Adaptive Sleep–Wake Discrimination for Wearable Devices

Walter Karlen*, *Member, IEEE*, and Dario Floreano, *Senior Member, IEEE*

***Abstract*—Sleep/wake classification systems that rely on physiological signals suffer from intersubject differences that make accurate classification with a single, subject-independent model difficult. To overcome the limitations of intersubject variability, we suggest a novel online adaptation technique that updates the sleep/wake classifier in real time. The objective of the present study was to evaluate the performance of a newly developed adaptive classification algorithm that was embedded on a wearable sleep/wake classification system called SleePic. The algorithm processed ECG and respiratory effort signals for the classification task and applied behavioral measurements (obtained from accelerometer and press-button data) for the automatic adaptation task. When trained as a subject-independent classifier algorithm, the SleePic device was only able to correctly classify 74.94 ± 6.76% of the human-rated sleep/wake data. By using the suggested automatic adaptation method, the mean classification accuracy could be significantly improved to 92.98 ± 3.19%. A subject-independent classifier based on activity data only showed a comparable accuracy of 90.44 ± 3.57%. We demonstrated that subject-independent models used for online sleep–wake classification can successfully be adapted to previously unseen subjects without the intervention of human experts or off-line calibration.**

***Index Terms*—Adaptation, context awareness, personal health, physiological signal classification, point-of-care, wearable.**

I. Introduction

MONITORING sleep–wake behavior of subjects at home allows the early detection of sleep disorders and is reducing health care costs [1]. Ambulatory health applications require comfortable devices that embed wearable sensors, electronics, and intelligent signal processing. The design of wearable sleep/wake discrimination systems is particularly challenging. The most common physiological signal used for sleep discrimination in clinical settings is the recording of brain activity with an EEG [2]. Unfortunately, EEG cannot be easily recorded with a wearable system and is subject to an increased level of noise. An alternative method is needed. It has also been shown that during sleep, intersubject differences in EEG [3] and cardiorespiratory signals [4], [5] are more pronounced than intrasubject variations. Consequently, any signal processing and classification algorithm tuned to a model user is bound to produce highly variable results in different persons. This suggests that on a mobile device an efficient user adaptation strategy is required.

A. Background

Sleep–wake behavior is normally monitored using polysomnographic analysis that includes the recording of EEG [6]. Polysomnography is usually conducted in sleep centers which requires the patient to stay overnight. More recently, portable recorders were used for ambulatory sleep recordings that allow the patient to go home overnight. The portable systems are modular, supporting a multitude of sensors required for polysomnographic analysis. Recent attempts to integrate sensors and electrodes into textiles made the recorders more wearable. Despite these advances, the devices often remain bulky. Furthermore, the portable systems were only used for recording and not for signal processing or classification. Instead of polysomnographic recordings, the less accurate actigraphy method is often used for long-term sleep studies [7], [8]. Actigraphy is a passive measure of sleep/wake behavior. Miniature accelerometers in a watch-like device are used to record the movement patterns of the subject. These wristbands are small, light weight, and low power and therefore easy to wear over multiple days. Several classification algorithms have been suggested for actigraphy analysis [9]–[13]. However, they do not provide real-time detection of sleep and wake. Furthermore, they often incorrectly classify low activity tasks (e.g., reading or watching television) as sleep because the measured behavioral quiescence is not unique to sleep [8], [11]. Furthermore, actigraphy is not a good tool for detecting wakefulness in subjects with irregular or fragmented sleep schedules [14].

We have previously demonstrated online sleep/wake classification based on power spectral density estimates of ECG, respiration effort (RSP) [4], and optionally accelerometer (ACC) signals [15]. We showed that if an artificial neural network (ANN) is trained and tested later on the same user, a mean correct sleep–wake classification of 94.23 ± 1.65% can be achieved [15]. However, when the ANN classifier was tested on data from users who did not contribute to the training of the classifier, the accuracy dropped significantly to 88.59 ± 6.66%. This indicated that at least some of the signals do not generalize well for other users and a single model cannot be used for accurate classification in a larger population. In our previous work, we remarked that an ANN could be trained for each user individually [4]. However, obtaining the necessary training dataset with accurate

Manuscript received July 11, 2010; accepted November 16, 2010. Date of publication December 17, 2010; date of current version March 18, 2011. This work was supported by the Solar Impulse project grant of Ecole Polytechnique Fédérale de Lausanne (EPFL). The work of D. Floreano was supported in part by the CURVACE Project sponsored by the Future and Emerging Technology Division within the seventh Framework Programme for Research of the European Commission under FET-Open Grant 237940. *Asterisk indicates corresponding author.*

*W. Karlen is with the Electrical and Computer Engineering in Medicine Group, the University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: walterk@ece.ubc.ca).

D. Floreano is with the Laboratory of Intelligent Systems, Institute of Microengineering, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland (e-mail: dario.floreano@epfl.ch).

