2014 IEEE International Conference on Healthcare Informatics

Towards Benchmarked Sleep Detection with Inertial Wrist-worn Sensing Units

Marko Borazio, Eugen Berlin, Nagihan Kücükyildiz, Philipp Scholl and Kristof Van Laerhoven
Embedded Sensing Systems
Technische Universität Darmstadt
Email: {borazio, berlin, nagi, scholl, kristof}@ess.tu-darmstadt.de

***Abstract*—The monitoring of sleep by quantifying sleeping time and quality is pivotal in many preventive health care scenarios. A substantial amount of wearable sensing products have been introduced to the market for just this reason, detecting whether the user is either sleeping or awake. Assessing these devices for their accuracy in estimating sleep is a daunting task, as their hardware design tends to be different and many are closed-source systems that have not been clinically tested. In this paper, we present a challenging benchmark dataset from an open source wrist-worn data logger that contains relatively high-frequent (100Hz) 3D inertial data from 42 sleep lab patients, along with their data from clinical polysomnography. We analyse this dataset with two traditional approaches for detecting sleep and wake states and propose a new algorithm specifically for 3D acceleration data, which operates on a principle of Estimation of Stationary Sleep-segments (ESS). Results show that all three methods generally over-estimate for sleep, with our method performing slightly better (almost 79% overall median accuracy) than the traditional activity count-based methods.**

I. Introduction

Sleep is an essential part of our life – almost one third of it we spend sleeping – and has been identified to be crucial to our health for a variety of reasons [1]–[3]. Sleep deprivation is known to lead to stress, a disturbed circadian rhythm, weight loss and, eventually, to death. The importance of finding out more about the way we sleep and if our sleep is sufficient is thus not limited to traditional disciplines such as somnology, neurology or psychiatry: providing a better picture on how well we sleep is relevant to all. Many off-the-shelf commercial devices can be bought for this purpose in a wristband form factor, from relatively compact devices such as the fitbit One[1], the Nike+ FuelBand[2] or the Jawbone UP[3] that are primarily aiming at fitness and activity tracking, to clinically evaluated devices such as the Actiwatch (Cambridge NeuroTechnology, Cambridge, UK) [4]. Some evaluations of commercial devices have shown that sleep information obtained from many such devices, such as total sleep time (TST), is not sufficiently accurate for sleep disorder assessment, (e.g., [5], which compares the fitbit to a commonly used actigraph for sleep evaluation). Such devices might be a benefit for private use, but have been found to overestimate sleep by a large margin [6].

Actigraphy devices are mainly used in the diagnosis of sleeping disorders like sleep apnea, but not necessarily deployed by every sleeping lab, since these devices tend to be expensive to acquire, to maintain, and to replace. The golden standard in medicine to observe sleep remains polysomnography [7], where the patient has to spent at least one night in a sleeping lab while being monitored through typically more than 20 different sensors. However, such an environment is often uncomfortable to sleep in, as these sensors often need to be wired to the side of the bed and is very different from what the patient is used to at home. Additionally, this method is expensive, time-consuming, and the ”first-night-effect” [8], i.e., a bad perception of sleep due to a novel environment, is inevitable. Therefore, alternative solutions, e.g., accelerometer-based wrist-worn devices that might provide additional long-term information on top of polysomnography, are being pursued and investigated as a complementary instrument.

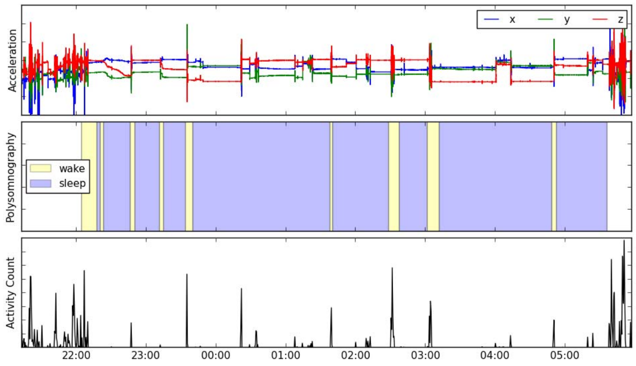

Fig. 1. An illustration of timeseries data from a wrist-worn 3D accelerometer (top) and polysomnography, suggesting that there is strong correlation between sleep-wake phases (middle) and amount of activity (bottom). This paper presents a benchmarking dataset to evaluate and reproduce results for such algorithmic approaches to detect sleep and wake phases from accelerometer data, and proposes a novel algorithm that is compared with 2 traditional ones.

Accelerated developments in the use of inertial sensors in cars and personal computing devices has led to the introduction of many commercial inertial loggers that can be worn around the wrist for several weeks at a time, and which are used to monitor both sleep and physical activity of the wearer. Most of these products are intended to be used in preventive health care scenarios by the users themselves to track and quantize their lifestyle. Verification of these commercial devices for clinical trials is rarely a priority. These devices generally estimate the times when the user was asleep, and several also contain models of sleeping cycles and individual stages (such as Rapid Eye Movement - REM - and Non-REM). These models and their different hardware solutions are mostly closed-sourced,

[1] http://www.fitbit.com, last access 06/2014
[2] http://nikeplus.nike.com, last access 06/2014
[3] http://www.jawbone.com, last access 06/2014

978-1-4799-5701-9/14 $31.00 © 2014 IEEE
DOI 10.1109/ICHI.2014.24

which makes the validation of the used algorithms challenging. Additionally, benchmark datasets with raw acceleration data from a wrist worn device with ground truth as polysomnography outputs, so that detection algorithms can be pitted against each other and results can be reproduced, are not publicly available. One exception is a dataset that includes inertial data [9], though the authors focused in this paper on a device that was worn around the waist, which is a common procedure for detecting the body posture outside a sleeping lab.

We argue in this paper that wearable wrist-worn activity loggers can be deployed as an additional instrument to complement the traditional polysomnography observation method, on the premise that the internal algorithms that estimate sleep and sleep stages are verified to work on benchmark data. For this purpose, we have *gathered a dataset* from a variety of 42 patients with sleeping disorders that have worn an inertial data logger at the wrist while also being observed via polysomnography to obtain the actual sleep ”ground truth”. The dataset contains the raw inertial data from a common MEMS (MicroElectroMechanical System) accelerometer at a rate of 100Hz and a sensitivity of $\pm 4g$ to enable generic testing of estimation algorithms. Comparing the traditional algorithms to detect sleep and wake cycles (as for example being used by the Actiwatch - see study in [10] - and the Mini-Motionlogger - see study in [11]), we present *a novel algorithm for sleep-wake phase detection*, showing that a 3D accelerometer-based device can yield detection accuracy of 74% by explicitly detecting segments of idleness, as opposed to being based on detected activity counts as is prevalent in related work.

The remainder of this paper is structured as follows: In Section II, we present relevant research to our work, while pointing out how it contributes to the field of sleep studies. Then, in Section III, we present implementation details of our adaptations of two well-known algorithms that will be compared with our method for sleep-wake detection. Section IV describes our first contribution in this paper: a benchmark dataset that has been recorded with 42 patients, suffering from a variety of sleeping disorders, to evaluate sleep-wake detection methods. Following the dataset description, we present the details and results from our study in Section V, discussing our findings in detail in Section VI. After the discussion and an outlook on this work, we conclude this work in Section VII.

II. Related Work

Several research groups investigated the use of actigraphy for sleep disorder assessment [12]–[14]. The results indicate that actigraphs can be used in addition to polysomnography, especially if it is important to monitor the patient in his or her usual environment, over longer stretches of time, or in paediatric treatment. Actigraphs can give insights into different sleeping disorders, such as sleep-wake disorders, sleep-schedule disorder, periodic limb movement (PLB), narcolepsy and sleep apnea. On the other hand, actigraphs are known to be less accurate in detecting wake segments during sleep and for sleeping disorders that exhibit vast amounts of such motionless periods such as in insomnia [15]. Generally, actigraphs return data that consist of activity counts, a measurement that has not been standardized across devices, and interpreted by each actigraphy manufacturer individually, making it challenging to compare the different algorithmic approaches with each other.

