Using smartwatch sensors to support the acquisition of sleep quality data for supervised machine learning

Cinzia Bernardeschi, Mario G.C.A. Cimino,
Andrea Domenici*, and Gigliola Vaglini

Department of Information Engineering,
University of Pisa, 56122 Pisa, Italy
{c.bernardeschi,m.cimino,a.domenici,g.vaglini}@ing.unipi.it

Abstract. It is a common practice in supervised learning techniques to use human judgment to label training data. For this process, data reliability is fundamental. Research on sleep quality found that human sleep stage misperception may occur. In this paper we propose that human judgment be supported by software-driven evaluation based on physiological parameters, selecting as training data only data sets for which human judgment and software evaluation are aligned. A prototype system to provide a broad-spectrum perception of sleep quality data comparable with human judgment is presented. The system requires users to wear a smartwatch recording heartbeat rate and wrist acceleration. It estimates an overall percentage of the sleep stages, to achieve an effective approximation of conventional sleep measures, and to provide a three-class sleep quality evaluation. The training data are composed of the heartbeat rate, the wrist acceleration and the three-class sleep quality. As a proof of concept, we experimented the approach on three subjects, each one over 20 nights.

Key words: Sleep monitoring, smartwatch, sleep quality estimation

1 Introduction

Supervised learning is a promising approach to a wide range of problems related to monitoring health conditions and diagnosing pathologies [1]. This work is part of a research effort aimed at designing a supervised learning architecture to detect sleep behavior shift. Behavior shift is a pattern used in broad-spectrum assessment of initial signs of disease or deviations in performance [2][3]. The availability of labeled training data is fundamental for supervised learning [4]. In the case of sleep behavior, the training data are sets of sensor data, each labeled with the related sleep quality. The objective of this study is to propose a support to the collection of per-night sleep quality data. In the literature it is well known that, besides the sleep state, other events and cognitive experiences

* Corresponding author

may influence the judgment, resulting in a cognitive bias [5]. The aim of this study is to support human judgment by collecting additional data during the monitored nights, via a familiar device, and to extract three-class sleep quality evaluations to be compared to those reported by the subject.

Polisomnography is a standard approach to sleep monitoring. It involves recording multiple physiologic variables at specialized centers, scored by human examiners on the basis of standardized criteria. It can be used for a few nights, which are insufficient for sleep habits. It is intrusive, which may disturb sleep. Consequently, it is not accurate for sleep behavior [6]. The diagnosis may vary depending on the examiner with a 20% variance [7]. Another approach is *actigraphy*, which is based on a watch-like device equipped with motion accelerometer, to monitor motion-related sleep disorders. Normal subjects show more than 90% correlation between actigraphy and polisomnography [6].

In the literature, there is a growing interest in the possibility of gathering sleep data from wearable devices. Recently developed smartwatches have been used for monitoring sleep patterns variation, because they can also feature sensors such as heartbeat rate monitor, wrist acceleration recorder, pedometer, magnetometer, barometer, ambient thermometer, oxymeter, skin conductance and temperature sensors, and GPS locator [8]. More specifically, heart rate and body movement are known to vary greatly during sleep and have a close relationship with sleep stages. Indeed, the autonomic nervous system significantly affects heart rate, and body movement is linked to sleep level [6]. It is known that wrist watch-shaped devices monitoring motion and pulse can measure sleep quality with sufficient accuracy. For example in [7] the authors evaluate a good correlation between sleep stages estimated using a wrist device and using polysomnography, by observing 45 subjects.

In this paper, smartwatches are used to gather data on individuals' physiological parameters. The main contribution of this preliminary work is a method and a software prototype to derive an identification of sleep stages as *wake*, *rapid eye movement* (REM), and *non rapid eye movement* (NREM), which are used to rate subject's sleep quality into three classes: *Normal*, *mediocre* and *scarce*. This way, human judgment can be supported by software-driven evaluation in order to identify the final training data set used for supervised learning.

