Wrist movement analysis for long-term home sleep monitoring

Qiang Pan, Damien Brulin, Eric Campo

Wrist movement analysis for long-term home sleep monitoring

Qiang Pan, Damien Brulin, Eric Campo
LAAS-CNRS / University of Toulouse, France
Full address: 7, avenue du Colonel Roche, 31400, Toulouse, France

Abstract

In this paper, we present several original methods for classifying sleep stages, including threshold based and k-means clustering based methods. The proposed algorithms use only acceleration data from non-dominant wrist, resulting in a classification into 4-sleep stages ("awake", "light sleep", "deep sleep" and "REM (Rapid eye movement)") for overnight sleep. We validate our methods by referring to the results of "Fitbit" and subjective feedbacks from volunteers on quality of sleep. Our algorithms compute the duration of each sleep stage to evaluate changes in sleep quality between different nights. A method of calculating a sleep score based on the duration of sleep and the duration of each sleep stage is proposed, which facilitates the evaluation of sleep quality by a single score. 5 volunteers were recruited for the tests. Among all the test nights, the proposed algorithm based on k-means clustering shows a superior or equivalent performance compared to the "Fitbit" results. These promising results allow us to consider a new non-intrusive method for users and medical staff to monitor the evolution of sleep quality through long-term follow-up. In addition, to evaluate the performance of our proposed system in terms of sleep stage classification, we use the PSG (Polysomnography) sleep monitoring gold standard to monitor the sleep of one of the volunteers throughout the night in a hospital's professional sleep laboratory. This experiment shows that the proposed 5km2 (5 iterations of k-means clustering with k=2) method and the threshold method are in good agreement with the PSG results. The accuracy of awake, REM, light sleep and deep sleep detection reaches respectively 0.78, 0.96, 0.75 and 0.97 by the Threshold method. More specifically, both methods we propose show good performance in the detection of awake and deep sleep. This longitudinal monitoring can help to detect abnormal changes in sleep that are usually a sign of a change in health status.

Keywords: Sleep monitoring, wrist movement, threshold, k-means, home monitoring

1 Introduction

With the increasing social pressure and the aging of the population, more and more people are suffering from sleep problems. Good sleep quality is an important factor of good health. It has been reported that sleep disturbance is highly correlated with health deterioration (Dregan & Armstrong, 2011). According to the American Academy of Sleep Medicine (AASM) (Thorpy, 2017), there are about 90 sleep disorders, including insomnia (one third of the

population), sleep apnea syndrome (2-4%), restless leg syndrome (6%), narcolepsy (0.04%), sleep paralysis (6%), nocturnal terrors, confusional arousals and nightmares (2.2-5%) (Ohayon, 2007). Sleep disorders and sleep dysregulation can lead to medical consequences such as cardiovascular (arrhythmia, hypertension, stroke), metabolic (diabetes, obesity) and psychiatric (depression, irritability, addictive behaviors) disorders (Stephansen et al., 2018). Poor sleep quality can affect physical and mental performance, judgment and mood, and is the main preventable factor in accidents (Krieger, 2017). In consequence, effective and continuous sleep monitoring is of great significance for timely understanding and follow-up of our health condition (Leng et al., 2020; Krističević et al., 2018; Vail et al., 2009). In recent years, sleep stage classification has been a topic extensively studied as one of the most critical steps in the effective diagnosis and treatment of sleep disorders. Obtaining the time spent in the different sleep stages in the ordinary daily life environment is of great significance for research and commercial applications. For example, obtaining an accurate sleep architecture can provide better information to guide behavioral changes and provide recommendations related to sleep improvement (Daskalova et al., 2018). PSG is nowadays the gold standard for sleep monitoring and sleep stages classification (*R* or *REM* for Rapid Eye Movement, *N1* to *N3* for Non-Rapid-Eye Movement and *W* for Wake). However, it is very invasive, expensive and time consuming to implement. Hence, it is very difficult to use the PSG method as a home and long-term sleep monitoring device. This is why, in this study, we try to develop a non-invasive, less expensive sleep monitoring device suitable for home use and long-term monitoring. So far, many researchers and technicians have tried to develop simple and non-intrusive systems which allow overcoming these issues. Guettari et al. (2017) adopt self-organizing map (SOM) algorithm—Kohonen maps to achieve classification of signal segments as three phases of sleep: deep/paradoxical sleep (*R, N3*), agitated and light sleep (*N1*, *N2*) and awake phase (*W*) based on body movements signal during sleep collected from a thermopile sensor, it shows classification accuracy of 87%. Gu et al. (2015) leverage conditional random field (CRF) model to classify sleep stages into wake, light sleep, deep sleep and REM based on features of signals from microphone, accelerometer and light sensor, the detection accuracy of system is 64.55%. The term “classifier” refers to an algorithm or function that associates input data to a specific type of category. The classifier can build a classification model based on existing data. The model matchs the data to a given category or apply it to data prediction. Therefore, it is possible to build an appropriate classification model based on the concept of the classifier to achieve sleep stages classification. Chambon et al. (2018) use softmax classifier to classify sleep stages into wake, N1, N2, N3 and REM based on EEG (Electroencephalography), EOG (Electrooculography), ECG

(Electrocardiography) and EMG (Electromyography) signal from PSG. It achieves best classification performance with an accuracy of 80% when using data from 6 EEG with 2 EOG (left and right) and 3 EMG chin channels. Kumar et al. (2018) propose a coarse-to-fine-level envisioned speech recognition framework to classify images imagined by participants using an RF (Random forest) classifier based on the EEG signal. Recognition accuracies of 85.20% and 67.03% were recorded for the coarse-level and fine-level classifications, respectively. It divides the classification task into two steps: coarse and fine classification. Compared to direct classification, this approach allows different but more appropriate parameters at each classification stage in order to achieve better final performance. However, the items of the text class to be recognized have different colors and fonts. Therefore, it is impossible to determine whether the different characteristics of the collected EEG signals are due to the differences between the characters themselves or to the differences in the colors or fonts seen by the participants. Güneş et al. (2010) adopt k-means clustering as feature weighting processing and then use k-nearest neighbors and decision tree as classifier to discriminate sleep stages into awake, REM, N1, N2, N3 based on EEG signal. The best recognition accuracy is 82.21% when using k-NN classifier with k=30. Van et al. (2001) adopt modified k-means clustering to automatically detect sleep stages among awake, N1, N2, N3 and spindles based on EEG signal. The result shows that stage awake and stage N1 can be clearly distinguished, the clusters corresponding to stage N2 and stage N3 are somewhat overlapping. Velicu et al. (2016) used Kushida's algorithm derived equation to process the wrist activity data as the discriminator for wake/sleep, wake/REM and light/deep by applying three different thresholds. For a test of 3 hours and 43 minutes, approximately 2 sleep cycles were detected, each around 110 minutes. In the study of Kalkbrenner et al. (2019), body sound microphone attached to the subject's neck to record tracheal body sound is used in order to detect respiratory and heart beats and to extract cardiorespiratory features. Inertial measurement unit including accelerometer and gyroscope attached to the thoracic belt is used to record movements and sleep positions in order to extract movement features. Then a linear discriminant (LD) classifier is used for automated sleep staging obtaining 56.5% accuracy for Wake/REM/light sleep/deep sleep classification compared to PSG. Beattie et al. (2017) use a wrist-worn device that measures wrist movement using a 3D accelerometer and measures heart rate using an optical pulse photoplethysmograph (PPG). They also estimate breathing rate with measured heart rate. Based on the extracted features of left-wrist movement, heart rate and breathing rate, an overall accuracy of 69% for automated sleep staging of Wake, Light (N1 or N2), Deep (N3) and REM (REM) is obtained by using linear discriminant

classifier. A systematic review on sleep monitoring can be found in our review paper (Pan et al., 2020), where state of the art and a comparative study are included.

In the literature, most research adopts supervised machine learning methods that typically require large amounts of learning data to train the classifier and computation to implement the model. However, some works adopt unsupervised methods such as k-means clustering to achieve sleep stage classification. It is usually based on signals directly related to sleep stages such as EEG signal which is very intrusive and not easy to collect in a home environment. Acceleration data measured by an accelerometer can be used to measure the intensity of movement. The more intense the movement, the more the acceleration data changes over time. In this paper, we propose a k-means clustering approach using only the acceleration data from a sensor worn on the wrist to achieve sleep classification into four classes: wake, light sleep, deep sleep and REM. The k-means clustering method requires a relatively smaller amount of computation (Xu & Tian, 2015) which could make the algorithm implementation easier and more efficient. Moreover, the acceleration data of a wrist sensor is very easy to collect. The subject only needs to wear a small and lightweight watch like on his wrist, very suitable for home environment and long-term monitoring.

The paper is organized as follows: section 2 describes the module (or device) developed to collect movements and presents the data preprocessing; section 3 focuses on the threshold-based methods; section 4 focuses on k-means-based method; section 5 presents the tests and results on five volunteers and also the definition of a sleep score; section 6 presents a conclusion and some perspectives.