The calculation of the activity count can vary substantially, depending on the device that is being used. Several research efforts, e.g., [16], have performed comparison studies to map raw accelerometer data to activity counts, such as those from the Actiwatch 7, to show that it is possible to use a 3D accelerometer to calculate the activity counts. This has been evaluated on sleeping data that have been recorded with an accelerometer and the Actiwatch in parallel, and is based on the findings of [17], where accelerometer data during the day were matched to an actigraph output. More recent work [18] took a similar approach to derive the activity count solely from inertial data to be able to use traditional sleep detection and sleep parameter algorithms. The algorithms were evaluated by data obtained from 15 healthy subjects in their home environment, showing high agreement rates between epochs (i.e., observed time intervals in sleep research) for an actigraph and a MEMS.

Based on this activity count measure, two validated algorithms have been introduced in previous research that calculate sleep parameters as well as sleep-wake cycles in actigraphs: (1) Oakley’s from 1997 [19] is used for the Actiwatch, and (2) Cole et al. from 1992 [20] is the basic approach for the Mini-Motionlogger actigraph. For both algorithms, the sleep-wake cycle is calculated offline, requiring the data to being downloaded after recording. Many sleep studies make use of these former mentioned devices, detailing how accurate these devices can detect sleep for a large variety of disorders [10], [11], [21], [22]. These algorithms also form the basis for many novel devices that are equipped with a 3D MEMS accelerometer, as opposed to the traditional actigraphs that contain an omni-directional accelerometer.

Cole et al. in their approach make use of the zero-crossing technique [22] to calculate first the activity counts for a specified epoch, i.e., a time interval in which activity counts are being calculated. The activity counts per epoch are used to determine the total activity count D by considering a 7 minute window according to the following equation:

$$\begin{aligned} D = P * (&A_{-4}W_{-4} + A_{-3}W_{-3} + A_{-2}W_{-2} + \\ &A_{-1}W_{-1} + A_0W_0 + A_{+1}W_{+1} + A_{+2}W_{+2}) \end{aligned} \quad (1)$$

which essentially detects sleep whenever $D < 1$. In this formula, P is the scaling factor and W the weighting for each activity count, calculating the weighted sum over the epochs 4 minutes prior (A_{-x}) and 2 minutes after (A_{+x}) the current epoch (A_0). The common parameters according to [23] are: $P = 0.0033$, $W_{-4} = 1.06$, $W_{-3} = 0.54$, $W_{-2} = 0.58$, $W_{-1} = 0.76$, $W_0 = 2.3$, $W_{+1} = 0.74$ and $W_{+2} = 0.67$.

Oakley presented a similar approach in his paper [19] to detect sleep and wake phases, making use of amplitude-based activity counts. The algorithm examines the epochs that are in the 2 minutes before and the 2 minutes after a scored epoch:

$$A = \frac{1}{25}a_{-2} + \frac{1}{5}a_{-1} + 2a_0 + \frac{1}{5}a_{+1} + \frac{1}{25}a_{+2} \quad (2)$$

where A is the total activity count for the scored epoch a_0, a_{-x} is the activity count before the scored epoch and a_{+x} the one after, with $x \in [1, 2]$ minutes. Each surrounding epoch is multiplied by a weighting factor ($\frac{1}{25}$ and $\frac{1}{5}$). The sensitivity threshold for A can be set to high (80), medium (40) or low

(20) sensitivity, detecting sleep whenever $A <$ threshold. Low and medium sensitivity thresholds correlate with a high degree to sleep estimated by polysomnography [4].

We will focus in this paper solely on the algorithms and their ability to accurately detect sleep and wake phases, from *any source* of inertial sensor data. Since the two aforementioned algorithms were specifically designed for use on signals from particular omni-directional accelerometer sensors, we first needed to adapt these for raw 3D MEMS acceleration sensors. The next section will provide the details on how these algorithms were re-implemented, based on comparison studies by other researchers, for detecting sleep and wake phases in raw accelerometer data such as those from the open-source inertial data loggers deployed in this paper.

III. Algorithm Details for the Comparison Study

The data processing chain for all algorithms considered in this paper's comparative study can be described in four distinct steps: (1) Obtaining the raw 3D accelerometer data, (2) band-pass filtering the data, (3) calculating the algorithm-specific features per epoch and (4) applying the sleep detection algorithms on this feature set. Figure 2 depicts plots from this process up to the feature extraction for the algorithms by Cole et al. and Oakley. Before any of the algorithms are being used, we filter the raw acceleration data to remove any noise present in the raw data, as well as any high-frequency motion artefacts, as argued for in [24]. For this purpose we use a Butterworth bandpass filter with different low- and high-cut frequencies that have been experimentally verified for sleep and wake phase detection in [16]. The activity counts for both methods have been implemented as follows:

Oakley. In order to be able to evaluate against the algorithm by Oakley, we use Virkkala's approach presented in [16] to estimate activity counts for Oakley's algorithm equivalent to the Actiwatch output. The activity count is estimated by using the z-axis[4] only to determine the maximum absolute value inside 1-second windows. These per-second values are accumulated over the observed epoch length and scaled by two parameters, x and y, accordingly: $A = x * G + y$, where A = total activity count equivalent to the Actiwatch activity count, G = activity count over the epoch length derived from inertial data, x = 66 and y = -3.3. The scaling factors have in the experiments of [16] been estimated for the commonly-used epoch length of 30 seconds for a wrist-worn, 3D accelerometer-based device, which is why we will use this same epoch length for our comparison study to calculate the activity count. The use of a different epoch length will require the reassessment of the scaling factors.

Cole et al. For the algorithm by Cole et al., we have implemented a windowed zero-crossing count on the inertial data to obtain the activity count as it is used in equation (1). Researchers in [25] detailed that the zero-crossing for this purpose on the accelerometer's z axis was conducted to obtain the activity count. We have replicated this approach here with these parameters, counting the zero-crossings on the filtered data for every 1-second interval.