This method is implemented by a *sleep stage estimator* (SSE), a Matlab application that analyzes smartwatch recordings of *heartbeat rate* (HBR) and *acceleration* of the subject's wrist. As a first step, the algorithm produces a simplified hypnogram from such recordings.

The SSE design is based on basic notions about sleep stages. During wake, body movement, as recorded by electromyography, is frequent, voluntary, and continuous. During REM sleep, movement is nearly absent, as it is potentially directed by dream activity but inhibited. During NREM sleep, movement occurs in episodic and involuntary posture shifts [9]. The use of heartbeat, recorded by electrocardiography, is a recurrent subject in the sleep staging literature. Heart rate variability (HRV) is significantly higher in REM sleep than in NREM sleep [10]. HRV has been also used to understand autonomic changes during different sleep stages in [11]. Moreover, a decision support system for sleep classification based on HRV has been presented in [12]. The process of falling asleep presents fluctuation in vigilance. Autonomic function changes during the wake-to-sleep transition, as reflected by the instantaneous HRV, are studied in [13]. Sleep staging is also performed in [14] on the basis of two channels of EEG.

The majority of the studies in the literature use clinical monitoring equipment and a very constrained experimental setting. In contrast, the present proposal has been conceived for a non-intrusive monitoring method, based on a general-purpose wearable device and minimally affecting the subject's everyday life. Clearly, the proposed sleep stage estimation provided by a smartwatch can deliver a broad-spectrum score of sleep quality, and not a precise evaluation of medical symptoms.

The paper is organized as follows: Section 2 covers materials and methods, Section 3 describes synthetically the SSE prototype design, Section 4 illustrates the sleep quality model, and discusses how the approach supports the selection of a training set. Section 5 draws conclusions and future work.

2 Materials and Methods

Figure 1 represents the overall method, from left to right. During physiological sleep, both subject perception and device sensing-logging contribute to the data acquisition.

On the side of the subject's judgment, sleep/wake perception is a process of discrimination that involves cognitive interpretation of physiological and psychological data [5]. According to this assumption, unreliable perception may occur when heterogeneous, incomplete, dynamic, uncertain aspects of experience are taken into consideration. The major events are manually annotated, either at the moment or later. In the next morning, an overall judgment of the night's sleep quality is given, scored as *normal*, *mediocre* and *scarce*.

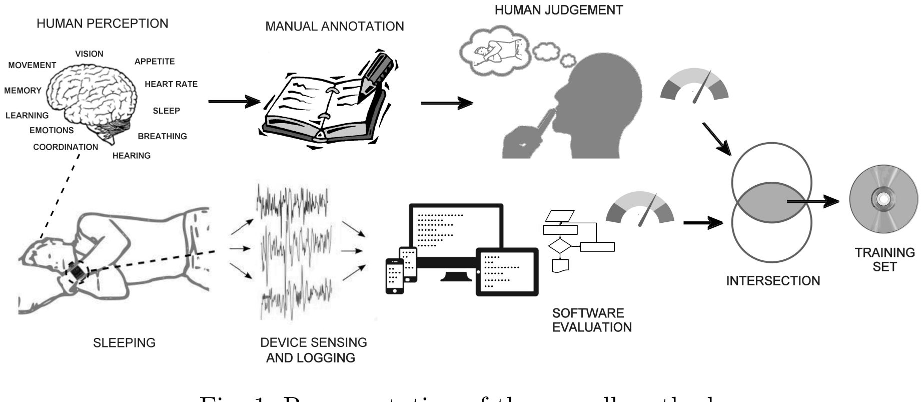


Fig. 1: Representation of the overall method.