Digital Object Identifier 10.1109/TBME.2010.2097261

sleep/wake labels for the supervised training of the ANN was very time consuming. This was because the procedure required setting up sensors for home video, electrooculography (EOG), and electromyography (EMG), and a technician manually analyzed the recordings. Further, the applications for the classifier were limited to people that were willing to undergo the training procedure.

To address intersubject differences in automated sleep scoring and sleep disorder classification from polysomnography recordings, different classification models for different subject groups are used [16]. Typically, clustering algorithms were used to associate the biomedical signals from a new subject with a subject group [17]. Subject groups were built off-line from previously classified signals stored in a database. This approach required significant amount of processing and storage resources. The need for large datasets with accurate prelabeled data also required considerable time investments and human intervention. Such a clustering and database approach is, therefore, not conceivable for an autonomous wearable system. An adaptation method for off-line actigraphy analysis of sleep and wake has been suggested [12]. The density of movements of the subject was calculated to adjust two thresholds used for the sleep–wake discrimination. The movement density was calculated off-line over the whole duration of the recording. Further, the described experiments only analyzed the periods when the subjects were in bed. This method of gathering *a priori* knowledge for the algorithm adaptation is neither practical nor available for wearable real-time applications. Another possible adaptation strategy was the tuning of the classification threshold of an ANN [4]. This simple method required only one parameter to be adjusted. The tuning was very limited, was performed off-line, and did not allow for adaptation to possible changes in the wearer's physiology. This tuning resulted in a statistically insignificant increase of the mean accuracy of only 1.43% for the given data and ANN topologies [4].

We introduce a new way to improve the classification accuracy with an online algorithm because the subject-independent networks did not show the desired accuracy for new users [15]. We decided to modify directly the ANN weights with a learning algorithm as it was used for the off-line training of subject-independent classifiers. To adjust the weights in a supervised manner, some *a priori* knowledge about the user's sleep/wake state is required. Video analysis, polysomnography, or any other known sleep detection methods described earlier would need some off-line analysis by a human or machine expert and were not suitable for an online data labeling on a wearable system. Equally, unsupervised clustering methods like the one used in [18] are too computationally intensive. Further, it is very unlikely that unsupervised training can find a more accurate classifier than a supervised, subject-specific training. We, therefore, suggest two new feedback methods to gather *a priori* knowledge and to automatically label the recorded data online. The feedback methods make use of typical behaviors that are used to differentiate sleep from wake by observation. Typical behaviors are:

1) specific body posture;
2) maintained behavioral quiescence;

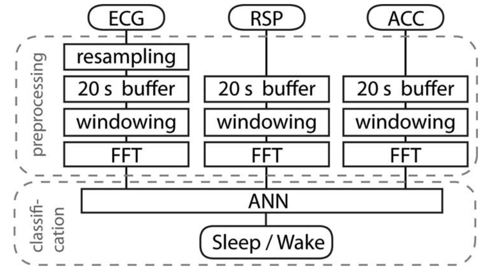

Fig. 1. Overview of the sleep/wake classification algorithm. Raw ECG, RSP and ACC signals are recorded and stored in a buffer for 20 s. Then a window function is applied and a short-time fast Fourier transformation (FFT) is used to calculate the spectral power density. The resulting frequency data are fed to a feed-forward ANN with a tangent-sigmoid transfer function. A symmetric classification threshold separates the ANN output into sleep or wake.

3) elevated arousal threshold;
4) state reversibility with stimulation [19].

We decided to monitor the user's activity because it can be passively recorded with an accelerometer. We also monitor user's reactivity to give an estimate of the arousal threshold. This measure can be easily recorded with a reaction task.

In the following sections, we present the research and development of algorithms for user-adaptive sleep–wake discrimination and experimental classification results. The experiments were performed on a wearable, energy efficient device called *SleePic* (derived from *sleep* and *programmable interface controller*). The SleePic device has been custom designed for our experiments. It is composed of a chest-worn belt that records ECG, RSP, and three-axis ACC, and a wristwatch that acts as a user interface with LEDs and a button. A detailed description of the SleePic hardware can be found in [20]. The SleePic embeds the previously developed sleep/wake classification algorithm and a newly developed method to adapt to different users automatically. The adaptation method does require only minimal user interaction and does not need the supplementary and constraining video, EMG, and EOG recordings that were used in our previous studies [4], [15]. The presented methods demonstrate the first step toward the development of context-aware personal health devices that are able to adapt to the user autonomously.

II. Algorithm Description

The goal of our work was to develop an algorithm for cardiorespiratory sleep/wake classification that is able to adapt to intersubject differences automatically. The algorithm had to be power efficient so that it could run on a wearable device. Further, for a high user acceptance, the algorithm had to rely on low user interaction to minimally disturb the subject in her/his daily activities. The algorithm was composed of two stages:

1) sleep/wake discrimination with an ANN classifier (see Fig. 1);

2) adaptation procedure that automatically labeled data segments and adapted the ANN to the user (see Fig. 2).