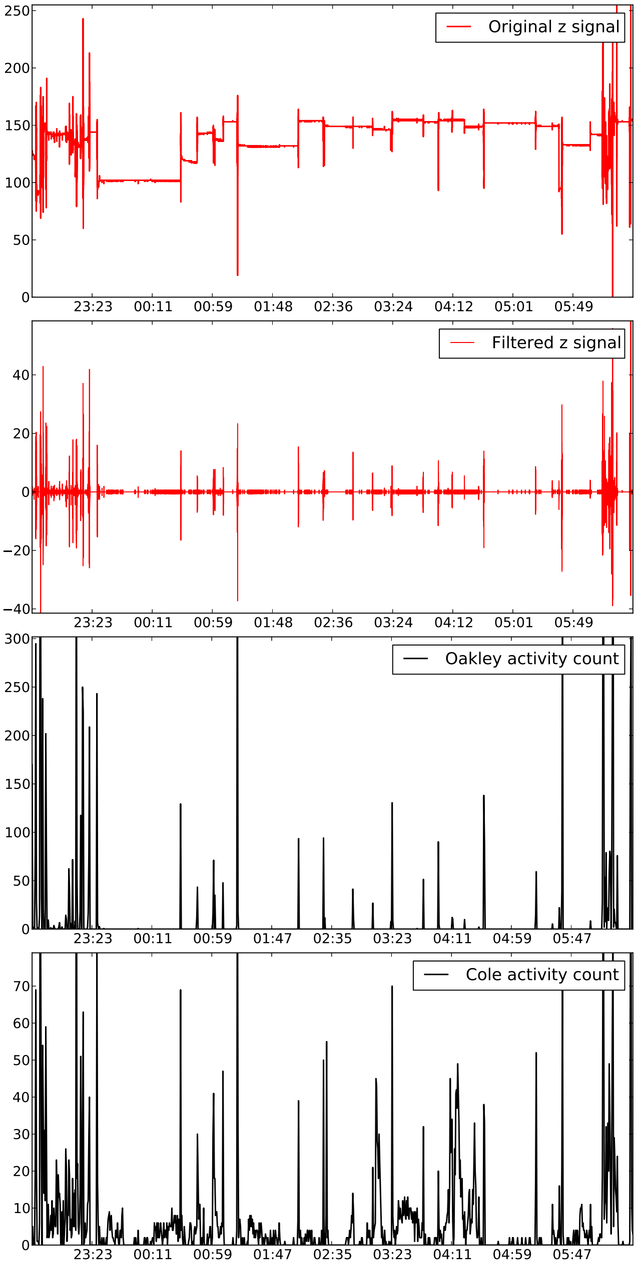

Fig. 2. Timeseries of the data abstraction steps performed in this study, to compare methods that detect sleep-wake phases based on 3D acceleration. The raw accelerometer data (top) is first treated with a band-pass filter (middle, in red), after which method-specific features, called activity counts, are computed per epoch (bottom, in black) to detect the sleep and wake phases.

The activity counts for both algorithms have been accumulated over epochs of 30 seconds, enabling the use of these two algorithms with the exact same parameters as introduced in previous work. It is important to note that the results from these re-implementations might still slightly deviate from the algorithms' designs in the way they are embedded in the Actiwatch and Mini-Motionlogger devices, since they operate on essentially different sensor modalities. However, the two independent studies that our implementations are based upon ([16] and [17]) report encouraging approximation results between the respective activity count methods (embedded in

[4] This is taken to be the axis that is perpendicular to the hand palm.

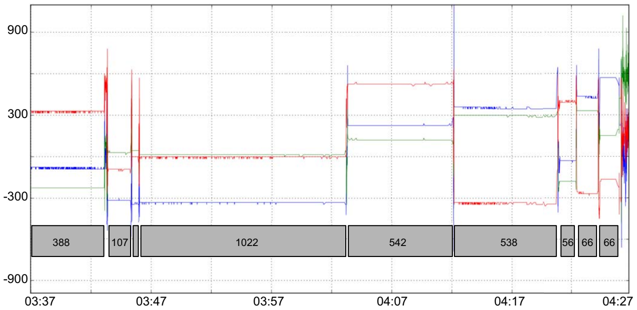

Fig. 3. Our approach focuses on the detection of sustained periods of idleness during sleep, in which the 3D acceleration signals remain flat. We use these segments and their duration (grey bars at the bottom of the plot, duration in seconds) as a basis for sleep detection, as opposed to sliding a fixed-width window over the data as in traditional activity count-based analysis methods.

hardware) and their 3D MEMS accelerometer-based reproduced variants, indicating that differences can be expected to be small.

The ESS approach. We present and evaluate in this paper a third alternative method, called ESS (Estimation of Stationary Sleep-segments), that is inherently different from the previous two methods since it does not rely on activity counts (whether produced by amplitude or zero crossings) over pre-defined time spans. Instead, it relies on the presence of long periods of idleness that in 3D acceleration data manifest themselves as flat horizontal signals. These are typically interchanged now and then with short transitions where the patient changed her sleeping posture. These segments are then used similarly to the epochs in the previous two methods, along with their duration (in seconds) as weights. Figure 3 illustrates this concept on typical sleep data from a 3D acceleration sensor (in mg) recorded at 100Hz over a time span of 50 minutes before awakening (at 4:27am): The intervals between motion segments are typically quite long during normal sleep. The detection of these segments is based on the following method:

$$S_\delta = \begin{cases} 1, & \text{if } \sqrt{\frac{1}{99}\sum_{i=1}^{100}(z_i - \overline{z})^2} > \delta, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Our approach consists essentially out of two steps: The first one applies a strong low-pass filter to the data to identify the segments in which there are no movement patterns present in the accelerometer data, as detailed in the formula above. Similar to the implementation for Oakley and Cole et al., we use solely the z-axis readings to which the filter is applied to. This is achieved by a sliding window approach in which the standard deviation (STD) is calculated over a 1-second interval, after which the resulting value is thresholded for δ. The second step then performs the identification of entire segments and collects these segments' start and stop times along with their length in seconds in a lookup table for later reference. A second threshold parameter is required here to select the minimal length (in seconds) for such intervals in which the accelerometer values remain unchanged. The source code (written in Python) for all three algorithm implementations as used in the remainder of this paper is available for download at: http://www.ess.tu-darmstadt.de/ichi2014.

The following section will focus on our experimental setup, specifically the wearable logging platform that was used in collecting our benchmark dataset, the methodology of the data collection process, as well as further details on the type of patients who participated in this study.

IV. Experimental Setup

For a comparative study between the two methods that were described in the previous section and our proposed sleep-wake phase detection method, we collected 3D acceleration data from 42 patients spending a night in a sleeping lab, while being supervised by somnologists and being monitored with polysomnography. The latter can be used as so-called "ground truth" for the patients' actual sleep-wake phases, as well as provide more details on the individual sleep phases (such as REM vs. Non-REM). The 3D acceleration logging platform with which these data were recorded is a wrist-worn device, for which both the hardware design as well as the firmware are publicly available.

A. Wrist-Worn 3D Accelerometer Logger

Gathering data over a long time span with a high frequency rate is a challenge, since it requires a device that (1) is sufficiently small and light so that it can be worn comfortably, (2) can store data internally over potentially longer periods for later (offline) evaluation and (3) from which the raw sensor data can be extracted, without preprocessing. It is hard to find such devices that meet all requirements, especially since most of these devices are closed-source and give only the possibility of looking at preprocessed data via the company's software or by uploading data directly to their web server. Therefore, we decided to rely on a custom-built prototype: the wrist-worn data logger measures 3D acceleration samples with a default sensitivity range of $\pm 4g$ that are sampled at 100Hz, together with the readings from an ambient light sensor (that could be used later on in detection of dark environments). The on-board accelerometer, the ADXL345 from Analog Devices, can be reconfigured from sensitivity ranges from $\pm 2g$ up to $\pm 16g$ and supports sampling rates of several thousands of samples per second. The whole unit fits in a plastic enclosure that protects the module and is small enough to be worn comfortably on the wrist and is attached with an elastic strap to it.

These sensor data are compressed and stored directly on the embedded SD card for later retrieval via a USB port. As soon as the logging device is thus attached to a host computer, its data can be downloaded and the small battery can be recharged. The 180mAh Li-Polymer battery lasts approximately two weeks while continually logging at 100Hz, which can be extended significantly by lowering the sampling rate (though for this paper's experiment purposes, we only needed to log data for maximally two days). The designs, both hardware and software, of this particular logger are open-source and available via http://www.ess.tu-darmstadt.de/hedgehog to support reproduction of our experiments.