On the side of computerized evaluation, the smartwatch senses physiological data with a heartrate monitor and an accelerometer that measures the acceleration in ms^{-2}. The smartwatch used to collect data was an LG Watch Urbane (LG-W150). The heart rate monitor uses an optical sensor to detect peaks in blood flow and computes the heart rate over an interval of time established by the constructor. The computed heart rate value is updated every tenth of a second, and the resulting time series is sampled by the SSE at one-second intervals. The acceleration data are sampled at 10 Hz. The SSE produces as an output: standard deviation of the acceleration magnitude (σ_{a}), variance of the heartbeat rate (σ^2_{hbr}), and approximate sleep staging. Also the respective plots are produced. Subsequently, a sleep quality evaluation is carried out, to produce a score with the above mentioned categories. The result can be sent to a mobile application and to a desktop application via Bluetooth and USB, respectively.

Finally, the resulting scores are compared: any night log whose computed score matches the perceived sleep quality becomes an entry of the final training set. Discordant scores are not considered for the training set. The method has been applied to three subjects of different age for 20 nights.

3 A sleep stage estimator

This section illustrates the criteria adopted by the SSE to estimate sleep stages. In essence, the subject is considered: (i) in WAKE stage, when motion level is high; (ii) in a REM stage, when motion level is low and pulse level and variability are both high; (iii) otherwise it is considered in a NREM stage.

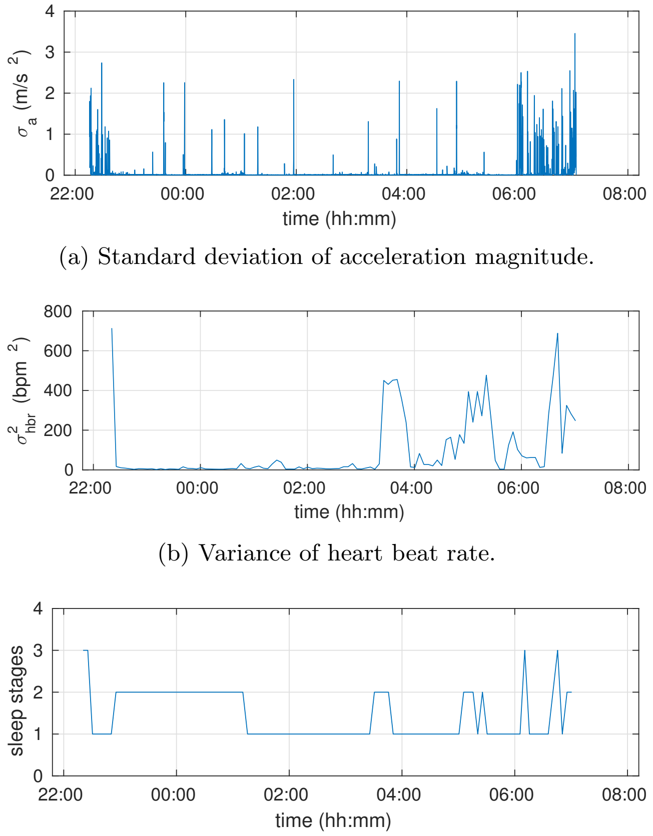
Starting from the three above criteria, the automatic evaluation of sleep stages is made according the following rules. First, the standard deviation of the acceleration magnitude (σ_{a}) is computed over a three-second sliding window. The variance is used in order to reduce the influence of the constant component due to gravity and of accelerometer noise. The mean values ($\bar{\sigma_{\mathrm{a}}}$) of σ_{a} and HBR ($\bar{h}$) are computed over the complete series. Then, average m_{hbr} and variance σ^2_{hbr} for HBR and average m_{asd} for σ_{a} are computed for each five-minute interval.

Each interval is marked as a wake, REM, or NREM period, by comparing the computed values of mean and variance against some thresholds: $m^{\mathrm{th}}_{\mathrm{hbr}}$ for HBR mean, $v^{\mathrm{th}}_{\mathrm{hbr}}$ for HBR variance, $m^{\mathrm{u}}_{\mathrm{asd}}$ and $m^{\mathrm{l}}_{\mathrm{asd}}$ for the upper and lower limits of standard deviation of the acceleration.