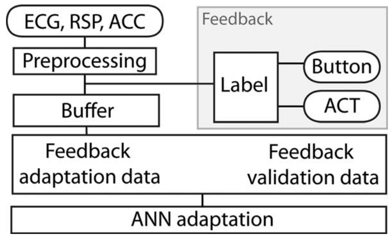

Fig. 2. Overview of the data flow for the automatic labeling of sleep/wake for the cardio-respiratory and accelerometer data (ECG, RSP, ACC) that were used for the adaptation of the ANN classifier. *A priori* label information is obtained from button and activity (ACT) data.

A. Sleep/Wake Discrimination

The sleep/wake classifier was based on the processing of cardio-respiratory signals. We also included the processing of ACC data because it is considered the most power-efficient signal to record on a wearable system. We have described the classifier in detail previously [4], [15] and present only the differences from the original version next.

The ECG, RSP, and ACC signals were sampled over segments of 20 s (see Fig. 1). The high sampling rate of ECG was reduced from 256 to 51.2 Hz to simultaneously fit all segments of the three signals into the RAM of the SleePic microcontroller. The 20-s segment size corresponded to 1024 sampling points (ECG) and 512 sampling points (RSP and ACC), respectively. The power of two size of the segments was favorable for the processing of the fast Fourier transform (FFT) on the microcontroller. On each segment, a Hamming window function was applied to reduce the border effects of the time–frequency transformation. The frequency content was extracted from each segment with an FFT. The content of the frequency bands obtained by the FFT were then fed to a feed-forward, single-layer ANN with a tangent-sigmoid output function stored in a lookup table. The size of the ANN varied depending on how many input signals were selected. A symmetric threshold was applied to classify the continuous ANN output into sleep and wake. The input network weights of the ANN were found to be redundant and not all necessary for the successful classification of sleep and wake [21]. Therefore, we created a different network topology for this study that used only the relevant input weights. In our particular case (single-layered network), all input features $i \in 1, \ldots, N$ were considered as relevant when the mean weight over all training runs 1 to M was larger than the median standard deviation of all layer weights of all runs, as follows:

$$S(w_i) = \begin{cases} 1, & \text{mean}(\vec{w}_i) \\ & \quad > \text{median}(std(\vec{w}_{1,\ldots,N})) \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $\vec{w}_i = (w_i^1, \ldots, w_i^M)$ and S is the selection function. The input size of the resulting pruned network was reduced to 8.3% of its original size. Only the reduced network topology was used in this study for the training, testing, and adaptation. We used the Levenberg–Marquardt back-propagation algorithm [22] to train the ANN and update the synaptic weights.

B. Adaptation Procedure

For the online adaptation of the ANN weights, *a priori* knowledge about the user's sleep/wake state was required. We developed two new feedback methods to automatically label the recorded data for supervised adaptation (see Fig. 2).

1) Activity Feedback: From actigraphy, we know that certain movement patterns can reliably be associated with a sleep or wake behavior [7], [8]. We, therefore, used actigraphy measures to obtain a number of labeled physiological data segments without user interaction at all. Inspired from the algorithm of Cole [10], the algorithm examined the current, four prior, and two posterior activity data segments (ACT) of 1-min size. An ACT consisted of the sum of activity zero-crossing within that segment. If the activity of the resulting 7-min window was very low (high), the algorithm considered the central 20 s as sleep (wake); otherwise, the algorithm did not label the data and the data were not selected for the adaptation set ($f = NaN$), as follows:

$$f(\text{ACT}) = \begin{cases} 1, & \text{if ACT}_{n-4,\ldots,n+2} \leq 1 \\ -1, & \text{if ACT}_{n-4,\ldots,n+2} \geq 10 \\ NaN, & \text{otherwise} \end{cases} \tag{2}$$

where $f = 1$ equals sleep and $f = -1$ equals wake.

2) Button Feedback: Humans are subject to an elevated arousal threshold during sleep [19]. A stimulation below this threshold will cause no reaction of the subject during sleep, but eventually will during wake. We applied this method to our algorithm by stimulating the wearer of the SleePic with a blinking LED on the SleePic Watch and simultaneously with a single, light vibration on the chest. Both stimuli could not be perceived during sleep. The stimuli were randomly generated by the SleePic every 15–60 min. If the wearer reacted to this stimulation by pressing the button on the SleePic Watch within 1 min ($button = 1$), he/she was considered as awake ($f = -1$). If a response was absent ($button = 0$), the wearer was either asleep or missed the stimulus. In that case, the ACT within the stimulus period was analyzed. If it was below or equal to a threshold of one zero-crossing, the wearer was considered as asleep ($f = 1$) and otherwise, no automatic labeling was performed ($f = NaN$) as follows:

$$f(button) = \begin{cases} -1, & \text{if } button = 1 \\ 1, & \text{if } button = 0 \\ & \quad \text{and } ACT \leq 1 \\ NaN, & \text{otherwise.} \end{cases} \tag{3}$$

The data from the 20-s prior to the stimulus were labeled accordingly to avoid training on the button pressing movement patterns that may arise during this period.

III. Methods and Materials

SleePic was used to demonstrate and test the developed algorithms. The algorithms did run in real time on SleePic, but because of the nature of the experimental design, it was not possible to perform the computing tasks in real time. Instead, the computing was done *post hoc*. This procedure did not alter the performance of the algorithm.