B. Study Participants

We gathered data from 42 sleeping lab patients aged between 28 and 86 years, suffering from a variety of sleeping disorders (though most were later diagnosed with primarily

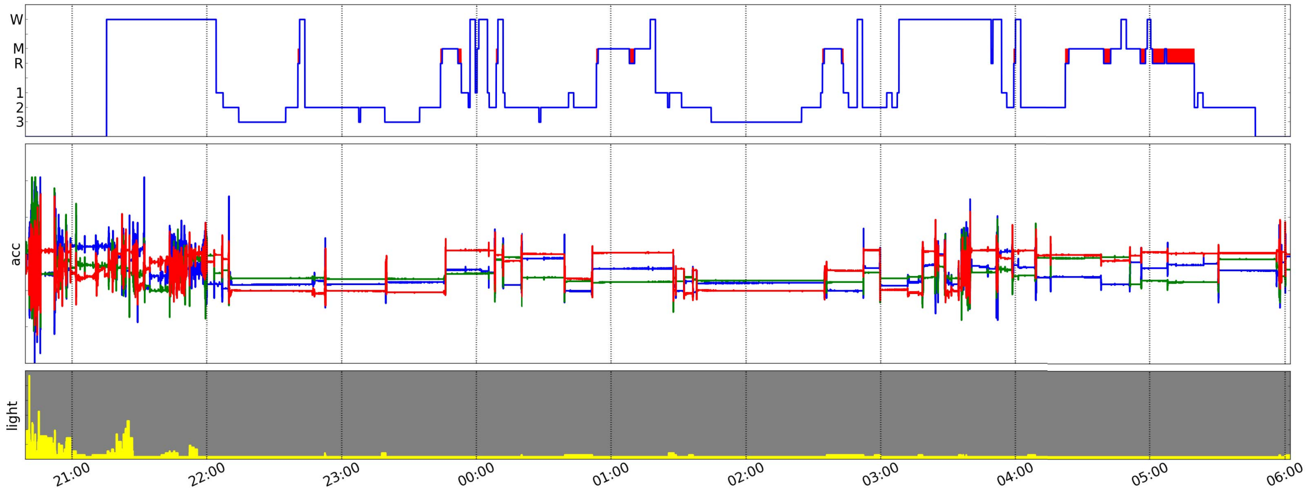

Fig. 4. TOP: Sleep phases from a 24 year old female subject, showing awake (W), movement (M), REM (R, red rectangles in the plot) and the Non-REM phases (1, 2, 3). MIDDLE: Inertial data from the sensor worn at the dominant wrist. BOTTOM: Light sensor values for the entire recording period.

sleep apnea syndrom (SAS), restless leg syndrome (RLS), or narcolepsy). In total, we recorded 45 nights' worth of data, whereby three patients wore the sensor for two nights in a row, attesting for over 409 hours of 100Hz acceleration samples, annotated with ambient light readings (which are not used in this paper's comparative study, but have been included in the benchmark dataset for incorporation in later algorithms) and the polysomnography details. Table I summarizes the main details on the patients that participated, indicating also how much data we obtained from the wrist sensor.

TABLE I. SOME KEY PROPERTIES OF THE COLLECTED BENCHMARK DATA (SAS = SLEEP APNEA SYNDROM, RLS = RESTLESS LEG SYNDROM).

number of nights:	45
number of patients:	42
gender distribution:	22 male, 20 female
age distribution:	24 - 86
disorders (diagnosed):	Insomnia, narcolepsy, SAS, RLS
data (in minutes):	24475

The patients were recruited by staff at the sleeping lab and monitored at least over one night via the standard polysomnography method, as well as with the wrist sensor. After a short introduction on how the wrist sensor works and what type of data it captures, each patient was asked to start wearing the sensor unit at least one hour before going to sleep and to take it off one hour after waking up. The patients signed a privacy policy and a consent form, allowing the scientific use of the obtained anonymized data from both the polysomnography as well as from the wrist sensor. They were furthermore given documents that describe the experiment in detail and stipulate how the data will be anonymized afterwards in order to enable sharing of the dataset with other researchers for future studies.

In general, the acceptance of wearing the wrist-worn device in addition to the polysomnography set-up was high: Many patients expressed interest in future studies of the device and responded positively to the idea of having such devices complement polysomnography for recording in their usual home environment.

C. Data Collection Method

Data obtained in this study consists of over 409 hours of inertial data and polysomnography data. The sensor was instructed to be worn on the dominant wrist, although previous studies have shown that the wrist placement is not crucial in sleep studies [26]. Before the sensor distribution, the real-time clock embedded on the accelerometer-based device was configured to be aligned to the clock of the polysomnography system in the sleeping lab, in order to obtain a synchronized dataset. On return of the wrist-worn sensors, their data logs were visualized to the patients as part of the privacy policy (see Figure 4 for an example of such a visualization).

The patients' polysomnography data was scored in 30 second epochs by standard procedure obtained after a few days of monitoring, since the medical staff had to analyze the data first, after which the doctor had to summarize these diagnostic findings. The data consists of the patients' demographic information, the wake and sleep stages, with sleep being displayed by REM and Non-REM (sleep phases 1-3) and sleep characteristics, e.g., total sleep time (TST).

The sleep phases are divided into three different stages, labelled '1', '2' and '3'. Additionally, periods of particularly high amounts of limb movement are marked in the dataset ("M"), as well as when the polysomnography detected a wake phase ("W"). Note that traditionally, a fourth sleep stage '4' is sparsely present in the dataset's polysomnography section, although this sleep stage is not actively being used since 2012 in the sleeping labs we have collaborated with. The dataset we obtained does not contain this sleep stage.

The following section will present the evaluation results on the recorded data, showing the performance of each of the two activity count-based algorithms, as well as our novel method presented in the previous section.

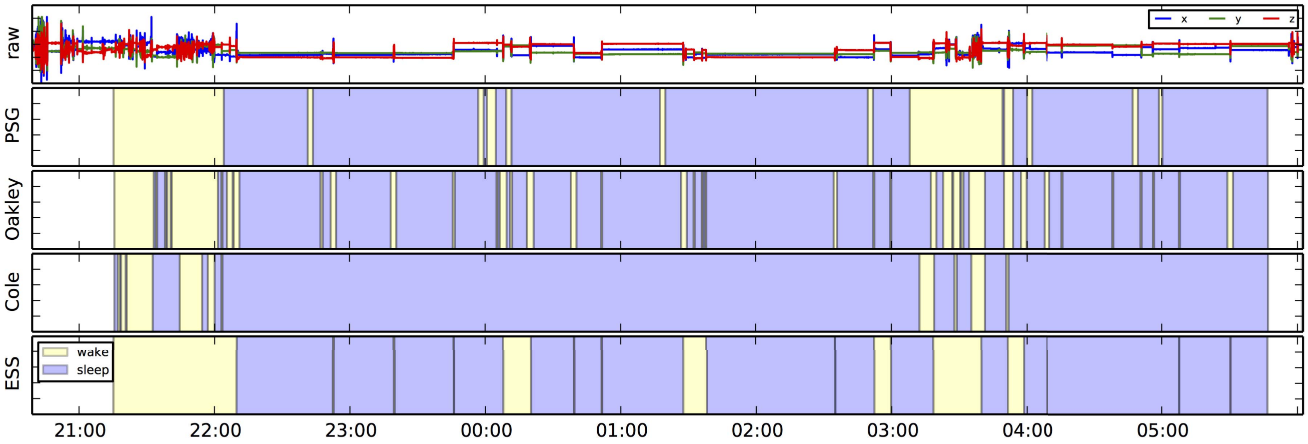

Fig. 5. Sleep-wake estimation results for a 24 year old female suffering from narcolepsy. Displayed are the evaluation of the activity count based algorithm (Oakley and Cole et al., middle plots) and the ESS algorithm (bottom plot) compared to the polysomnography (PSG) output. Additionally, we see the raw 3D inertial data of the wrist-worn sensor (top plot). Precision and recall are similar for all three algorithms (87%-99.9%).