The criterion for staging is the following: a five-minute interval is considered a wake period if the difference δ_{asd} between m_{asd} and $\bar{\sigma_{\mathrm{a}}}$ is greater than $m^{\mathrm{u}}_{\mathrm{asd}}$. Otherwise, the interval is considered a REM period if (i) the difference δ_{hbr} between m_{hbr} and $\bar{h}$ is greater than $m^{\mathrm{th}}_{\mathrm{hbr}}$, and (ii) σ^2_{hbr} is greater than $v^{\mathrm{th}}_{\mathrm{hbr}}$, and (iii) δ_{asd} is smaller than $m^{\mathrm{l}}_{\mathrm{asd}}$. Otherwise, the interval is marked as NREM.

The resulting marks are then recorded and plotted as a hypnogram, assigning the numerical values of 3, 2, and 1 to wake, REM, and NREM periods, respectively. The threshold values have been chosen so as to maximize the matching between the stages identified by a human observer applying the three above criteria, and those generated by the SSE.

Figure 2 shows the plots of the standard deviation of the acceleration magnitude, the HBR variance, and the estimated hypnogram of a sample night of a subject, respectively. After a short NREM sleep, the hypnogram shows a two-hour REM stage, followed by fairly long periods of NREM sleep separated by REM stages.



(a) Standard deviation of acceleration magnitude.

(b) Variance of heart beat rate.

(c) Hypnogram (3: wake, 2: REM, 1: NREM).

Fig. 2: Recording of a sample night (Sept 4, 2015).

4 Sleep quality evaluation

The data used in the experiment were collected from three subjects of different age: a healthy man aged 22 (subject A), a man aged 72 (subject B), affected by minor age-related ailments, and a woman aged 88 (subject C), affected by arterial hypertension.

Human judgment on per-night sleep quality is based on three cognitive criteria: (i) the time interval between lights off and falling asleep; (ii) the number of nocturnal awakenings; (iii) the feeling of being rested after sleep. An overall judgment of the night's sleep quality is usually scored as normal (N), mediocre (M) and scarce (S). The subjects recorded their per-night sleep quality, according to the above criteria, in a sleep diary.

The software-evaluated sleep quality model is based on an estimation of wake, REM, and NREM stages. More specifically, the model is based on the following criteria: (i) sleep latency, defined as the time interval between lights off and the fall asleep; (ii) wake ratio is the ratio of the wake time divided by total time in bed; (iii) REM sleep ratio is the ratio of the REM sleep time divided by total time in bed; (iv) NREM sleep ratio is the ratio of the NREM sleep time divided by total time in bed. All these values are computed by the SSE.

We remark that human judgment and the software-evaluated sleep quality model share two criteria, namely sleep latency and nocturnal awakenings. But they also depend on two independent features, namely feeling rested after sleep and the ratio of REM and NREM stages, respectively. Thus, they are both incomplete when taken separately.

Table 1 shows the values of wake ratio W (%), REM sleep ratio R (%), NREM sleep ratio NR (%), and sleep latency SL (min) for the three subjects and for each night N. We remark that no distinction is made between wake and shallow NREM phases, which are very difficult to identify. For this reason, the value of W includes shallow NREM phases.

The table also reports the scores assigned by the subject (*diary quality*, DQ) and computed by the software (*computed quality*, CQ). Computed scores are evaluated using thresholds, which may vary for different subjects as shown in Table 2, where $\mathrm{W}^{\mathrm{th}}_{\mathrm{min}}$ and $\mathrm{W}^{\mathrm{th}}_{\mathrm{max}}$ are the lower and upper thresholds for the wake time ratio and R^{th} and $\mathrm{SL}^{\mathrm{th}}$ are for the REM sleep time ratio and sleep latency. Note that R is not considered for the elderly subjects, since it decreases with age and in practice disappears. The calculation of CQ is made according tho the following rules. Classes are ordered for decreasing quality, i.e., N, M, S.