A. Subjects and Recordings

Following informed consent, eight volunteers (two females and six males) aged 24–30 years wore the SleePic system. The subjects were in good health and reported no cardio-respiratory disease or any sleep disorders. The subjects came to the laboratory in the evening and were instructed about the experiment procedure and how to wear the device. The subjects wore the SleePic device for a minimum of 36 h that included two nights. They were allowed to remove the belt during heavy sport or when showering. During the whole experiment, the subjects performed a randomly scheduled reaction task using the button on the SleePic Watch. The subjects were asked to sleep at home. After the recording, the subjects returned the SleePic recording system to the laboratory, were debriefed, and filled out a questionnaire about the usability and comfort of the system.

Because of the ambulatory nature of the experiment, the subjects were expected to move freely and perform normal daily activities. Therefore, we did not consider the possibility of recording EEG signals for reference. Instead, the subjects had to maintain a logbook by indicating the system-off times, their sleep times, and particular events related to the system that may happen during the experiment. Additionally, a technician installed an infrared video camera in the bedroom to record the sleep behavior during bedtime. A technician analyzed the logbook and video recordings and labeled the wake/sleep periods in 10-s intervals. Afterward, the technician removed data epochs from the SleePic data for periods where the SleePic was not worn. When the SleePic recording device failed to record any data, the missing data epochs were also discarded. However, signals with movement artifacts or other task-dependent disturbances were not discarded, since they might contain useful information for the classification. With one subject, the sensor belt became too loose, which was not detected immediately during the recording. Therefore, an additional 3.5 h with bad data were discarded for this recording. The cardio-respiratory and activity data obtained from the SleePic system were used for the classification experiments. The expert sleep/wake labels obtained from the logbook and the video analysis were solely used for the performance assessment of the algorithm.

B. Classification Experiments

We conducted a series of experiments to evaluate the adaptation strategies. Data from the SleePic system were used to train and adapt three different network topologies. The topologies differed in the input signal vector. The networks topologies consisted either of the features from the cardio-respiratory signals (ECGRSP), the activity features (ACC), or the combination thereof (ECGRSPACC).

1) Subject-Independent Experiments: In a first set of experiments, we replicated the generalization experiments from our previous study [4] but using SleePic data. The expert-labeled data from each subject were randomly divided into a *training dataset* (80%) and a *validation dataset* (20%). The *test dataset* consisted of the full dataset of each subject (see Fig. 3). In order to prevent any performance bias, training and test datasets from the same dataset were never used simultaneously within an experiment. We trained networks for each ANN topology by using the training datasets of all subjects but one.[1] The performance of the network was evaluated after each training iteration on the validation datasets from the same subjects in the training dataset. The training was stopped if the performance of the validation did not increase for more than five iterations. The test datasets from the remaining subject were used to measure the performance of the network after the training was completed. We repeated the experiment until every subject was once in the test dataset (eight times). Ten independent runs for each experiment and subject were performed from different initial network weight values. Initialization of the weights was done with the Nguyen–Widrow method [23].

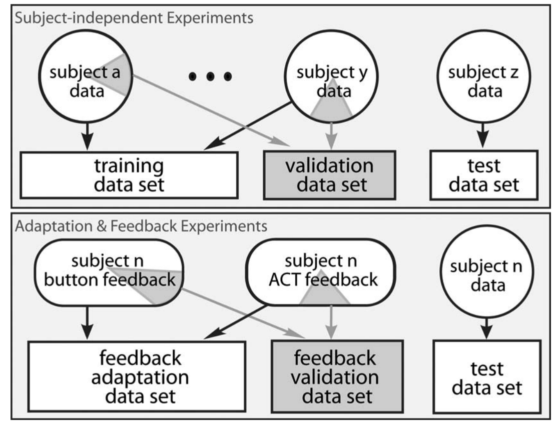

Fig. 3. Classification experiments. In the subject-independent experiments (top), data from all subjects but one contributed to training (80%) and validation (20%) datasets. The data from the remaining subject consisted of the test dataset. In the adaptation experiments (bottom), the data from one subject were again used as a test dataset. The feedback adaptation set and the feedback validation set were composed of the button and activity (ACT) feedback from the same subject. For the feedback experiments, the button and ACT feedback data were used separately to build the feedback adaptation and validation sets. All the experiments were repeated until every subject was once in the test dataset.

2) Adaptation Experiments: In this set of experiments, the feedback adaptation dataset was used individually to adapt the generalized networks obtained in the subject-independent experiments. The feedback data (labeled by the Button and Activity Feedback algorithms) were combined and randomly split into a *feedback adaptation dataset* (80%) and a *feedback validation dataset* (20%). For each subject, ten different feedback datasets were generated (see Fig. 3, bottom). The feedback validation dataset was used to stop the training and avoid overfitting. The same data as in the subject-independent experiments were used as a test dataset. The best network obtained from the subject-independent experiments for each topology was used as a start network for the adaptation procedure.[2] Ten independent runs for each feedback adaptation and validation dataset were generated. This was repeated for each subject (eight times) and each topology, making a total of 240 runs.