2 Data acquisition and preprocessing

2.1 Sensing device and settings

A smart module (Figure 1) already designed in our team is used as the sensing device. This module is an embedded system powered by a button battery (3V) whose basic components are:

- A NRF51822 microcontroller containing a 32-bit ARM Cortex M0 processor and a 256kB flash memory. This microcontroller is equipped with a low energy Bluetooth module, + 4dBm power and a sensitivity of -93dBm, for data transfer.
- A 2MB non-volatile FRAM memory for data backup during sleep.
- An low-power ADXL362 tri-axial accelerometer.
- I/O ports to interface other sensors, depending on the parameters to be observed.

Programs are written in C using Keil μVision.

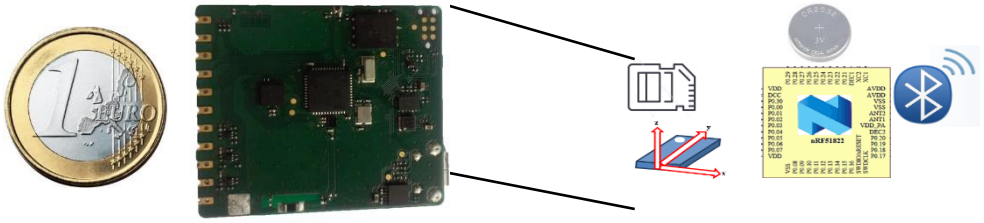

Figure 1: Integrated sensing device used.

The ADXL362 accelerometer of the smart module is adopted to acquire acceleration data in the experiment. The parameters of the ADXL362 are:

-Measurement range: -2g ~ 2g,

-Output data rate (ODR): 12.5 Hz,

-Output resolution: 8 bits.

Although the ODR is 12.5Hz, we only sample the output acceleration data every second. This reduces the amount of data and limits storage space, which could be a great advantage for long-term monitoring applications.

We position the smart module on the non-dominant wrist, wearing it like a watch as shown in Figure 2. After switching on the smart module, it will first search for the corresponding Bluetooth slave device (in this case a PC) to try to pair with it. If the smart module can pair with the Bluetooth slave device within 10 seconds (usually 10 seconds is enough for pairing if the Bluetooth slave device is advertising), it will start to send stored data in FRAM to the Bluetooth slave device for further processing. Otherwise, it will start to acquire acceleration data every second and store it in FRAM.

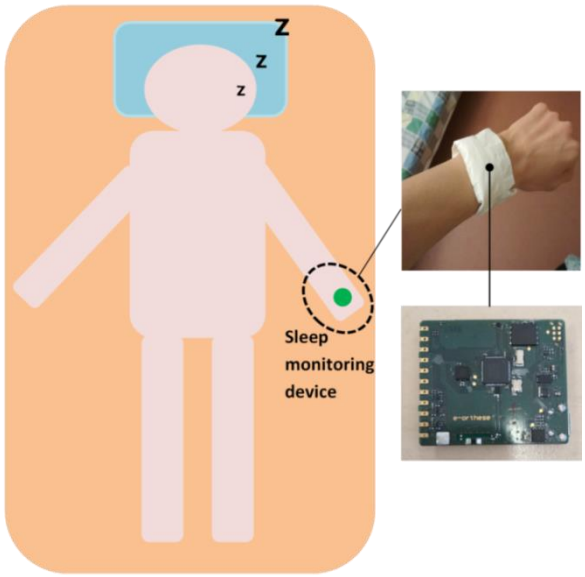

Figure 2: Position of the smart module

2.2 Data preprocessing

With acceleration values Ax, Ay, and Az, a corresponding movement level M_i for sample i, will be calculated by equation (1), where N is the number of samples.

$$M_i = |Ax_{i+1} - Ax_i| + |Ay_{i+1} - Ay_i| + |Az_{i+1} - Az_i|, \ i = 1, \ 2, \ 3, ..., \ N-1 \tag{1}$$

The overnight movement level data is cut into 30-samples epochs, noted as S_j (j = 1, 2, 3,…, L, L being the total number of epochs for one night). Each epoch is the shortest unit for further sleep stage classification, which has a duration of 30s, as in Rechtschaffen and Kales Guidelines Guidelines (Rechtschaffen & Kales, 1968). Using a sleep stage classification algorithm, each epoch will be classified as awake, light sleep, deep sleep and REM.

In each epoch, movement levels of the corresponding 30 samples are summed to obtain an epoch movement level EM_j, as in equation (2).

$$EM_j = \sum_{k=1}^{30} M_{jk}, \ \ j = 1, \ 2, \ 3, ..., \ L \tag{2}$$

Where j is the index of epochs, L is the number of epochs.

As sleep is a process that is continuously changing, it is necessary to associate the previous and subsequent periods when analyzing the sleep state at a given time. Thus, for each epoch, 9 epochs are considered before and after it. A weighted *PM* value is defined (see equation (3)) to further facilitate sleep analysis.

$$\begin{aligned} PM_j = {} & e^{-0.25}EM_{j-9} + e^{-0.5}EM_{j-8} + e^{-1}EM_{j-7} + e^{-0.25}EM_{j-6} + e^{-0.5}EM_{j-5} + e^{-1}EM_{j-4} + e^{-0.25}EM_{j-3} + \\ & e^{-0.5}EM_{j-2} + e^{-1}EM_{j-1} + e^{0}EM_j + e^{-1}EM_{j+1} + e^{-0.5}EM_{j+2} + e^{-0.25}EM_{j+3} + e^{-1}EM_{j+4} + e^{-0.5}EM_{j+5} + \\ & e^{-0.25}EM_{j+6} + e^{-1}EM_{j+7} + e^{-0.5}EM_{j+8} + e^{-0.25}EM_{j+9}, \ \ j = 10, \ 11, \ 12, ..., \ L-9 \end{aligned} \tag{3}$$

The data preprocessing scheme is illustrated in Figure 3.

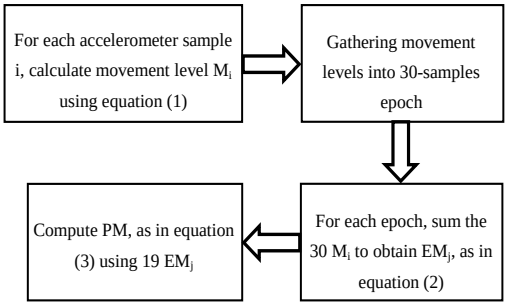

Figure 3: Overnight data preprocessing scheme

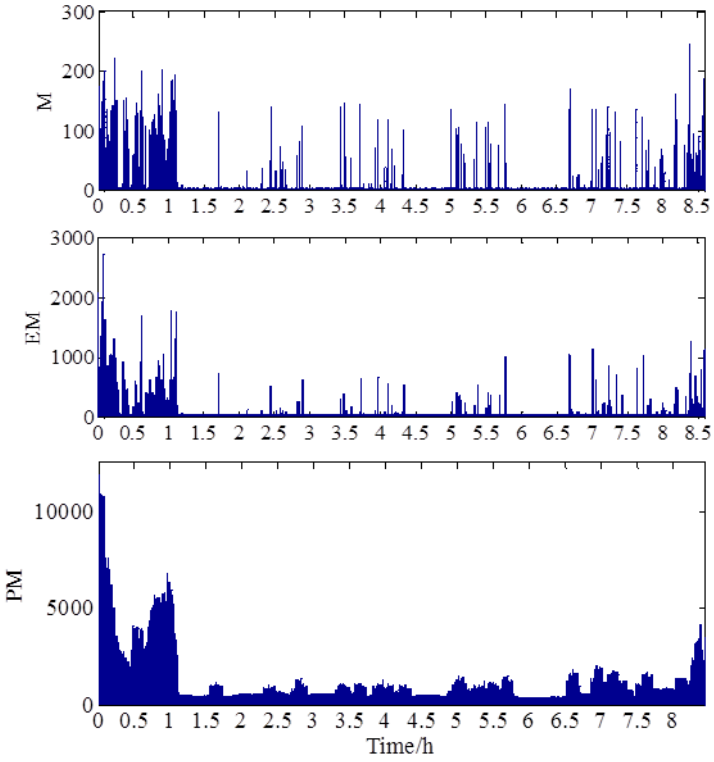

Figure 4: Illustration of M, EM, PM for a same night

As shown in Figure 4, for the overnight data, M is very scattered which further complicates the real sleep analysis. On contrary, PM has a more orderly data evolution which is helpful for the further sleep stages classification. Based on the calculated PM, we have implemented two methods for classifying sleep stages: the threshold method and the k-means method. The threshold method is the simplest and most direct approach. Initially, it was used as an attempt to determine performance. After testing, it was found to be good, so we decided to adopt it. The K-means clustering is also a very common and traditional clustering method. This method is easy to implement and its working mechanism is easy to understand. A better understanding of this clustering method allows us to know how to use it, for example, to choose the parameters and how to organize the clustering framework for better results. The threshold method and the k-means clustering method are described in detail in the following sections.

3 Threshold method

3.1 Sleep and Awake discrimination

Wrist movement can be considered an indicator of wakefulness (Christine & Monique, 2006). The amount of wrist movement can therefore be a sign of sleep or awake state. We define $T_{S/W}$ as a threshold (Figure 5(a) which shows an overnight PM evolution) for discriminating ‘Awake’ and ‘Sleep’ epochs, which is 1350 determined from experimental observation and

testing. When the PM value of an epoch is greater than $T_{S/W}$, the epoch is classified as 'Awake'. Otherwise, it is classified as 'Sleep' and the discrimination process continues to refine the classification.