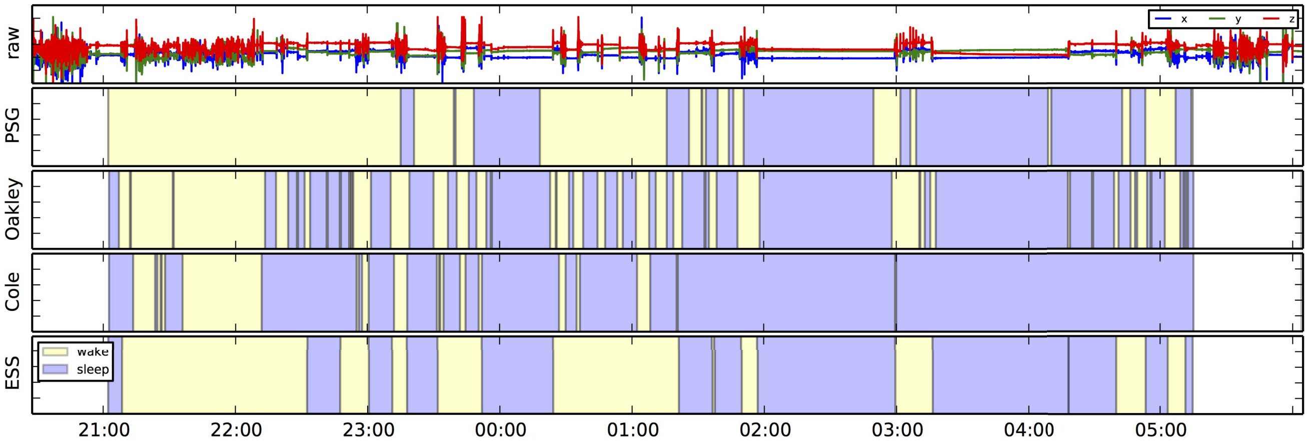

Fig. 6. Sleep-wake estimation results for a 69 year old male suffering from SAS. Displayed are the evaluation of the activity count based algorithm (Oakley and Cole et al., middle plots) and the ESS algorithm (bottom plot) compared to the PSG output. Additionally, we see the raw 3D inertial data of the wrist-worn sensor (top plot). Especially in the beginning of the recording is a sleeping segment detected, which was due to immobility of the patient while starting the PSG. ESS shows clearly detected wake segments, outperforming the other two algorithms by 10%-20% in accuracy.

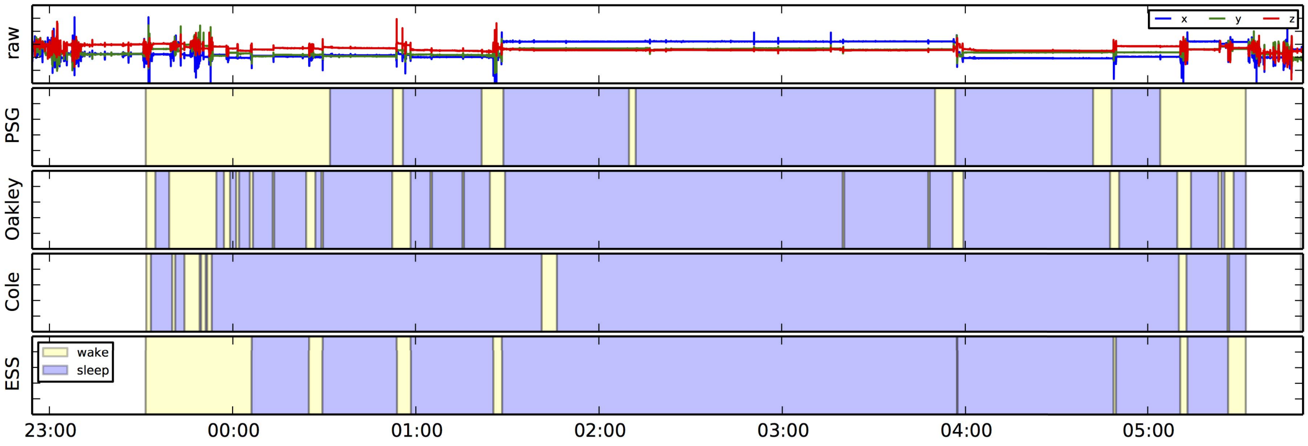

Fig. 7. Sleep-wake estimation results for a 72 year old male patient suffering from SAS. Displayed are the evaluation of the activity count based algorithm (Oakley and Cole et al., middle plots) and the ESS algorithm (bottom plot) compared to the PSG output. Additionally, we see the raw 3D inertial data of the wrist-worn sensor (top plot). ESS clearly detects the initial wake segment while Oakley and Cole et al. tend to detect short sleep intervals.

TABLE II. TOTAL SLEEP TIME (TST) AND SLEEP EFFICIENCY (SE) FOR ALL NIGHTS ACCORDING TO PSG, THE TWO ACTIVITY-COUNT BASED ALGORITHMS AND OUR APPROACH (ESS).

parameter	PSG	Oakley	Cole et al.	ESS
TST	287 min ±92	413 min ±42	406 min ±44	328 min ±88
SE	66% ±20%	94% ±3%	93% ±4%	76% ±18%

V. EVALUATION

In order to find the optimum minimum length for the intervals of non-movement, we defined six different interval thresholds (300, 360, 480, 600, 720 and 900 seconds) and evaluated their performance in regard to sleep detection, using the PSG dataset as ground truth. We take the accuracy for detecting sleep and wake phases into regard, and use the precision and recall to investigate further differences between the individual parameters' performances. We observe that the mean accuracy is best for the 600 interval which is why we choose this interval to determine immobile segments. Additionally, we set the standard deviation (STD) threshold to 6, as derived by experimental evaluation of different thresholds.

For each of the three sleep estimation algorithms we compare the results to the PSG output. Figure 5 shows the visual results for all three algorithms together with the raw data and the PSG estimation for sleep and wake (blue and yellow respectively). For Oakley's algorithm we use the most sensitive threshold of 20 to mark the epoch as sleep. Just by visual inspection we see that Cole et al. overestimates sleep and fails to detect small wake segments. Oakley exhibits many, mostly short wake segments within sleep intervals but detects most of the sleep. Interestingly, ESS detects the initial wake segment which is almost identical to the PSG segment, while sleep is being detected accurately. Quantitative results confirm the observations: accuracies vary for all three in the range of 82% - 85%. More visual results are shown in Figures 6 and 7, indicating a better performance for the ESS algorithm in contrast to Oakley and Cole et al.

Additionally to the visual inspection, we investigated some sleep parameters to complete the dataset's description. We calculate total sleep time (TST) and sleep efficiency (SE) for each algorithm and compare these values. Total sleep time is the amount of sleep in minutes being detected by the algorithm. Sleep efficiency is the quotient of TST and total recording time (here: PSG start and PSG end). Table II shows the results for these parameters, indicating a very low mean value for TST as determined by PSG. In comparison to that, Oakley and Cole et al. tend to overestimate TST, while ESS represents a TST value between PSG and Oakley. Here, the inevitable problem can be observed: activity count based systems tend to overestimate sleep in general, as depicted in [5]. The same is shown here in SE: Oakley and Cole et al. exhibit high values of 94% and 93% respectively, while ESS is very close to the PSG SE. We can state here that all three algorithms differ from PSG, ESS less than Oakley and Cole et al.