[1]The training parameters for the subject-independent experiments were μ: 0.001; μ increase: 10; μ decrease: 0.1; μ max: 10^{10}; min gradient: 10^{-10}; max validation failures: 5.

[2]The training parameters for the adaptation experiments were μ: 0.001; μ increase: 2; μ decrease: 0.1; μ max: 10^{10}; min gradient: 10^{-10}; max validation failures: 2.

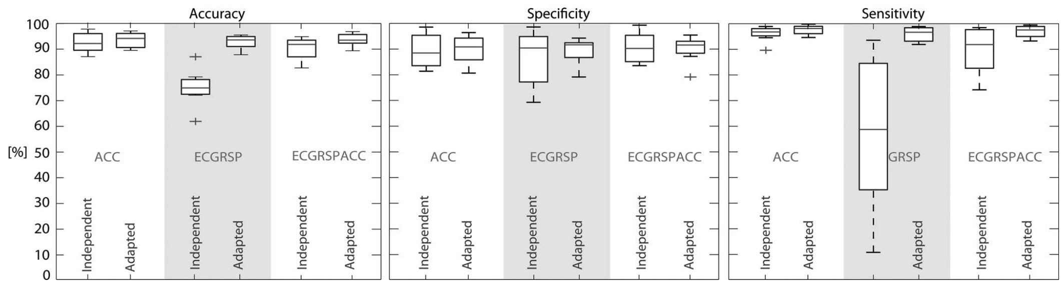

Fig. 4. Accuracy (left), specificity (middle), and sensitivity (right) of the three different ANN input topologies (ACC, ECGRSP, ECGRSPACC). Left boxes: results of subject-independent networks (independent). Right boxes: results when adapting the subject-independent networks with the adaptation dataset (adapted). The horizontal lines of each box are the lower quartile, median, and upper quartile values (from bottom to top). The whiskers represent the most extreme values within 1.5 times the interquartile range from the quartile. The outliers (crosses) are data with values beyond the ends of the whiskers.

3) Feedback Experiments: With the feedback experiments, we analyzed the individual contributions from each feedback strategy to the performance of the networks. For this, the adaptation experiments were repeated with only the automatically labeled data from the Button Feedback or the Activity Feedback in the adaptation and validation dataset, respectively.

C. Performance Assessment

To evaluate the performance of the classifiers, we calculated the accuracy (fraction of all correctly classified segments), the sensitivity (fraction of correctly classified sleep segments), and the specificity (fraction of correctly classified wake segments). These performance measures were calculated on all 20-s segments for each experiment and topology, as follows:

$$\text{accuracy} = \frac{\#\ \text{true sleep seg} + \#\text{true wake seg}}{\#\ \text{all seg}} \tag{4}$$

$$\text{sensitivity} = \frac{\#\ \text{true sleep seg}}{\#\ \text{all sleep seg}} \tag{5}$$

$$\text{specificity} = \frac{\#\ \text{true wake seg}}{\#\ \text{all wake seg}}. \tag{6}$$

We estimated the quality of sleep with the sleep efficiency parameter, calculated as the total classified sleep time divided by the time in bed.

IV. Results

A total of 250 h of valid SleePic recordings were obtained (37% sleep and 63% wake). On average, 1.83 labeled segments per hour were obtained from the Button Feedback (58 ± 13.2 labels per recording), containing 36.95 ± 10.27% sleep labels. This corresponds to an equivalent sleep/wake proportion as for the entire recording. Using the labeling rule from the Button Feedback method, neither false-positive nor false-negative labels were generated. The automatic labeling from the Activity Feedback contained on average 14 labeled segments per sleep hour and 8.9 per wake hour (335.88 ± 99.85 per recording). Using this labeling rule, the algorithm generated a total of 27 false sleep and 6 false wake labels that corresponded to an error rate of 1.3%. The wrong labels were not discarded for the adaptation experiments.

The accuracy, sensitivity, or specificity of the subject-independent experiments with topologies containing the frequency features of ACC data were statistically better than the topology without the ACC (left boxes in Fig. 4; Student's t-test, $p < 0.05$, for both cases).

TABLE I
MEAN PERFORMANCE OF FEEDBACK METHODS [% ± SD]

Feedback	accuracy	sensitivity	specificity
ACC			
Subject-independent	90.44 ± 3.57	91.39 ± 3.91	89.92 ± 4.34
Button	91.63 ± 4.41	94.69 ± 6.59	87.54 ± 7.89
Activity	92.62 ± 3.08	97.14 ± 2.49	88.25 ± 6.26
Button & Activity	92.98 ± 3.19	96.71 ± 2.44	88.97 ± 6.44
ECGRSP			
Subject-independent	74.94 ± 6.76	57.53 ± 28.06	86.58 ± 10.89
Button	76.64 ± 9.25	69.65 ± 23.71	79.80 ± 12.41
Activity	91.06 ± 3.44	95.79 ± 2.63	87.15 ± 4.80
Button & Activity	91.12 ± 3.43	95.16 ± 3.17	87.20 ± 4.94
ECGRSPACC			
Subject-independent	90.23 ± 4.29	89.58 ± 8.54	90.48 ± 5.56
Button	91.59 ± 4.33	92.63 ± 6.92	90.57 ± 4.52
Activity	92.67 ± 2.83	96.25 ± 3.04	89.62 ± 4.72
Button & Activity	92.94 ± 3.37	96.09 ± 3.63	90.42 ± 4.72