3.2 Deep sleep and Light sleep/REM discrimination

A lower movement level corresponds to a deeper sleep state (Pollak et al., 2001). It is possible to define a threshold PM value or standard deviation of several continuous PM values to discriminate light sleep from deep sleep. REM sleep is characterized by an activated brain in a paralyzed body, but muscle twitches often accompany REM sleep (Carskadon & Dement, 2005). So we can suppose that the overall movement level during REM is very low, but the standard deviation of movement level could be relatively high because of the muscle twitches. Based on the above analysis, we think that deep sleep is characterized by the lowest standard deviation of movement level, which could be used as a feature to distinguish it from light sleep and REM. Hence, it is possible to define a threshold of standard deviation of several continuous PM values to distinguish deep sleep from light sleep and REM. For epochs first classified as 'Sleep', 6-epochs groups G are formed (representing 3-minutes data). For each G, the standard deviation (SD) of PM values is calculated. If SD is less than a threshold $T_{D/LR}$ (as illustrated in Figure 5(b)), epochs in G are classified as 'Deep sleep'. The value of $T_{D/LR}$ is 10 derived from testing, observation and correction.

3.3 Light sleep/REM discrimination

Light sleep and REM sleep are characterized by relatively high and low movement levels respectively. So, a threshold on PM value can be used to discriminate them. After the two previous steps, remaining epochs noted as H can be classified as 'Light sleep' or 'REM'. To discriminate these two stages, a 500 value threshold $T_{L/R}$ is defined (as illustrated in Figure 5(c)), derived from tests, observations and threshold adjustments. When the PM of H is greater than $T_{L/R}$, it will be classified as 'Light sleep' otherwise as 'REM'.

The overall procedure of classification is described in Figure 6.

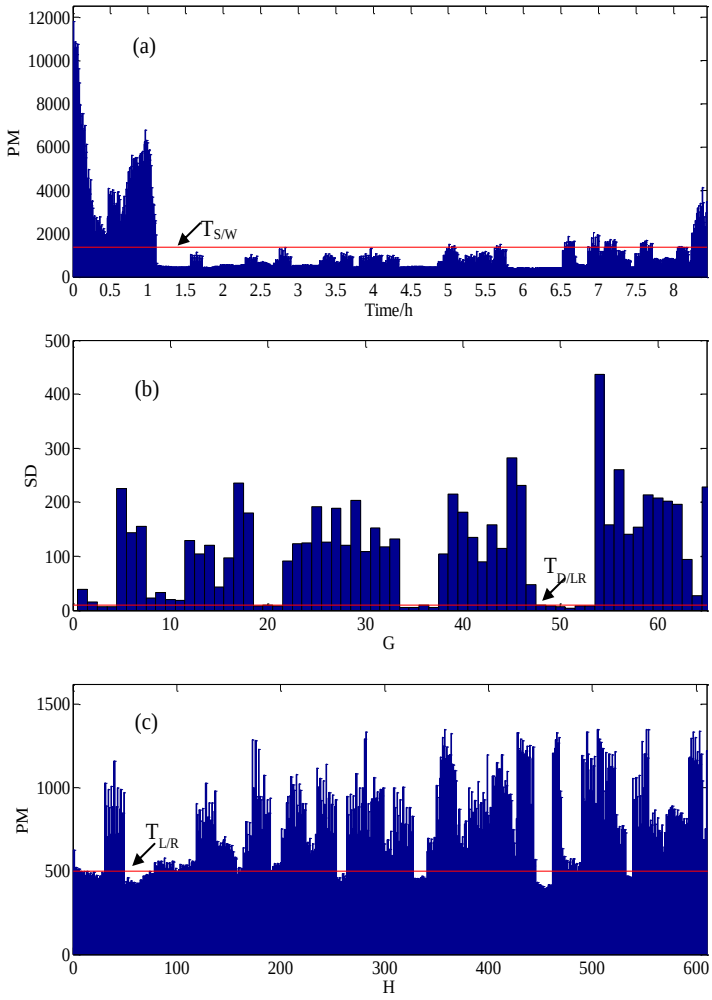

Figure 5: Illustration of three thresholds in sleep stages discrimination

a) Sleep/Awake threshold b) Deep sleep threshold c) Light sleep/REM threshold

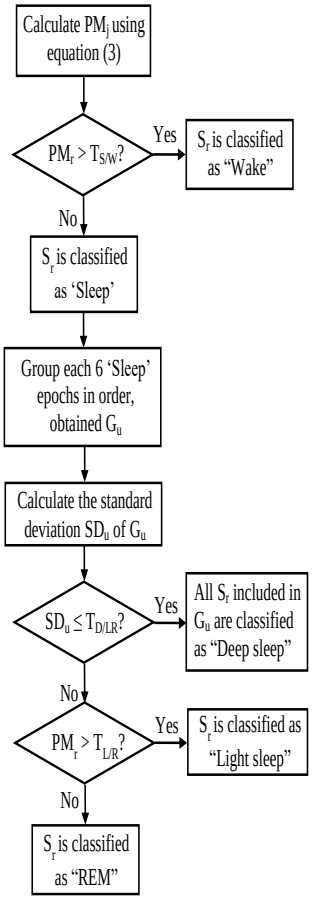

Figure 6: Procedure of sleep stages discrimination

3.4 Detection of falling asleep and waking up

Based on the "sleep" and "awake" detection implemented by the threshold method described above, we have defined the falling asleep point and the waking up point to indicate the beginning and the end of an overnight sleep respectively. Once monitoring begins, if "sleep" state lasts at least 5 minutes, the first point of the 5 minutes will be considered as the starting point of falling asleep, noted as asleep point. Starting from the end of recording and doing backward checking, if the "sleep" state lasts at least 5 minutes, it will be considered as the last epoch of "sleep". So the next epoch is considered as the starting point of the awakening, noted as awakening point. The epochs between asleep point and awakening point are defined as sleep segment.

3.5 Optimization processing

After obtaining the result of the sleep stages classification using the procedure shown in Figure 6, some steps are necessary to optimize results:

1) Modify all epochs before falling asleep point to be "awake".
2) Modify all epochs after awakening point to be "awake".

3) When ‘light sleep’ does not last more than 1 minute and there is an “awake” state before and after, set this ‘light sleep’ period so that it is classified as ‘awake’.
4) When “REM” lasts less than 1 minute and there is a ‘light sleep’ before and after, set this “REM” period as a ‘light sleep’ one.

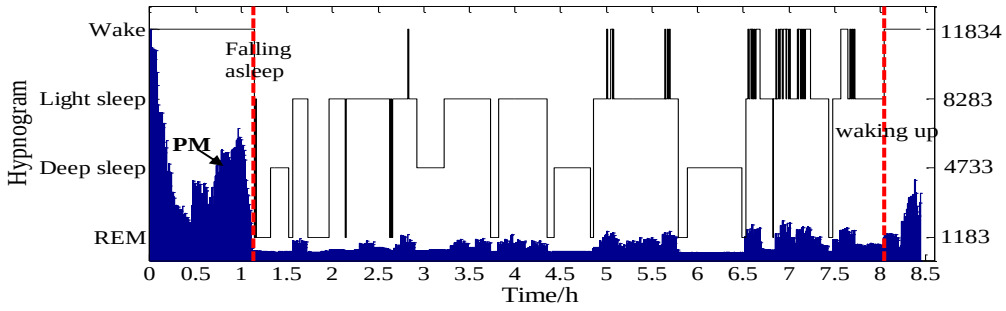

Figure 7: Result of the “Threshold method” for an overnight sleep

Figure 7 shows a result of the “Threshold method” including detection of time of falling asleep, waking up and hypnogram with corresponding PM evolution.

The “Threshold method” uses three thresholds which are all absolute values to classify sleep stages. It means that the same thresholds will be applied to different people. This is difficult to make this a universally applicable method because people's movements during sleep are different, the amplitude and frequency of movements have individual differences that can be caused by factors such as body height and weight, gender, physical condition, age, etc. In order to develop a universal method suitable for different people, we plan to test k-means clustering that makes this possible.

4 K-means method

4.1 K-means clustering

As a classic machine learning method, k-means clustering (MacQueen, 1967) has been commonly used in fields as varied as image segmentation, data compression, wireless sensor network routing, data mining, etc. It is an effective method to automatically classify dataset into k-groups based on the similarity of features of each data group. First, it randomly selects k initial cluster centers C_i and then iteratively performs the following steps:

1. Assign each sample s_i to its nearest clustering center;
2. Update each C_i clustering center with the mean of samples currently in the cluster.

The algorithm converges when the assignment of samples to clusters no longer change. For the k-means clustering algorithm, the selection of initial cluster centers could significantly affect the final clustering result. As the initial clustering centers are randomly selected, the clustering result also has some uncertainty. During the experiments, we found that the final

clustering results using randomly selected cluster centers did not generally change much, but in a few cases the final clustering results were very far from the others. In order to prevent that these rare cases become the final clustering result, we repeat the same clustering procedure several time, and then determine the final clustering result by voting, as described in the "Voting rule" section.