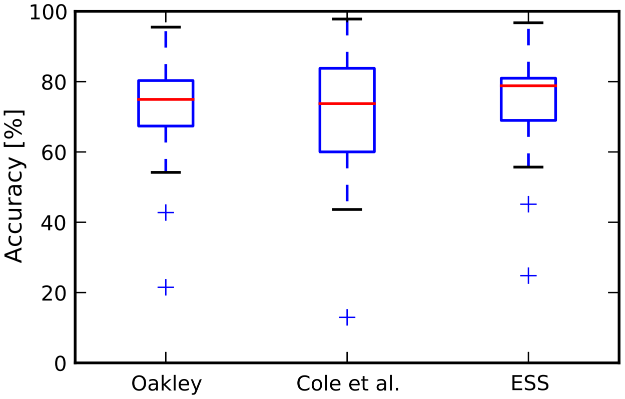

Fig. 8. Accuracy results for the ESS (median: 78.83%) algorithm compared to Oakley (median: 74.94%) and Cole et al. (median: 73.75%). Both median (red line in the box) and interquartile results indicate that the proposed ESS outperforms the traditional actigraphy-based approaches, while all approaches still leave ample of room for improvement.

Accuracy results for all algorithms are shown in Figure 8 as boxplots. We observe a median (red line in the box) for ESS that is slightly higher than for Oakley and Cole et al. Median accuracy values for the three approaches ESS (78.83%), Oakley (74.94%) and Cole et al. (73.75%) are all close to each other. Overall, however, ESS' performance is slightly better over all participants in the study, which can be also seen by observing the interquartile borders. Note here, that for each algorithm an outlier is visible (lowest '+' for each boxplot): This is explained by a 69 year old female patient, who lay awake most of the time while being recorded, while her inertial data log exhibited almost no movements. For this strong outlier case, all three algorithms estimated sleep instead of wake.

Additionally, we show in Figure 9 precision and recall results for both sleep and wake segments (left plots: precision and recall for sleep, right plots: precision and recall for wake). Recall is the portion of sleep (or wake) segments that were correctly identified as sleep (or wake) during the classification. We observe for sleep that precision results are slightly better for ESS (median: 78.95%) and Oakley (median: 78.29%), whereas Cole et al. rest at 72.91%. We can highlight here that all three algorithms perform similarly in retrieving sleep segments from the given dataset. Cole et al. exhibits a high recall for sleep (median: 98.74%), which can be explained by the fact that Cole mostly detects sleep throughout the whole dataset, while Oakley (median: 92.93%) and ESS (median: 94.12%) highlight wake states more often. Interestingly, wake is being detected with a high variety in precision for all three algorithms, showing a higher recall for Oakley and ESS (both over 20% higher than Cole et al.).

The approach suggested by Cole et al., while detecting almost all the sleep intervals, fails to detect the relevant wake segments resulting in a much lower recall. The problem for detecting wake segments in this dataset in particular from a sleeping lab is visible in the results and is a challenge for most sleep-wake detection algorithms. It is also important to note here that the ESS algorithm keeps an adequate balance of detecting sleep and wake segments in the dataset, as opposed to Oakley and Cole et al., which tend to neglect wake segments as shown in the results of Figure 9.

We will now discuss our findings, highlighting both the benefits and limitations of using the dataset presented in this study, and give insights into further improvements for algorithms that analyse these data.

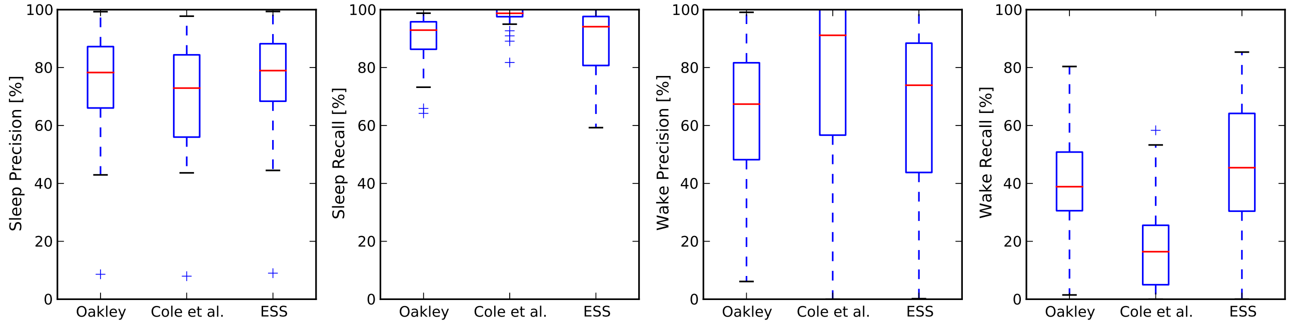

Fig. 9. Precision and recall results (left: sleep, right: wake) for ESS compared to Oakley and Cole et al. Sleep precision (leftmost plot) for Oakley (median: 78.29%) and ESS (median: 78.95%) are on the same level, while Cole et al. exhibits a better recall (median: 98.74%). For wake the results are inverted: Oakley (median: 38.86%) and ESS (median: 45.42%) show a higher recall, while Cole et al. yields a higher precision (median: 91.14%).

VI. Discussion

This work compares two commonly used sleep detection algorithms to the results of the ESS approach on sleep-wake detection with data from sleeping lab patients. The accuracies for detecting sleep and wake segments are very promising already, as shown in the previous section, yet we believe that parameters can be optimized to improve on the detection of such sleep and wake intervals. We will focus in this section on finding good candidates for these parameters, as well as what impact the dataset has for future studies.

A. Dataset

The dataset recorded for this study is a challenging one: First of all, most of the subjects observed suffer from a sleep disorder (diagnosed after their visit to the sleep lab), which makes it difficult to determine when the patient is really awake or just exhibiting spontaneous muscle contractions. This we observed for example for a 69 year old female subject, suffering from sleep apnea syndrome (SAS). According to PSG the patient was sleeping, but this sleep was interrupted by various incidents which let the sleep-wake algorithm detect wake segments even though the patient was sleeping. Second, the dataset includes also healthy patients (though a minority at 5/42 in total), which makes it a rich dataset not only on various sleeping disorders. All sleeping lab patients were diagnosed several days after their stay in the sleeping lab, we did not include healthy patients in the dataset on purpose. Additionally, three patients had to spend two consecutive nights in the sleeping lab, which produced particularly useful data as it should minimize the ”first-night-effect” drastically on the second night.

Our dataset with PSG data is enriched with acceleration data from a wrist-worn sensor and enables follow-up research to use data that is not only based on activity counts similar to actigraphs, but also on more fine-grained and relatively high-resolution (100 Hz) signals. Especially in the paediatric sleep research [27] such a set-up could lead to a more reliable detection of sleep-wake segments, since children have been observed to move more during sleep.

Some limitations of the dataset have to be considered here as well: (1) The benchmark dataset contains recordings that are mostly from patients spending one night in the sleeping lab, which is bound to have a drastic impact on the dataset. Normal sleep that approximates that of the patient’s home environment usually is achieved on the second night in the sleeping lab. According to the medical staff at the sleeping labs, the patients’ diagnoses could be obtained by spending only one night in the sleeping lab. A minor part of the dataset was obtained from patients who did spend two subsequent nights in the sleeping lab, but most of the data can be regarded as atypical and challenging for detecting sleeping patterns. Further evaluation needs to be conducted on how such an effect is influencing the overall results. (2) We assessed only few healthy patients, which is why the behavior of the ESS algorithm still has to be investigated under ”normal” circumstances, i.e., observing sleep in the home environment. This obviously becomes a challenge when ground truth data (PSG recordings) are needed to asses and evaluate the algorithm. The wrist-worn sensor presented in this study can be used for long-term studies without the need of recharging it for at least two weeks straight, depending also on the recording settings, especially the sensor module’s sampling frequency. Due to this focus on sleeping lab patients, the results on this dataset are therefore likely not representative for healthy subjects.