1. CQ is initially set to the best quality, i.e., N.
2. for each W, SL, R:
   - If $W > \mathrm{W}^{\mathrm{th}}_{\mathrm{max}}$, CQ is set to the next lower class.
   - If $W < \mathrm{W}^{\mathrm{th}}_{\mathrm{min}}$, CQ is set to the next higher class.
   - If $SL > \mathrm{SL}^{\mathrm{th}}$, CQ is set to the next lower class.
   - If $R < \mathrm{R}^{\mathrm{th}}$, CQ is set to the next lower class.

The values causing CQ to move to the next lower and higher class are highlighted with minus and plus signs, respectively, in Table 1.

As a result, by comparing the values of DQ and CQ, we note that: (i) Subject A has 4 discordant cases on 20, i.e., nights 4, 13, 15, and 17); (ii) Subject B has 5 discordant cases on 20, i.e., nights 1, 4, 5, 14, and 16; (iii) Subject C has 12 cases of discordance on 20.

Table 1: Summary data.

N	subject A						subject B						subject C					
	W	R	NR	SL	DQ	CQ	W	R	NR	SL	DQ	CQ	W	R	NR	SL	DQ	CQ
1	-51	41	59	25	M	M	+24	0	100	5	M	N	+50	75	25	10	M	N
2	+27	-5	95	10	N	N	+34	0	100	10	N	N	70	30	70	60	N	N
3	39	24	76	15	N	N	-49	11	89	10	M	M	+58	89	11	15	M	N
4	+19	17	83	5	M	N	40	5	95	10	M	N	+59	29	71	10	M	N
5	+28	18	82	5	N	N	-57	3	97	25	N	M	+38	9	91	15	N	N
6	40	39	61	5	N	N	37	14	86	10	N	N	+53	84	16	10	S	N
7	+32	-9	91	5	N	N	39	2	98	20	N	N	+52	96	4	5	M	N
8	+25	32	68	10	N	N	35	2	98	5	N	N	61	7	93	10	M	N
9	38	19	81	5	N	N	-45	15	85	10	M	M	-85	13	87	35	M	M
10	+30	-0	100	15	N	N	-61	0	100	25	M	M	-76	50	50	10	S	M
11	35	11	89	5	N	N	33	31	69	10	N	N	63	23	77	5	N	N
12	+14	65	35	5	N	N	-61	0	100	20	M	M	+37	76	24	-75	M	M
13	-47	-0	100	5	N	S	-72	5	95	15	M	M	+54	100	0	15	N	N
14	+21	18	82	5	N	N	-54	0	10	30	N	M	68	16	84	15	N	N
15	+28	-0	100	15	M	N	-68	4	96	15	M	M	+19	33	67	10	M	N
16	35	20	80	5	N	N	-56	3	97	10	N	M	-80	41	59	25	N	M
17	42	-0	100	10	N	M	31	0	100	20	N	N	+46	54	46	40	M	N
18	+28	-0	100	15	N	N	42	0	100	5	N	N	+45	35	65	15	M	N
19	+19	-8	92	10	N	N	-60	0	100	10	M	M	70	33	67	10	M	N
20	+31	31	69	10	N	N	41	13	88	10	N	N	+57	5	95	10	N	N

Table 2: Thresholds for CQ quality assessment.

Subject	W^{th}_{min}	W^{th}_{max}	R^{th}	SL^{th}
A	35	45	10	30
B	35	45	-	30
C	60	70	-	60

A closer look at the discordant cases of Subject C can be usefully made, to obtain a better insight of sleep quality. The diary score is always lower than the software-computed one, and it can be ascribed to a potential cognitive bias.

5 Conclusions

Generating a reliable set of training data for sleep quality evaluation is a challenging problem, mainly due to misperceptions by a subject's judgment. In this paper we propose a novel approach to identify a reliable sleep quality data set for supervised machine learning by comparing sleep quality estimates derived from computation on physiological parameters and from human judgment. The approach is based on a software sleep stage estimator which exploits a subject's physiological parameters provided by commercially available smartwatches. Training data can be selected considering the subset of data for which the human judgment and the computed estimation concur. The experimental study on three subjects shows the viability of the approach. As a future work, the system will be cross-validated on a higher number of subjects. Further, information provided by the pedometer can be exploited to improve the software-driven evaluation.