The k-means method is applied to sleep epochs to obtain a hypnogram which contains "Awake", "Light sleep", "Deep sleep" and "REM". The sleep epochs begin from the time where we fall asleep until we wake up, which is detected by the threshold method described in section 3. As far as we know, there are several works (Diykh et al., 2016; Van & Philips, 2001; Güneş et al., 2010) that adopt k-means method to classify sleep stages using the EEG signal, but no one uses the wrist movement signal.

4.2 Feature extraction

A 2-dimension feature based on *PM* is used for k-means clustering. We directly use *PM* as the first dimension of the feature. All PMs are grouped sequentially, and each group contains six *PM* values. The standard deviation of the PM values in each group is used as the second dimension of feature for the corresponding six epochs in the group. In other words, the second dimension of feature of the six epochs in one group is the same, which is the standard deviation of their corresponding PM values.

4.3 Sleep stages clustering

The overall procedure of this "k-means" method includes 5 iterations of k-means clustering with k=2, noted as 5km2.

We have also tried to classify the four sleep stages directly using only one iteration of k-means clustering with k=4, noted as 1km4.

We have adopted "Fitbit charge 2^{TM}" as the reference device to evaluate the results of the "Threshold", "5km2" and "1km4" methods. The "Fitbit charge 2^{TM}" is a commercial device that has been compared with the PSG (polysomnography) gold standard and validated as promising for sleep stages and sleep-wake detection (Zambotti et al., 2018). "Fitbit Charge 2^{TM}" showed a sensitivity of 0.96 (accuracy to detect sleep), a specificity of 0.61 (accuracy to detect wake), an accuracy of 0.81 in detecting N1+N2 sleep ("light sleep"), an accuracy of 0.49 in detecting N3 sleep ("deep sleep"), and an accuracy of 0.74 in detecting rapid-eye-movement (REM) sleep (Zambotti et al., 2018). The classification results of the Fitbit, threshold method, 5km2 and 1km4 methods will be presented in section 5.3, and the hypnograms and pie chart for the proportion of each two-night sleep stage obtained by each method are presented in Figures 7 and 8 respectively.

According to the study (Carskadon & Dement, 2005), for normal young adults who live on a conventional sleep-wake schedule and without sleep disorders:

- Waking up during sleep usually represents less than 5% of the night.
- Light sleep generally accounts for about 47% to 60% of sleep.
- Deep sleep generally constitutes about 13% to 23% of sleep.
- REM sleep usually accounts for 20% to 25% of sleep.

It has been found that the proportion of light sleep should be much higher than that of deep sleep. However, given the experimental results, the 1km4 method still obtains too much deep sleep time and not enough light sleep time, which is contradictory with the results of the study (Carskadon & Dement, 2005) obtained by the PSG method. The 5km2 and threshold methods have comparable results to those of the study (Carskadon & Dement, 2005). We will therefore present the 5km2 method in detail.

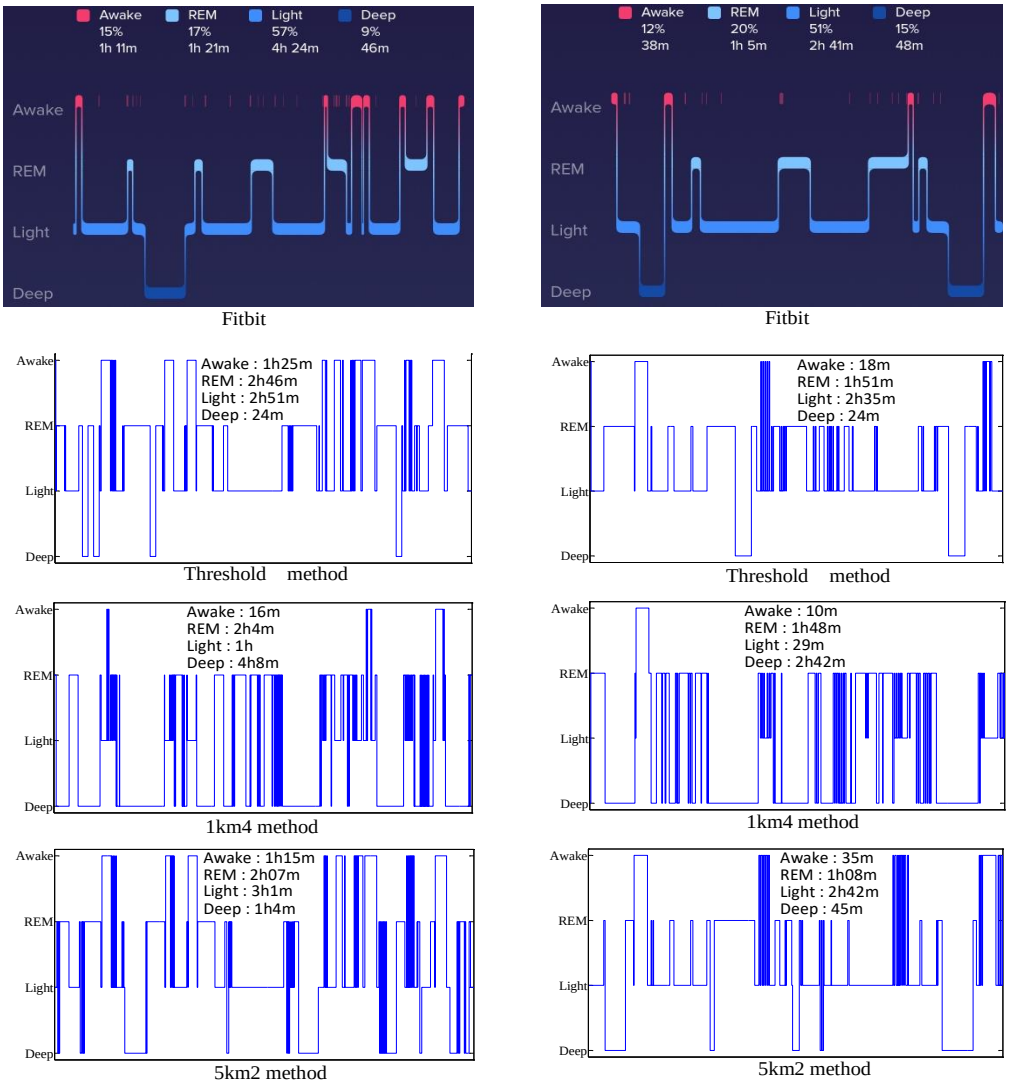

Figure 8: Sleep stages classification result of four methods for two nights

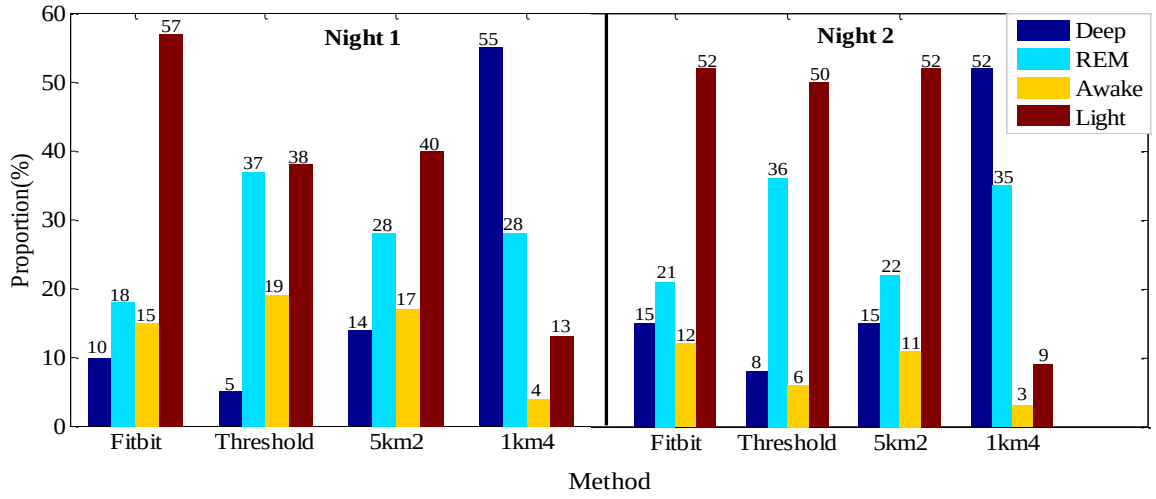

Figure 9: Proportion of each sleep stages derived from four methods for two nights

4.3.1 Awake

After a k-means clustering (k=2) on sleep segment defined by the "Threshold" method, the cluster with the highest mean value of *PM* is classified as "awake". The other cluster is S1.

4.3.2 Light sleep

The "Light sleep" comes from 2 sources. Firstly, after a second k-means clustering on the previous cluster S1, the cluster with the highest mean value of *PM* is classified as "Light sleep". The other cluster is noted S2. Then, a third k-means clustering is performed on S2, the cluster with the highest mean value of *PM* is defined as quasi-REM noted S3, the cluster with the lowest mean value of PM is defined as quasi-Deep sleep noted S4. Finally a fourth k-means clustering is performed on S3, and the cluster with the highest mean value of PM is also classified as "Light sleep". In summary, "light sleep" corresponds to cluster S2 and the last mentioned cluster.