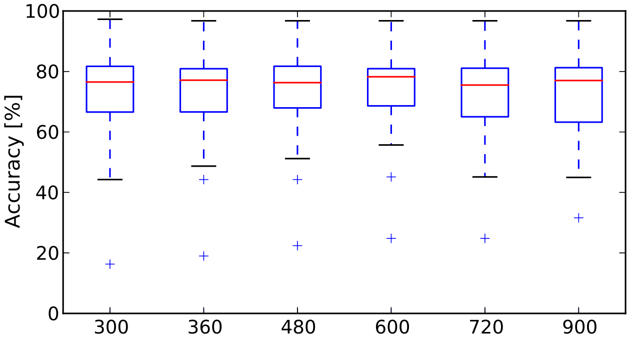

Fig. 10. Different idleness segment thresholds (x-axis, in seconds) compared among each other, indicate that 600 seconds (10 minutes) is the optimum interval length threshold for the ESS algorithm.

B. Parameters

As mentioned in the previous section, we determined an immobile segment length of 600 to mark the segment as sleep. This immobility threshold when varied yields different accuracy results on our dataset, as depicted in Figure 10 for the

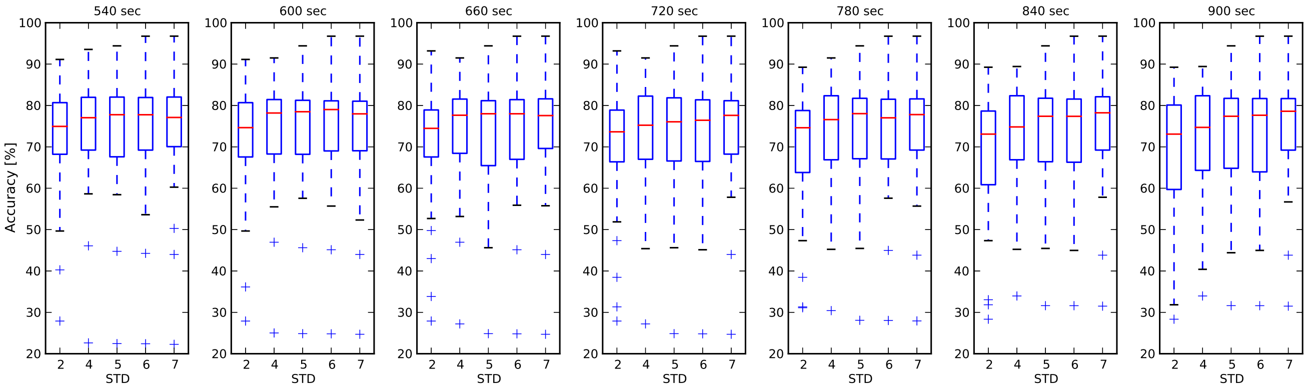

Fig. 11. Accuracy results for different STD thresholds (2, 4, 5, 6 and 7) for idleness interval length thresholds (540, 600, 660, 720, 780, 840 and 900 seconds). These two parameters were varied to achieve the optimum recognition rate for detecting both sleep and wake segments across all patients.

aforementioned thresholds (300, 360, 480, 600, 720 and 900). We observe an increase of the median until 600, after which it slightly drops again. Whether other thresholds can improve on the results has still to be investigated. For this purpose we have to take into consideration the STD threshold for detecting immobile signals within the inertial data. In Figure 11 we show the accuracy results for each immobile threshold length and the individual STD thresholds (2, 4, 5, 6 and 7). A too small STD leads to a lower accuracy, while the highest depicted here (7) stays steady over all immobile lengths. Not shown here is that for lower thresholds, we receive a very high precision for sleep, but a very low one for wake, which would contradict a system that should detect both sleep and wake segments. We can conclude that depending on the scenario, the thresholds can be varied and that for this study, the optimum thresholds of 6 for the STD and 600 for the idleness interval length return the most optimal results.

Activity count-based methods such as the ones by Oakley and Cole et al. take the surrounding epochs into consideration to smooth out the sleep detection over longer intervals of time. Such a step is not implemented in the ESS algorithm. One possibility to embed this is to determine all the stationary segments and detect smaller movement segments in-between these segments, which could be filtered out by setting a specific windowed threshold for movement (e.g., 2-3 seconds), that are marked as sleep. Nevertheless, these are considerations for future studies that need to be evaluated more thoroughly.

This paper's topic was limited to the identification of sleep and wake phases present in the 3D accelerometer data. Since the paper's dataset also contains more fine-grained sleep phase annotations, an interesting further line of investigation would be to take a deeper look into algorithms that not only determine sleep-wake intervals, but also estimate further phases such as REM and Non-REM, based on wrist-worn accelerometer data. As an initial investigation on how indicative the presence of activity in the data is for particular sleep phases, a histogram was constructed that shows the number of occurrences for each sleep phase per variance bin, as depicted in Figure 12 by a distribution of variances over 1 second for all sleep phases (SP1-3 and REM) including wake segments. As can be observed, REM (red) occurs only on the low ranges of variance, indicating that this phase exhibits low variances only (which is in support of what is know about limb motion during the REM phase). However, how well any algorithms could manage to estimate REM/Non-REM phases needs much more investigation on the dataset itself.

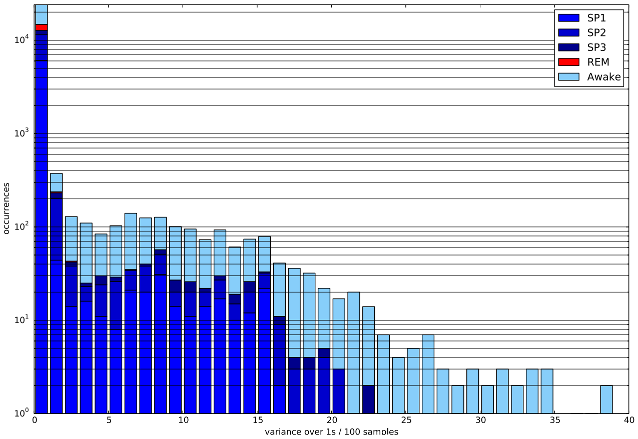

Fig. 12. Histogram of variance occurring in the different sleep phases REM and Non-REM (SP1-SP3) and wake phases.

VII. Conclusions

With the advent of 3D accelerometer MEMS chips that are both small and power-efficient enough to be included in wearable devices, long-term monitoring of sleep and wake phases has become an attractive and cost-effective instrument to complement traditional sleep lab studies using polysomnography. The systematic evaluation of algorithms that detect sleep and wake phases in such accelerometer data is still lacking, however, as current personal sleep devices and systems on the market are closed-source and not meant to be clinically deployed.

This paper contributes to such systematic evaluation of detection algorithms by presenting a challenging and publicly-available dataset[5] with over 409 hours worth of polysomnography-annotated 3D acceleration data at 100Hz for

[5] http://www.ess.tu-darmstadt.de/ichi2014

42 sleep lab patients, recorded with an open-source logging platform. We furthermore presented a novel method to detect sleep and wake phases for such platforms and compared this method with two traditional activity count-based methods on the benchmark dataset. Results show that the ESS algorithm achieves an overall median accuracy of almost 79% for detecting sleep and wake intervals. Compared to the other two methods of Oakley and Cole et al., relevant wake segments are detected with a higher confidence.

Future work that is currently ongoing includes improvement possibilities for the presented detection approach: performance could for instance be expected to improve with additional information, coming either from further on-board sensors (such as the ambient light sensor on the presented wrist-worn logging platform) or by including patient-specific models on sleeping disorders and personal routines, e.g., on usual sleep times.

The dataset and source code for the three evaluated algorithms is available at http://www.ess.tu-darmstadt.de/ichi2014 to support reproducing this paper's experiments and to facilitate investigation and evaluation of further methods.

