515–520.
[47] M. J. Johnson and A. S. Willsky, "Bayesian nonparametric hidden semi-Markov models," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 673–701, 2013.
[48] K. Kumar, R. Sinha, M. G. Chandra, and N. K. Thokala, "Data-driven electrical load disaggregation using graph signal processing," in *Proc. IEEE Annu. India Conf. (INDICON)*, Dec. 2016, pp. 1–6.
[49] J. Pang, G. Cheung, A. Ortega, and O. C. Au, "Optimal graph laplacian regularization for natural image denoising," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 2294–2298.
[50] C. Yang, A. Kerr, V. Stankovic, L. Stankovic, P. Rowe, and S. Cheng, "Human upper limb motion analysis for post-stroke impairment assessment using video analytics," *IEEE Access*, vol. 4, pp. 650–659, Apr. 2016.
[51] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs: Frequency analysis," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3042–3054, Jun. 2014.



BOCHAO ZHAO received the B.Eng. degree (Hons.) from the University of Strathclyde, Glasgow, U.K., in 2014, where he is currently pursuing the Ph.D. degree in electronic and electrical engineering. His research interests include graph signal processing and energy disaggregation. His current focus is on disaggregation for electricity profiles (electricity usage per hour).

KANGHANG HE received the B.Eng. degree (Hons.) from the University of Strathclyde, Glasgow, U.K., in 2015, where he is currently pursuing the Ph.D. degree in electronic and electrical engineering. His research interests lie in graph signal processing applications includes power disaggregation, eye trackers, and electroencephalography.

LINA STANKOVIC (M'04–SM'12) received the B.Eng. degree (Hons.) in electronic communications engineering and the Ph.D. degree from Lancaster University in 1999 and 2003, respectively. She is currently a Senior Lecturer with the Department of Electronic and Electrical Engineering, University of Strathclyde. She has co-authored four book chapters and over 130 peer-reviewed research articles. Her main research areas lie in signal information processing and data science, with a focus on smart monitoring and meaningful information extraction from sensor network arrays, as applied to energy data for load modeling, profiling and disaggregation, biomedical data for motion assessment, and seismic data for event detection and classification. She is a member of the EPSRC Peer Review College. She is an Associate Editor to IEEE TSIPN, and previously an Area Editor for Elsevier *International Journal of Electronics and Communications*.

VLADIMIR STANKOVIC (M'03–SM'10) received the Dr.Ing. (Ph.D.) degree from the University of Leipzig, Leipzig, Germany, in 2003. From 2003 to 2006, he was with Texas A&M University, College Station, first as a Research Associate and then as a Research Assistant Professor. From 2006 to 2007, he was with Lancaster University. Since 2007, he has been with the Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, U.K., where he is currently a Reader. He has co-authored four book chapters and over 160 peer-reviewed research papers.

His research interests include graph signal processing, data science, energy disaggregation, signal information processing for smart homes and buildings, and image analysis. He was an IET Technical and Professional Network in Vision and Imaging Executive Team Member and a member of the IEEE Communications Review Board. He was a Symposium Chair of the IEEE SmartGridComm 2013, a Technical Program Committee Co-Chair of Eusipco 2012, a General Chair of the IEEE Multimedia Signal Processing Workshop 2017, and an Area Chair of the IEEE International Conference on Image Processing 2018 and the IEEE International Conference on Visual Communications and Image Processing 2018. He was an Associate Editor of the IEEE COMMUNICATIONS LETTERS and the IEEE TRANSACTIONS ON IMAGE PROCESSING. He is currently an Editor-at-Large of the IEEE TRANSACTIONS ON COMMUNICATIONS and an Area Editor of Elsevier *Signal Processing: Image Communication*.