4.3.3 Deep sleep

On the quasi-Deep sleep noted S4, a new k-means clustering is carried out. The cluster with the lowest mean value of *PM* is classified as "Deep sleep".

4.3.4 REM

The "REM" also comes from 2 sources. On the one hand, after a k-means clustering for quasi-REM S3, the cluster with the lowest mean value of *PM* is classified as "REM". On the other hand, with the k-means clustering performed over quasi-Deep sleep S4, the cluster with the highest mean value of *PM* is also classified as REM.

The procedure of sleep stages clustering is illustrated in Figure 10.

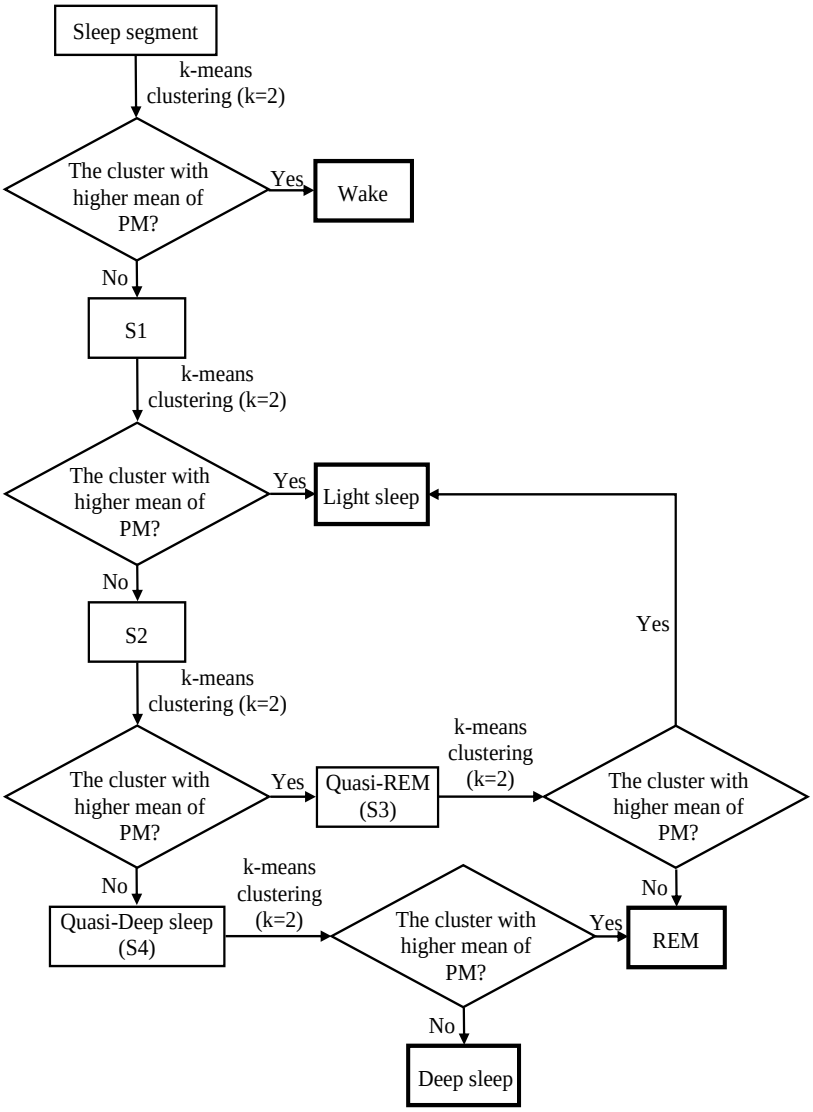

Figure 10: Procedure of sleep stages clustering

4.4 Voting rule

For the k-means clustering, the primary clustering centers are randomly selected, which gives randomness to the final result. In our experience, most of the time the final results will be the same or very close whatever the randomly selected primary clustering center is. However, in a few extreme cases, the final sleep stages distribution will be very different from the one reported in (Carskadon & Dement, 2005). To eliminate these extreme cases, a voting rule has been designed.

First of all, we performed the above-mentioned sleep stages clustering procedure ten times. Thus, for each epoch, we obtained 10 clustering results. Then, for the 10 clustering results, the classification to which the epoch finally belongs is determined by a majority vote.

5 Experiment results

5.1 Computational complexity and processing time

For an algorithm, too high a computational complexity will increase the time required to execute the algorithm which will affect the real-time performance, while also increasing the cost of the hardware platform. The computational complexity of all the proposed methods is relatively low. The computational complexity of the “Threshold” method is of constant order, i.e., O(1), and the computational complexity of both the “1km4” and “5km2” methods is of linear order, i.e., O(n). In addition to analyzing the theoretical computational complexity of the proposed methods to demonstrate their efficiency, we also experimentally measured the execution time required for these methods.

In the experiment, the “Threshold”, “1km4” and “5km2” algorithms are all implemented on the same computer with “Intel i7-2600 CPU @ 3.40GHz, 8GB RAM” on “MATLAB R2011b”. For 8-hour night-time data processing, the time spent on the "Threshold method", "1km4 method" and "5km2 method" is 1.04s, 1.73s and 1.84s respectively. We can see that the processing time for all the algorithms is very short, less than 2 seconds, and the hardware platform required is common. This indicates the proposed methods can be easily used in a real time application, giving fast results.

5.2 Falling asleep/waking up detection analysis

We recruited 5 young adults (3 women, 2 men) as test subjects. The physical factors of the subjects are presented in Table 1. The Mean±STD of the 5 subjects’ age, weight and height is 29.8±2.2 years, 56.4±7.9 kg and 167.8±9.9 cm. A total of 15 overnights sleep are tested by four sleep stage classification methods, namely “Fitbit”, “Threshold (the method presented in section 5 which uses 3 thresholds)”, “5km2” and “1km4”. The “Threshold”, “5km2” and “1km4” methods are all implemented solely based on wrist movement data. Considering all nights, we collected 30 results of falling asleep/waking up detection. The difference between Fitbit and the proposed “Threshold method” in detected time of falling asleep and waking up are illustrated in Table 2.

Table 1, Physical factors of subjects

Subjects	Gender	Age	Weight (kg)	Height (cm)	Physical condition
1	Male	31	65	176	Normal
2	Male	28	60	180	Normal
3	Female	27	45	156	Normal
4	Female	31	60	164	Normal
5	Female	32	52	163	Normal

Table 2: Comparison between the Fitbit method and the "Threshold method" for the detection of the time of falling asleep and the time of waking up

Subjects	Nights index	Methods	Time of falling asleep	Time of waking up	Difference in time of falling asleep/min	Difference in time of waking up/min
1(male)	1	Fitbit	02:34	09:22	0	5
		Threshold	02:34	09:17		
	2	Fitbit	00:21	08:19	4	5
		Threshold	00:25	08:14		
	3	Fitbit	02:06	09:32	1	0
		Threshold	02:05	09:32		
	4	Fitbit	00:10	07:51	1	7
		Threshold	00:11	07:44		
2(female)	5	Fitbit	23:50	07:19	8	0
		Threshold	23:58	07:19		
	6	Fitbit	03:09	09:29	2	0
		Threshold	03:07	09:29		
	7	Fitbit	03:09	09:29	2	0
		Threshold	03:07	09:29		
3(female)	8	Fitbit	02:52	12:50	3	0
		Threshold	02:55	12:50		
	9	Fitbit	03:06	09:30	0	4
		Threshold	03:06	09:26		
4(male)	10	Fitbit	02:53	08:05	4	0
		Threshold	02:57	08:05		
	11	Fitbit	00:49	08:00	3	3
		Threshold	00:52	07:47		
	12	Fitbit	01:21	08:09	1	0
		Threshold	01:22	08:09		
	13	Fitbit	01:50	08:07	3	12
		Threshold	01:53	07:55		
5(female)	14	Fitbit	23:33	08:07	14	5
		Threshold	23:47	08:02		
	15	Fitbit	23:57	09:01	2	49
		Threshold	23:59	08:12		
	16	Fitbit	-	-	-	-
		Threshold	22:21	08:12		

Table 3 shows that 25 out of 30 results have a time difference of not exceed 5 minutes. For night 14, subject 5 reported going to bed around 11.30pm, then watching his smartphone for 10 minutes and then falling asleep. The time of falling asleep detected by the proposed "Threshold method" is therefore more accurate than Fitbit's for that night. For night 15, the difference in waking time is 49 minutes. However, subject 5 reported waking up around 8:00am that morning. Therefore, the 09:01am wake-up time determined by Fitbit is clearly incorrect. Subject 5 agrees with the wake-up time determined by waking up time of the "Threshold method" proposed for night 15.

Table 3: Number of results of the sleep and wake-up detection in different time difference ranges

Time difference	≤ 5 min	> 5 min, ≤ 10 min	> 10 min,≤ 15 min	> 15 min
Number of results	25	2	2	1

5.3 Sleep stages classification analysis

The results of the sleep stage classification are presented in Table 4. They are compared with users' self-reported feedbacks.