Acknowledgments

We would like to thank the director Dr. med. Karl-Otto Steinmetz and the medical staff of the Elisabethenstift sleep laboratory (Darmstadt, Germany), who have enabled us to perform this study, as well as the 42 patients who participated in this study. We would also like to thank Jussi Virkkala for providing us with a test dataset as well as the original scripts of his study on activity counts [16]. This work was sponsored by the project **Long-Term Activity Recognition with Wearable Sensors** (LA 2758/1-1), which is funded by the German Research Foundation (DFG).

References

[1] S. Banks and D. F. Dinges, "Behavioral and Physiological Consequences of Sleep Restriction," *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*, vol. 3, no. 5, p. 519, 2007.

[2] J. M. Siegel, "Sleep Viewed as a State of Adaptive Inactivity," *Nature Reviews Neuroscience*, vol. 10, no. 10, pp. 747–753, 2009.

[3] H. P. Van Dongen, G. Maislin, J. M. Mullington, and D. F. Dinges, "The Cumulative Cost of Additional Wakefulness: Dose-response Effects on Neurobehavioral Functions and Sleep Physiology from Chronic Sleep Restriction and Total Sleep Deprivation," *Sleep*, vol. 26, no. 2, pp. 117–129, 2003.

[4] C. A. Kushida, A. Chang, C. Gadkary, C. Guilleminault, O. Carrillo, and W. C. Dement, "Comparison of Actigraphic, Polysomnographic and Subjective Assessment of Sleep Parameters in Sleep-disordered Patients." *Sleep Medicine*, vol. 2, no. 5, pp. 389–96, Sep. 2001.

[5] H. E. Montgomery-Downs, S. P. Insana, and J. a. Bond, "Movement Toward a Novel Activity Monitoring Device." *Sleep and Breathing*, vol. 16, no. 3, pp. 913–7, Sep. 2012.

[6] C. P. Pollak, W. W. Tryon, H. Nagaraja, and R. Dzwonczyk, "How Accurately Does Wrist Actigraphy Identify the States of Sleep and Wakefulness?" *Sleep*, vol. 24, no. 8, pp. 957–65, Dec 2001.

[7] C. A. Kushida, M. R. Littner, T. Morgenthaler, C. A. Alessi, D. Bailey, J. Coleman, L. Friedman, M. Hirshkowitz, S. Kapen, M. Kramer, T. Lee-Chiong, D. L. Loube, J. Owens, J. P. Pancer, and M. Wise, "Practice Parameters for the Indications for Polysomnography and Related Procedures: An Update for 2005." *Sleep*, vol. 28, no. 4, pp. 499–521, Apr. 2005.

[8] H. W. Agnew, W. B. Webb, and R. L. Williams, "The First Night Effect: An EEG Study of Sleep," *Psychophysiology*, vol. 2, no. 3, pp. 263–266, 1966.

[9] J. Tilmanne, J. Urbain, M. V. Kothare, A. V. Wouwer, and S. V. Kothare, "Algorithms for Sleep-wake Identification Using Actigraphy: A Comparative Study and New Results." *Journal of Sleep Research*, vol. 18, no. 1, pp. 85–98, Mar. 2009.

[10] A. R. Weiss, N. L. Johnson, N. a. Berger, and S. Redline, "Validity of Activity-based Devices to Estimate Sleep." *Journal of Clinical Sleep Medicine (JCSM), Official Publication of the American Academy of Sleep Medicine*, vol. 6, no. 4, pp. 336–42, Aug. 2010.

[11] L. de Souza, A. A. Benedito-Silva, M. N. Pires, D. Poyares, S. Tufik, and H. M. Calil, "Further Validation of Actigraphy for Sleep Studies," *Sleep - New York*, vol. 26, no. 1, pp. 81–85, 2003.

[12] T. Morgenthaler, C. Alessi, L. Friedman, J. Owens, V. Kapur, B. Boehlecke, T. Brown, A. Chesson, J. Coleman, T. Lee-chiong, J. Pancer, T. J. Swick, M. Clinic, V. A. Greater, L. A. Healthcare, and L. Angeles, "Practice Parameters for the Use of Actigraphy in the Assessment of Sleep and Sleep Disorders: An Update for 2007." *Sleep*, vol. 30, no. 4, pp. 519–29, Apr. 2007.

[13] A. Sadeh, P. Hauri, D. Kripke, and P. Lavie, "The Role of Actigraphy in the Evaluation of Sleep Disorders." *Sleep*, vol. 18, no. 4, pp. 288–302, 1995.

[14] A. Sadeh, "The Role and Validity of Actigraphy in Sleep Medicine: An Update." *Sleep Medicine Reviews*, vol. 15, no. 4, pp. 259–267, 2011.

[15] K. L. Lichstein, K. C. Stone, J. Donaldson, S. D. Nau, J. P. Soeffing, D. Murray, K. W. Lester, and R. N. Aguillard, "Actigraphy Validation with Insomnia," *Sleep - New York*, vol. 29, no. 2, p. 232, 2006.

[16] J. Virkkala, "Using Accelerometers as Actigraphs," *Poster in ESRS'12*, 2012.

[17] V. T. van Hees, M. Pias, S. Taherian, U. Ekelund, and S. Brage, "A Method to Compare new and Traditional Accelerometry Data in Physical Activity Monitoring," *2010 IEEE International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, pp. 1–6, Jun. 2010.

[18] B. H. W. te Lindert and E. J. W. Van Someren, "Sleep Estimates Using Microelectromechanical Systems (MEMS)." *Sleep*, vol. 36, no. 5, pp. 781–9, May 2013.

[19] N. Oakley, "Validation with Polysomnography of the Sleepwatch Sleep/Wake Scoring Algorithm used by the Actiwatch Activity Monitoring System," *Technical Report*, 1997.

[20] R. J. Cole, D. Kripke F, D. Mullaney, and C. Gillin, "Automatic Sleep/wake Identification from Wrist Activity," *Sleep*, vol. 15, no. 5, pp. 461–469, 1992.

[21] K. Benson, L. Friedman, A. Noda, D. Wicks, E. Wakabayashi, and J. Yesavage, "The Measurement of Sleep by Actigraphy: Direct Comparison of 2 Commercially Available Actigraphs in a Nonclinical Population." *Sleep*, vol. 27, no. 5, pp. 986–9, Aug. 2004.

[22] G. Jean-Louis, D. F. Kripke, W. J. Mason, J. A. Elliott, and S. D. Youngstedt, "Sleep Estimation from Wrist Movement Quantified by Different Actigraphic Modalities," *Journal of Neuroscience*, vol. 105, no. 2, pp. 185–191, 2001.

[23] M. Takeshima, M. Echizenya, Y. Inomata, K. Shimizu, and T. Shimizu, "Comparison of Sleep Estimation Using Wrist Actigraphy and Waist Actigraphy in Healthy Young Adults," *Sleep and Biological Rhythms*, vol. 12, no. 1, pp. 62–68, Jan. 2014.

[24] S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, W. Moorcroft, and C. P. Pollak, "The Role of Actigraphy in the Study of Sleep and Circadian Rhythms." *Sleep*, vol. 26, no. 3, pp. 342–92, May 2003.

[25] W. Karlen, C. Mattiussi, and D. Floreano, "Improving Actigraph Sleep/wake Classification with Cardio-respiratory Signals," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*. IEEE, 2008, pp. 5262–5265.

[26] A. Sadeh, K. M. Sharkey, and M. A. Carskadon, "Activity-Based SleepWake Identification: An Empirical Test of Methodological Issues," *Sleep*, vol. 17, no. 3, pp. 201–207, 1994.

[27] S. P. Insana, D. Gozal, and H. E. Montgomery-Downs, "Invalidity of One Actigraphy Brand for Identifying Sleep and Wake Among Infants." *Sleep Medicine*, vol. 11, no. 2, pp. 191–6, Feb. 2010.