TABLE II
MEAN SLEEP EFFICIENCY AFTER ADAPTATION [% ± SD]

Expert	ACC	ECGRSP	ECGRSPACC
84.98 ± 8.66	93.07 ± 5.86	95.82 ± 6.42	92.39 ± 4.74

The adaptation method improved the accuracy of all topologies containing cardio-respiratory features as inputs (Student's t-test, $p < 0.01$). The adaptation method had a larger impact on the sensitivity than the specificity for both topologies containing cardio-respiratory signals (see Table I). The accuracy of the adaptation experiments showed no significant difference compared to the subject-independent experiments in the ACC topology (see Fig. 4; Student's t-test, $p > 0.40$).

No significant difference in accuracy, sensitivity, or specificity between the three feedback methods can be observed for the ACC and ECGRSPACC topologies (see Table I; Student's t-test, $p > 0.75$). Furthermore, the sensitivity for the Activity Feedback showed a reduced standard deviation compared to the Button Feedback. This indicates that the falsely labeled Activity Feedback labels had no negative influence on the adaptation performance. Button Feedback alone was not able to improve the accuracy of the ECGRSP topology.

The adaptation algorithm significantly overestimated sleep efficiency (see Table II). This means that most classifier models estimated the sleep quality of the subjects to be better than it was detected by the human expert.

V. Discussion

We aimed to design a power-efficient algorithm for wearable sleep/wake classification. The experiments successfully showed that the presented algorithm can be embedded in a wearable device with an autonomy of more than 36 h.

The classification test results in Fig. 4 indicated that the specificity of the ACC topology was higher than expected from the literature that analyzed previous actigraphy algorithms [8], [11]. Two effects might have contributed to this result: 1) the SleePic was measuring the activity of the subjects based on the movements of the body and not of the wrist. However, the location of measurement should not significantly change the detection of motor activity [24]; and 2) in addition to motor activity, the features computed by the SleePic algorithm also contained information about body position. The body position was encoded in the low-frequency component of the FFT preprocessing. This information was not available to algorithms in traditional wrist actigraphy. The high correlation between effective sleep efficiency (percentage of sleep when in bed) and ACC sensitivity (percentage of correct sleep classification) supported the hypothesis that the ANN classifier is using the body position as a valuable classification feature. (Kendall correlation $\tau = 0.90$, $p < 0.01$.) Further experiments including wrist actigraphy are required to study this effect in more detail.

The experiments showed that the detection of sleep was more difficult than wake (low sensitivity) and had the highest impact in reducing the subject-independent performance of the ECGRSP topology. This suggested that the intersubject variation of ECG and RSP was mostly present during sleep. In fact, the classification models that presented sensitivities below 50% belonged to two subjects that had sleep patterns that would correspond to wake patterns in the other subjects. This high discrepancy between subjects during sleep strengthens our postulation for the need for a user adapting device.

A. Subject-Independent Versus Adapted Systems

Our observations suggest that the ACC data were able to generalize well between different subjects and adaptation or retraining for a new subject was not necessary (see Fig. 4). The use of accelerometer data recorded from the chest might, therefore, be an appropriate alternative to a cardio-respiratory classifier that requires model adaptation. However, studies have shown that accelerometer data alone do not accurately classify wake states containing low activity [8], [11]. To evaluate this effect, additional experiments with subjects that present fewer movement patterns during wake are required.

Sleep patterns of ECG and RSP contained larger intersubject variations than the wake patterns. Adaption was able to address these variations. Specificity could not significantly be improved. This indicated that the wake patterns contained intrasubject differences, which were difficult to separate with the single-layer ANN classifier used for our experiments.

B. Button Versus Activity Feedback

The high labeling accuracies obtained with the Button Feedback (100%) and Activity Feedback (98.2%) indicated that both methods were robust strategies to obtain automatically labeled sleep/wake data. Although the labeling rules for the Activity Feedback were much more conservative than commonly used actigraphy algorithms, mislabeling could not be prevented. This had no effect on the adaptation performance.

We repeated the adaptation experiment with each labeling source to qualify the data obtained from the different feedback methods for the automatic labeling. The size of the adaptation dataset collected by the Button Feedback was too small to improve the classification of the ECGRSP topology (see Table I). The actively sampled Button Feedback required some attention of the wearer. Therefore, increasing the frequency of gathering this feedback could lead to more discomfort. The different strategies for the labeling might be also complementary. This can be explained by the nature of the feedback adaptation data. Whereas data from the Activity Feedback came only from clearly classifiable segments of sleep and wake passively sampled from accelerometer data, the randomly sampled Button Feedback data also contained segments that were more difficult to classify, e.g., where subjects displayed low activity when awake. Gathering of a larger adaptation set for more specific adaptation data in further experiments could be improved with a modified labeling rule. We suggest using a combined solution where the button pressing task is not activated randomly, but by using prior knowledge. The unthresholded output of the ANN is a possible source of prior knowledge. For example, if the output is close to the classification threshold where the classification uncertainty is increased, an additional reactivity test could be useful.