Table 4: Comparison of the four methods in sleep stages classification

Subject	Night	Method	Awake	Light sleep	Deep sleep	REM	Sleep score	Declarative feedback from the subject on his sleep
1(male)	1	Fitbit	71	192	62	86	75.0	Very poor sleep, awake sleep many times
		Threshold	108.5	163.5	18	112	50.0	
		5km2	30	158.5	117.5	97	83.4	
		1km4	14	20	262.5	106.5	73.1	
	2	Fitbit	67	247	74	107	79.5	Very tired before sleep, sleep much better than the first night, less sleep awake
		Threshold	66.5	154	66	181	62.9	
		5km2	29.5	134.5	117	187.5	73.7	
		1km4	17.5	32	321.5	97.5	77.3	
	3	Fitbit	59	264	46	81	73.2	Normal sleep
		Threshold	85	169	51	142	63.2	
		5km2	75	181	64	127	71.5	
		1km4	15.5	59.5	248	124	78.4	
2(female)	4	Fitbit	53	184	112	111	84.3	Very light sleep with distinct awake sleep
		Threshold	12.5	175.5	153	112	91.1	
		5km2	19.5	210	66	175.5	77.5	
		1km4	8.5	16	299.5	147	73.1	

		Fitbit	51	262	54	82	77.5	
	5	Threshold	41	265	54	81	79.6	Sleep better than last night (night 4)
		5km2	116.5	198	54	73.5	63.4	
		1km4	25.5	83.5	180	153	74.2	
		Fitbit	60	236	55	79	76.6	
	6	Threshold	29.5	95.5	132	151.5	71.4	Normal sleep with sleep awake
		5km2	23	259.5	35	112	77.4	
		1km4	19	59.5	222.5	128.5	76.5	
		Fitbit	15	216	59	90	79.2	
	7	Threshold	3	95.5	132	151.5	70.5	Very poor sleep, with distinct awake sleep
		5km2	61.5	146.5	6	169	41.8	
		1km4	18.5	44	240	80.5	71.5	
		Fitbit	56	373	61	115	62.4	
3(female)	8	Threshold	22.5	287	48	237	55.4	Sleep much better than last night (night 7)
		5km2	10.5	315	23.5	246.5	48.3	
		1km4	8.5	23.5	360.5	203	58.7	
		Fitbit	25	229	31	100	70.3	
	9	Threshold	11	102	42	225	49.8	Normal sleep
		5km2	10	197	35	139	67.7	
		1km4	9.5	32.5	203	136	65.7	
		Fitbit	38	161	48	65	60.0	
	10	Threshold	18	155.5	45	89	60.6	Normal sleep
		5km2	34.5	161.5	44.5	67.5	59.4	
		1km4	9.5	29	162	107.5	53.3	
		Fitbit	45	199	99	87	87.3	
	11	Threshold	30.5	243	90	51.5	83.2	Normal sleep
		5km2	136	152.5	94	33.5	57.6	
		1km4	33	76.5	192	114.5	76.3	
4(male)		Fitbit	59	195	82	79	80.7	
	12	Threshold	32	317.5	30	28	58.5	Very poor sleep, experience an unpleasant thing before sleep
		5km2	121	249.5	30	8	45.1	
		1km4	6.5	75.5	169.5	157	73.2	
		Fitbit	79	156	92	49	65.0	
	13	Threshold	23.5	279.5	36	22.5	55.6	Very poor sleep, feel very anxious before sleep which affecting the sleep
		5km2	126.5	127.5	6.5	102	34.6	
		1km4	45	79	150	88.5	65.8	
		Fitbit	43	272	33	166	66.6	
	14	Threshold	16	280.5	66	132	86.6	Good sleep, left bed at about 6:30 then returning to bed continue to sleep
5(female)		5km2	10.5	270	136.5	78.5	93.3	
		1km4	7.5	6	334.5	147.5	73.1	
	15	Fitbit	39	379	35	91	64.8	Poor sleep, many dreams

	Threshold	7.5	393	42	51	66.3	and awake during this sleep.
	5km2	135.5	189.5	37.5	132	51.5	Get up at nearly 8:15
	1km4	31	96.5	242	125	80.1	
	Fitbit	-	-	-	-		
16	Threshold	14.5	381.5	42	153	65.8	Normal
	5km2	141	341	8.5	101.5	31.9	
	1km4	45	111	139.5	296.5	46.3	

(In this table, the unit of the number representing the duration of the sleep stages is the minute).

Considering all nights, 10 nights show better results with the 5km2 method, 4 nights show comparable performance between the 5km2 method, “Fitbit” and “Threshold” methods, and 2 nights show lower performance for the 5km2 method compared to the “Fitbit” and “Threshold” methods. The test results for nights 2, 4, 7, 8 and 11 - 16 show that the “5km2” method appears to have superior performance in sleep stages classification.

On night 2, the volunteer reported having slept well. Comparing the results of the “Fitbit”, “Threshold” and “5km2” methods, the 5km2 has the least “Wake” and “Light sleep” and the most “Deep sleep” and “REM”, which is consistent with the subject's feedbacks.

During the fourth night, the volunteer felt that he slept poorly, with very light sleep and a distinctly awake sleep. It can be seen that the k-means method finds more “Light sleep” and less “Deep sleep” than the other two methods, which is more indicative of the subject's true state of sleep.

During night 7, the volunteer felt that he had a very little sleep, which is associated with distinct awake sleep. We note that the results of the 5km2 k-means method show a much higher “Wake” and a much lower “Deep sleep”. Compared to the “Fitbit” and “Threshold” methods, the 5km2 method better highlights sleep problems according to the subject's feedback.

During night 8, the volunteer mentioned better sleep compared to the previous night (night 7). Compared to the result of night 7, the results of the k-mean method show a significant decrease in “Wake” and an increase in “Deep sleep” on night 8. However, the other two methods even show a significant increase in “Wake” and a near or significant decrease in “Deep sleep”, which may not indicate an improvement in sleep quality.

The test results of nights 1 and 5 show that the k-means method is less effective in classifying sleep stages.

On night 1, the subject reports poor sleep and repeated awake sleep, but the k-means method gives the least “Wake” and the most “Deep sleep”, which is contrary to the actual sleep state.

On night 5, the subject sleeps better sleep than the previous night (night 4). However, the k-means method shows a dramatic increase in “Wake” and a slight decrease in “Deep sleep”, which is also contrary to the actual sleep state.

Nights 3, 6, 9 and 10 are considered by the subjects as normal sleeps. The results of the k-means method are comparable to those of two other methods for these nights.

On nights 12 and 13, subject 4 reports very poor sleep. For the “5km2” and “Threshold” results, we can see the significant decrease in deep sleep time on nights 12 and 13 compared to nights 10 and 11 which are considered normal sleep by subject 4. However, for the of “Fitbit” results, the deep sleep time even increase significantly on nights 12 and 13 compared to night 10.

As shown in Table 4, the deep sleep times obtained by “Fitbit”, “Threshold” and “5km2” increased, decreased and decreased respectively between nights 14 and 15. A study (Brand et al., 2014) showed that individuals are less awake after the onset of sleep and that people who sleep more deeply report less daytime sleepiness. It can therefore be assumed that being awake is negatively correlated with good sleep and that deep sleep is positively correlated with good sleep. According to the feedback of sleepers, the sleep of night 14 is good and the sleep of night 15 is bad. Therefore, the decrease in the duration of deep sleep from night 14 to night 15 may better reflect the real change in sleep quality between these two nights. However, in order to evaluate this assertion, additional experiments need to be conducted over several nights.

5.4 Sleep score

After obtaining hypnogram we can obtain the duration of each sleep stages, which is closely related to the quality of sleep. It is therefore possible to assess sleep quality by defining a sleep score based on the hypnogram, which helps users without relevant sleep knowledge to intuitively understand the results of their sleep monitoring.

For healthy sleep, the total sleep duration and the proportion of each sleep stage should be within a reasonable range. The appropriate sleep duration (Hirshkowitz et al., 2015) for individuals of different generation is shown in Table 5.

Table 5: Appropriate sleep duration for each generation

Generation	Appropriate sleep duration
newborns	14 ~ 17 h
infants	12 ~ 15 h
toddlers	11 ~ 14 h
preschoolers	10 ~ 13 h
school-aged children	9 ~ 11 h
teenagers	8 ~ 10 h
young adults and adults	7 ~ 9 h
older adults	7 ~ 8 h

In this paper, all volunteers belong to the young adult and adult generation. The normal proportion of each sleep stage for individuals of this generation without sleep complaints is shown in Table 6 (Carskadon & Dement, 2005).

Table 6: Proportion of normal sleep stages for young adults and adults

Sleep stage	Normal proportion
Awake	< 5%
Light sleep	47% ~ 60%
Deep sleep	13% ~ 23%
REM	20% ~ 25%

The definition of the symbols is presented in Table 7. These symbols are used in the flowchart of the sleep score calculation.