Acknowledgements

This work was partially supported by the PRA 2016 project "Analysis of Sensory Data: from Traditional Sensors to Social Sensors" funded by the University of Pisa. The authors thank Giovanni Pollina, Silvio Bacci and Silvia Volpe for their work on the subject during their thesis.

References

1. S. J. Redmond, Q. Y. Lee, Y. Xie, and N. H. Lovell. Applications of supervised learning to biological signals: ECG signal quality and systemic vascular resistance. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 57–60, Aug 2012.
2. Asier Aztiria, Golnaz Farhadi, and Hamid Aghajan. User behavior shift detection in ambient assisted living environments. *JMIR Mhealth Uhealth*, 1(1):e6, Jun 2013.
3. Paolo Barsocchi, Mario G.C.A. Cimino, Erina Ferro, Alessandro Lazzeri, Filippo Palumbo, and Gigliola Vaglini. Monitoring elderly behavior via indoor position-based stigmergy. *Pervasive and Mobile Computing*, 23:26 – 42, 2015.
4. Mario G. C. A. Cimino, Alessandro Lazzeri, and Gigliola Vaglini. *Improving the Analysis of Context-Aware Information via Marker-Based Stigmergy and Differential Evolution*, pages 341–352. Springer International Publishing, Cham, 2015.
5. D. Weigand, L. Michael, and H. Schulz. When sleep is perceived as wakefulness: an experimental study on state perception during physiological sleep. *Journal of Sleep Research*, 16(4):346–353, 2007.
6. Yunyoung Nam, Yeesock Kim, and Jinseok Lee. Sleep monitoring based on a tri-axial accelerometer and a pressure sensor. *Sensors*, 16(5):750, 2016.
7. Takuji Suzuki, Kazushige Ouchi, Ken-ichi Kameyama, and Masaya Takahashi. Development of a sleep monitoring system with wearable vital sensor for home use. In *BIODEVICES 2009 Int. Conf. on Biomedical Electronics and Devices*, pages 326–331, 2009.
8. John J Guiry, Pepijn van de Ven, and John Nelson. Multi-sensor fusion for enhanced contextual awareness of everyday activities with ubiquitous devices. *Sensors*, 14(3):5687–5701, 2014.
9. J.A. Hobson. Sleep and dreaming. *Journal of Neuroscience*, 10(2):371–382, 1990.
10. Francesco Versace, Manola Mozzato, Giuliano De Min Tona, Corrado Cavalero, and Luciano Stegagno. Heart rate variability during sleep as a function of sleep cycle. *Biol Psychol.*, 63(2):149–162, 2003.
11. Phyllis K. Stein and Yachuan Pu. Heart rate variability, sleep and sleep disorders. *Sleep Medicine Reviews*, 16(1):47–66, 2012.
12. Martin O. Mendez, Matteo Matteucci, Vincenza Castronovo, Luigi Ferini-Strambi, Sergio Cerutti, and Bianchi Anna M. Sleep staging from heart rate variability: time-varying spectral features and hidden markov models. *Int. J. Biomedical Engineering and Technology*, 3(3/4):246–263, 2010.
13. Zvi Shinar, Solange Akselrod, Yaronn Daga, and Armanda Baharav. Autonomic changes during wakesleep transition: A heart rate variability based approach. *Autonomic Neuroscience*, 130(12):17 – 27, 2006.
14. Syed Anas Imtiaz and Rodriguez-Villegas Esther. Automatic sleep staging using state machine-controlled decision trees. In *Proceedings of the IEEE Eng. Med. Biol. Soc. 2015*, pages 378–381. IEEE, 2015.