VI. Conclusion

We demonstrated that embedded, subject-independent models used for online sleep–wake classification can be successfully adapted to previously unseen subjects without the intervention of human experts. We have shown that for a topology that is based only on accelerometer data, the maximal accuracy can be reached when it is trained for a subject-independent application. An adaptation to a new user has only minimal effects on the performance of such a classifier. In the same way, we have shown that the ANN classifiers that were based on a cardio-respiratory signal topology can be improved significantly by adapting the neural weights. Although the accuracy of the adapted networks was not significantly higher than the ones from the accelerometer based networks, the use of cardio-respiratory signals for the classification could display an advantage when higher specificity is required.

The main achievement was the description and evaluation of two methods for automatic gathering of labeled data about the subjects' sleep/wake states. Both methods used measurements of typical behaviors that are associated with normal sleep and wake, notably an increased arousal threshold and maintained behavioral quiescence during sleep. The suggested methodology is based only on occasional button pressing of the subject and the measurement of the user's activity which makes the method power and computationally efficient.

The conducted experiments show some limitations. The duration of the experiments was not sufficient to assess the robustness of the adaptation algorithm. The duration did not allow us to monitor intrasubject variations and possible effects thereof on the adaptation and consequently on the classification performance. The SleePic device that embeds the described algorithms will also need to be tested on groups experiencing sleep disorders. For clinical applications, the system and algorithms will need to undergo further tests with more subjects and including data from a wider population.

Because of the simplicity and the low sensor requirements of the newly described method, it is not limited to the cardio-respiratory sleep/wake classification, but could also be used for automatic adaptation of other sleep discrimination algorithms.

ACKNOWLEDGMENT

The authors would like to thank G. Dumont, S. Wischmann, C. Mattiussi, and C. Brouse for their rich comments on previous versions of this document. They also thank all subjects who accepted to participate in this study.

REFERENCES

[1] H. Colten and B. Altevogt, *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*. H. R. Colten and B. M. Altevogt, Eds. Washington, DC: National Academies Press, 2006.

[2] R. Ogilvie, "The process of falling asleep," *Sleep Med. Rev.*, vol. 5, no. 3, pp. 247–270, 2001.

[3] J. Buckelmüller, H.-P. Landolt, H. H. Stassen, and P. Achermann, "Trait-like individual differences in the human sleep electroencephalogram.," *Neuroscience*, vol. 138, pp. 351–356, 2006.

[4] W. Karlen, C. Mattiussi, and D. Floreano, "Sleep and wake classification with ECG and respiratory effort signals," *IEEE Trans. Biomed. Circuits Syst.*, vol. 3, no. 2, pp. 71–78, Apr. 2009.

[5] S. Redmond and C. Heneghan, "Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 485–496, Mar. 2006.

[6] C. Iber, A. Chesson, and S. Quan, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specifications*. Westchester, NY: Am. Acad. of Sleep Med., 2007.

[7] 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. 957–965, 2001.

[8] A. Sadeh, "The role of actigraphy in sleep medicine," *Sleep Med. Rev.*, vol. 6, no. 2, pp. 113–124, 2002.

[9] J. Lötjönen, I. Korhonen, K. Hirvonen, S. Eskelinen, M. Myllymäki, and M. Partinen, "Automatic sleep-wake and nap analysis with a new wrist worn online activity monitoring device vivago WristCare.," *Sleep*, vol. 26, no. 1, pp. 86–90, 2003.

[10] R. J. Cole, D. F. Kripke, W. Gruen, D. J. Mullaney, and J. C. Gillin, "Automatic sleep/wake identification from wrist activity," *Sleep*, vol. 15, no. 5, pp. 461–469, 1992.

[11] L. de Souza, A. Benedito-Silva, M. Pires, D. Poyares, S. Tufik, and H. Calil, "Further validation of actigraphy for sleep studies," *Sleep*, vol. 26, no. 1, pp. 81–85, 2003.

[12] J. Hedner, G. Pillar, S. D. Pittman, D. Zou, L. Grote, and D. P. White, "A novel adaptive wrist actigraphy algorithm for sleep-wake assessment in sleep apnea patients," *Sleep*, vol. 27, no. 8, pp. 1560–1566, 2004.

[13] S. Edward, S. Nadezhda, S. Stephanie, N. Michael, and C. S. Group, "Activity-based sleep/wake identification in infants," *Physiol. Meas.*, vol. 25, pp. 1291–1304, 2004.

[14] J. Paquet, A. Kawinska, and J. Carrier, "Wake detection capacity of actigraphy during sleep," *Sleep*, vol. 30, no. 10, pp. 1362–1369, 2007.