Table 7: Definition of symbols

	Awake	Light sleep	Deep sleep	REM
Duration	D_W	D_L	D_D	D_R
Lower limit of normal proportion	P_{WL}	P_{LL}	P_{DL}	P_{RL}
Upper limit of normal proportion	P_{WU}	P_{LU}	P_{DU}	P_{RU}
Total sleep duration		T		
Lower limit of appropriate sleep duration		T_L		
Upper limit of appropriate sleep duration		T_U		
Sleep score		S		

The steps for calculating the sleep score are shown in Figure 11. The sleep score is calculated on the basis of the total sleep duration and the duration of each sleep stage. Depending on the normal range given in Tables 5 and 6, any parameter outside the range will result in a lower sleep score. Besides, within the normal range, a deeper and less awake sleep will result in a higher sleep score. After obtaining the sleep score, we rescale it to a range of 0 ~ 100, with the higher score meaning better sleep. The rescaling method is the last step in the diagram in Figure 10. Use the result of the 5km2 of night 1 in Table 4 as an example to calculate the sleep score. The duration of awake, light sleep, deep sleep and REM is 30, 158.5, 117.5 and 97 minutes. Thus, D_D is 117.5, D_W is 30. According to Table 6, P_{DL} is 0.13, P_{DU} is 0.23, P_{WL} is 0, P_{WU} is 0.05. According to the first step described in Figure 10, we can obtain the primary sleep score S_p from equation (4).

$$S_p = D_D \times (1 - \frac{P_{DL} + P_{DU}}{2}) - D_W \times (1 - \frac{P_{WL} + P_{WU}}{2}) \quad (4)$$

By introducing the value into the equation, we can obtain S_p=117.5×(1-(0.13+0.23)/2)-30×(1-(0+0.05)/2)=67.1. Then we check if the proportion of deep sleep, light sleep and REM are in the normal range or not.

The proportion of deep sleep is 29.2%. According to Table 6, the proportion of deep sleep is too high. This results in a reduction of the score by equation (5).

$$S_D = S_p - (D_D - T \times P_{DU}) \times (1 - \frac{P_{DL} + P_{DU}}{2}) \quad (5)$$

Where Sp has been calculated in the previous step, the value is 67.1. T is the total sleep duration which is the sum of each sleep stage duration, i.e. the sum of 30, 158.5, 117.5 and 97, i.e. 403. By introducing this value into the equation we can get SD=67.1-(117.5-403×0.23) ×(1-(0.13+0.23)/2)=46.76.

The proportion of light sleep is 39.3%. According to Table 6, the proportion of light sleep is too low. This results in a reduction of the score by equation (6).

$$S_{DL} = S_D - (T \times P_{LL} - D_L) \times (1 - \frac{P_{LL} + P_{LU}}{2}) \quad (6)$$

Where SD has been calculated in the previous step, the value is 46.76. According to Table 6, PLL is 0.47, PLU is 0.6. DL is 158.5. By introducing the value into the equation we can obtain SDL=46.76-(403×0.47-158.5) ×(1-(0.47+0.6)/2)=32.4.

The proportion of REM is 24.1%. According to Table 6, the proportion of REM is within the normal range. The score will not change at this step, namely:

$$S_{DLR} = S_{DL} \quad (7)$$

Finally, we check the total sleep duration is within the normal range or not. The total sleep duration T=403 minutes (6.7h) is not in the normal range for young adults and adults, which should be 7~9h according to Table 6. This will result in a reduction of the score by equation (8).

$$S_{DLRT} = S_{DLR} - (T_L - T) \quad (8)$$

Where SDLR has been calculated in the previous step, the value is -69.1. According to Table 6, TL=420 minutes (7h). By introducing the values to the equation SDLRT=32.4-(420-403)=15.4 can be obtained. The SDLRT is the raw sleep score, we rescale the raw sleep score in the range of 0~100 to obtain the final sleep score S. The rescaling is carried out by equation (9).

$$S = \frac{S_{DLRT} - S_{\min}}{S_{\max} - S_{\min}} \times 100 \quad (9)$$

Where S_{min} means the raw sleep score of a bad sleep, S_{max} means the raw sleep score of a very good sleep. We have set the sleep with a duration of five minutes and the five minutes are all light sleep as the worst sleep. For the worst sleep, the duration of awake, light sleep, deep sleep and REM is 0, 5, 0 and 0. We can then calculate the corresponding raw sleep score S_{min}= -417.2. And we set the sleep with the duration of upper limit, the deep sleep proportion is also upper limit, no awake and both the light sleep and REM are in normal range as the best sleep. For the best sleep, the duration of awake, light sleep, deep sleep and REM is 0, 280.8,

124.2 and 135. We can then calculate the corresponding raw sleep score S_{max}= 101.8. According to the equation, we can obtain the final sleep score S=(15.4-(-417.2))/(101.8-(-417.2))×100=83.4. This is the whole procedure of sleep score calculation with the given duration of each sleep stage.

We are trying to find a lower limit of sleep score for a good sleep. Here we define a lower limit for good sleep as sleep where the lower limit of the appreciate duration, the lower limit of the normal deep sleep proportion, the upper limit of the normal awake proportion, and light sleep, REM are both within the normal range. For the lower limit of good sleep, the duration of awake, light sleep, deep sleep and REM is 21, 252, 54.6 and 92.4 minutes. The corresponding sleep score is defined as the lower limit of the sleep score for good sleep also as the basis for good sleep, which is 85.1. The sleep scores calculated for all volunteers on the basis of the hypnogram given by four methods are presented in Table 4. As can be seen, for the total 16 test nights, only the sleep score of night 4 with the threshold method, night 11 with the Fitbit method and night 14 with the threshold and 5km2 methods is above the lower limit of the sleep score for good sleep. Thus, according to the method proposed for calculating the sleep score and the good sleep baseline, the rate of good sleep with the Fitbit method is 6.67% (1/15); the rate of good sleep with the Threshold method is 13.3% (2/15); the rate of good sleep with the 5km2 method is 6.67% (1/15). It should be pointed out that the 5 volunteers for the tests are all PhD students. One study showed that only 11.5% of the students surveyed met the criteria for good sleep quality (Buboltz et al., 2009). Thus, the relatively low rate of good sleep obtained by the methods we propose can be considered a reasonable result.

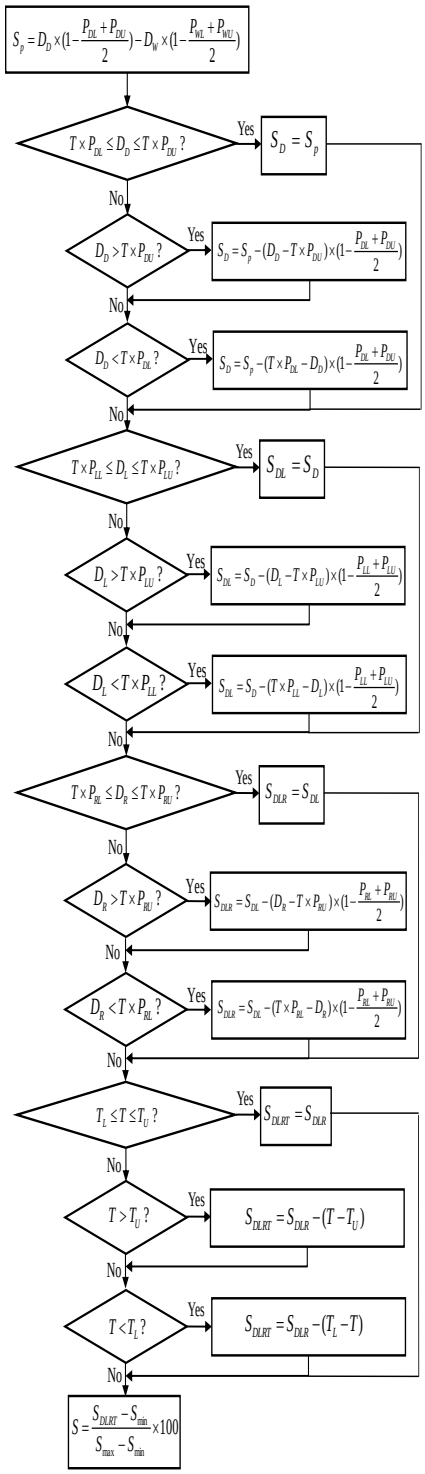

Figure 11: Procedure for calculating the sleep score

6 Performance evaluation with the PSG gold standard

6.1 materials and methods

The PSG is the gold standard for sleep monitoring, It is widely used to evaluate the performance of new sleep monitoring devices. In this study, we also use the PSG to test the two proposed methods. We use the sleep monitoring wristband developed by our group and

the PSG to simultaneously monitor the sleep of a volunteer in the sleep laboratory located in the university hospital center of Toulouse in France. The real sleep monitoring environment of this experiment and of the volunteer with the PSG and our wristband on his body is shown in Figure 12. The volunteer recruited is a 28-year-old man with a BMI (body mass index) of 18.3. A one-night test was carried out for a primary evaluation of the performance of the two proposed methods.

Figure 12. The volunteer with PSG and our wristband on body in real environment of this experiment.

6.2 Results

We compare the hypnogram obtained from the PSG and the two methods proposed, epoch by epoch. We compare hypnograms obtained by the threshold method and the 5km2 method with the hypnogram obtained by the PSG as shown in Figure 13.