[15] W. Karlen, C. Mattiussi, and D. Floreano, "Improving actigraph sleep/wake classification with cardio-respiratory signals," in *Proc. IEEE Annu. Int. Conf. Eng. Med. Biol. Soc.*, Vancouver, BC, Canada, vol. 2008, pp. 5262–5265.

[16] T. Penzel and R. Conradt, "Computer based sleep recording and analysis," *Sleep Med. Rev.*, vol. 4, no. 2, pp. 131–148, 2000.

[17] P. Anderer, G. Gruber, S. Parapatics, M. Woertz, T. Miazhynskaia, G. Klosch, B. Saletu, J. Zeitlhofer, M. J. Barbanoj, H. Danker-Hopfe, S.-L. Himanen, B. Kemp, T. Penzel, M. Grozinger, D. Kunz, P. Rappelsberger, A. Schlogl, and G. Dorffner, "An E-health solution for automatic sleep classification according to Rechtschaffen and Kales: Validation study of the Somnolyzer 24 $\times$ 7 utilizing the Siesta database," *Neuropsychobiology*, vol. 51, pp. 115–133, 2005.

[18] A. Krause, A. Smailagic, and D. Siewiorek, "Context-aware mobile computing: Learning context- dependent personal preferences from a wearable sensor array," *IEEE Trans. Mobile Comput.*, vol. 5, no. 2, pp. 113–127, Feb. 2006.

[19] W. F. Flanigan, "Behavioral," in *The Sleeping Brain: Perspectives in the Brain Sciences* (ser. Perspectives in the Brain Sciences). Los Angeles, CA: Brain Info. Service/Brain Res. Inst., 1972, vol. 1, pp. 14–18.

[20] W. Karlen and D. Floreano, "SleepPic. Hardware developments for a wearable on-line sleep and wake discrimination system," in *Proc. BIOSIGNALS 2011—Int. Conf. Bio-inspired Syst. Sig. Process.*, F. Babiloni, A. Fred, J. Filipe, and H. Gamboa, Eds., Rome, Italy: SciTePress, Jan. 26–29, 2011, pp. 132–137.

[21] P. Dürr, W. Karlen, J. Guignard, C. Mattiussi, and D. Floreano, "Evolutionary selection of features for sleep/wake discrimination," *J. Artif. Evol. Appl.*, pp. 1–10, 2009.

[22] M. Hagan and M. Menhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Trans. Neural Netw.*, vol. 5, no. 6, pp. 989–993, Nov. 1994.

[23] D. Nguyen and B. Widrow, "Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights," in *Proc. Int. Joint Conf. Neural Netw.*, Washington, DC, 1990, vol. 3, pp. 21–26.

[24] H. A. M. Middelkoop, E. M. Dam, D. A. den Doel, and G. Dijk, "45-Hour continuous quintuple-site actimetry: Relations between trunk and limb movements and effects of circadian sleep-wake rhythmicity," *Psychophysiology*, vol. 34, no. 2, pp. 199–203, 1997.

Walter Karlen (M'08) received the M.Sc. degree in microengineering in 2005, and the Ph.D. degree in computer, communication, and information sciences in April 2009, both from the Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

While doing Ph.D. work, he was a Research Assistant at the Laboratory of Intelligent Systems at EPFL. Concurrently, he was a Scientific Consultant for physiological monitoring in extreme environments for Solar Impulse S.A., Switzerland. He is currently a Postdoctoral Fellow at the Electrical and Computer Engineering in Medicine Group, University of British Columbia, Vancouver, BC, Canada. His research interests include biomedical signal processing, adaptive algorithms, embedded and wearable devices, and mobile biomedical sensors and systems. His research focuses on clinical and health applications in anesthesia, sleep, attention, and fatigue.

Dr. Karlen is a Fellow of the Canadian Institutes of Health Research (CIHR).

Dario Floreano (SM'05) received the M.A. degree in visual psychophysics from the University of Trieste, Trieste, Italy, in 1988, the M.Sc. degree in neural computation from the University of Stirling, Stirling, U.K., in 1992, and the Ph.D. degree in cognitive systems and robotics from the University of Trieste, in 1995.

He is currently the Director of the Laboratory of Intelligent Systems, School of Engineering and the Director of the Swiss National Center of Competence in Robotics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. He published almost 200 peer-reviewed technical papers and edited and coauthored several books, among which *Evolutionary Robotics* with S. Nolfi (MIT Press, 2000, 2004), *Bioinspired Artificial Intelligence* with C. Mattiussi (MIT Press, 2008, 2010), and *Flying Insects and Robots* with Zufferey, Srinivasan, and Ellington (Springer-Verlag, 2009). He is a cofounder of the company senseFly.com. His research interests focus on bioinspired artificial intelligence and robotics.

Dr. Floreano is a cofounder of the International Society for Artificial Life, Inc., a member of the Advisory Group to the European Commission for Future Emerging Technologies, and an Advisory Board Member of the Institute for Advanced Studies Collegium Budapest.