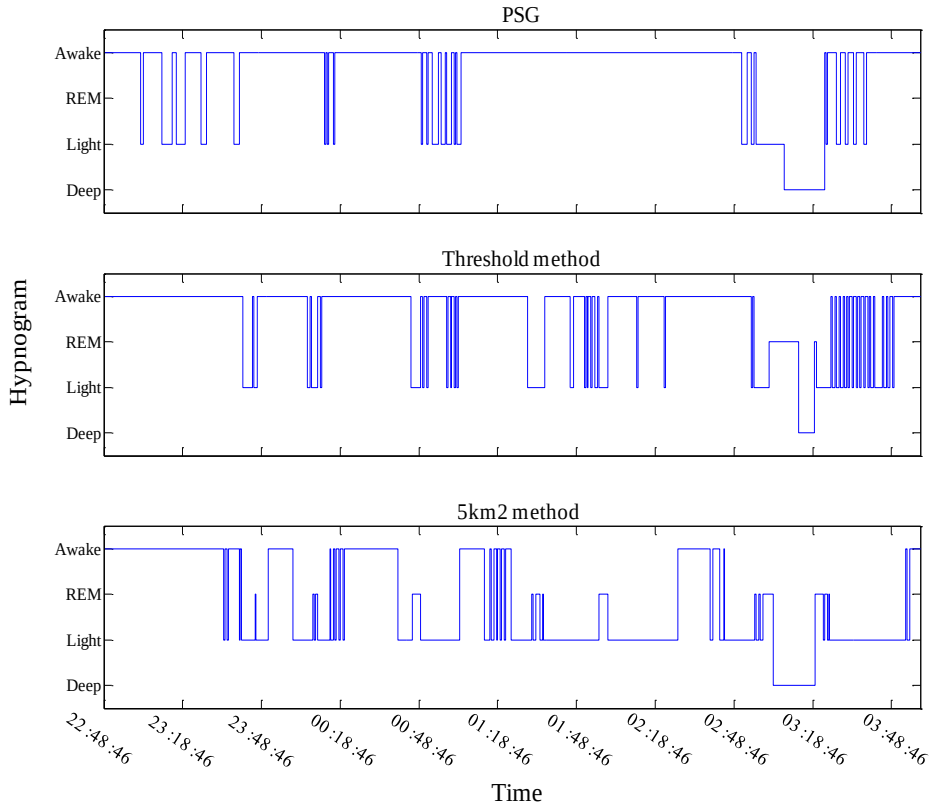

Figure 13. Hypnogram obtained from the PSG, Threshold method and 5km2 method.

Table 8: Cumulative duration (in min) of each sleep stage obtained from the PSG, threshold and 5km2 methods

	Awake	Light sleep	Deep sleep	REM
PSG	**254**	**41.5**	**15.5**	**0**
Threshold method	237	56	6	12
5km2 method	119	155.5	16	20.5

Unit: minute

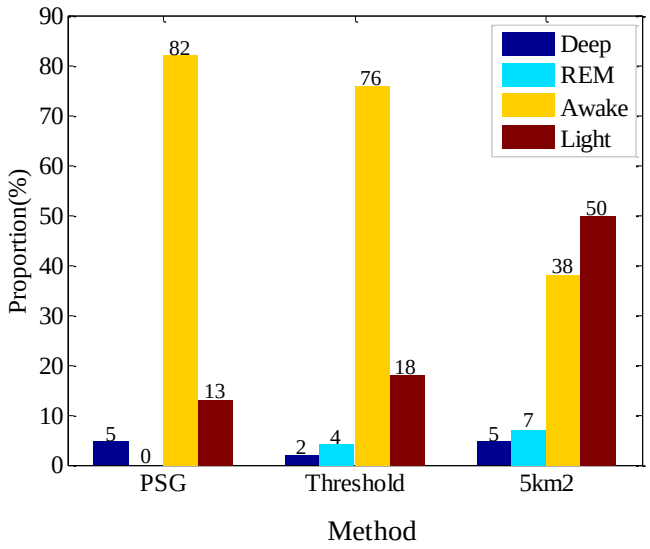

Figure 14. Proportion of each sleep stage obtained from the PSG, threshold and 5km2 methods.

Table 8 shows the duration of each sleep stage obtained by the PSG, the threshold method and the 5km2 method. Figure 14 shows the proportion of each sleep stage obtained from the PSG, the threshold method and the 5km2 method.

By observing the PSG hypnogram in Figure 13, we can see that the sleep of this night consists mainly of awake and light sleep. This result is similar to that of the threshold method and the 5km2 method. These two methods give a hypnogram consisting mainly of awake and light sleep. According to Table 8 and Figure 14, the difference is that the hypnogram of the threshold method contains relatively more awake (76%) which is closer to that detected by the PSG (82%) method, the hypnogram from the 5km2 method contains relatively less awake (38%) but more light sleep (50%). In the PSG hypnogram, the longest duration of deep sleep is located around 03:18:46. The threshold method and the 5km2 method also detected deep sleep at this time, but the proportion of deep sleep detected by the 5km2 method (5%) is higher and the same as that detected by the PSG (5%). In general, the hypnogram obtained by the threshold method and the 5km2 method have a similar profile to that of the PSG, and the 5km2 method is relatively more efficient than the threshold method. The PSG did not detect any REM epochs but both the Threshold method and the 5km2 method detected some REM epochs. However, the cumulative duration of the REM detected by the Threshold method and the 5km2 method is relatively short, being 12 minutes and 20.5 minutes respectively.

		Predicted (Threshold method)			
		Awake	Light	Deep	REM
True (PSG)	Awake	**423**	**85**	**0**	**0**
	Light	**51**	**21**	**0**	**11**
	Deep	**0**	**6**	**12**	**13**
	REM	**0**	**0**	**0**	**0**

		Predicted (5km2 method)			
		Awake	Light	Deep	REM
True (PSG)	Awake	**211**	**272**	**0**	**25**
	Light	**27**	**38**	**8**	**10**
	Deep	**0**	**1**	**24**	**6**
	REM	**0**	**0**	**0**	**0**

Figure 15. Confusion matrix of the two methods proposed

We compare the results of the sleep stage classification of the threshold method and the 5km2 method with the PSG, epoch by epoch. Two confusion matrices are created to show the result, as illustrated in Figure 15. To evaluate the agreement between the two proposed methods and the classification of sleep stages using the PSG method, Cohen's Kappa coefficient (κ) is calculated. According to the guidelines of Landis & Koch (1977), the Threshold method shows a fair agreement with the PSG ($\kappa = 0.24$), the 5km2 method shows a slight agreement with the PSG ($\kappa = 0.09$).

As shown in the confusion matrix between the Threshold method and the PSG, most awake epochs are correctly classified as awake, a small proportion are incorrectly classified as light sleep, no awake epochs are classified as deep sleep or REM. The light sleep epochs are mainly classified as awake or light sleep, but most are wrongly classified as awake. For deep sleep epochs, the classification results are scattered, but most are classified as deep sleep and REM. For the confusion matrix between the 5km2 method and the PSG, most awake epochs are classified as awake and light sleep, no one is wrongly classified as deep sleep but a very small amount is wrongly classified as REM. Most light sleep epochs are classified as awake and light sleep with a small amount classified as deep sleep and REM. Most deep sleep epochs are correctly classified, none are wrongly classified as awake and only one epoch is incorrectly classified as light sleep.

In general, the classification errors of these two methods exist mainly in the confusion between awake and light sleep, and between deep sleep and REM. It should be noted that neither of the two proposed methods involves confusion between deep sleep and awake, and there is only little confusion between deep sleep and light sleep. For physiological significance, deep sleep is very different from awake and light sleep. Therefore, confusion between deep sleep and awake, and confusion between deep sleep and light sleep can be

considered as serious error. Fortunately, both proposed methods have very few errors in this respect.

		True (PSG)	
		Awake	Not awake
Predicted (Threshold method)	Awake	TP=423	FP=51
	Not awake	FN=85	TN=63

		True (PSG)	
		Light	Not Light
Predicted (Threshold method)	Light	TP=21	FP=91
	Not Light	FN=62	TN=448

		True (PSG)	
		REM	Not REM
Predicted (Threshold method)	REM	TP=0	FP=24
	Not REM	FN=0	TN=598

		True (PSG)	
		Deep	Not Deep
Predicted (Threshold method)	Deep	TP=12	FP=0
	Not Deep	FN=19	TN=591

Figure 16. Confusion matrix for the recognition of each sleep stage with the threshold method

		True (PSG)	
		Awake	Not awake
Predicted (5km2 method)	Awake	TP=211	FP=27
	Not awake	FN=297	TN=87

		True (PSG)	
		Light	Not Light
Predicted (5km2 method)	Light	TP=38	FP=273
	Not Light	FN=45	TN=266

		True (PSG)	
		REM	Not REM
Predicted (5km2 method)	REM	TP=0	FP=41
	Not REM	FN=0	TN=581

		True (PSG)	
		Deep	Not Deep
Predicted (5km2 method)	Deep	TP=24	FP=8
	Not Deep	FN=7	TN=583

Figure 17. Confusion matrix for the recognition of each sleep stage with the 5km2 method

The confusion matrices for the recognition of each sleep stage using the Threshold method and the 5km2 method are shown in Figures 16 and 17. Six performance assessment indexes based on the confusion matrix are calculated, including sensitivity, specificity, accuracy, precision, balanced accuracy and F1 score, as presented in Table 9. These indexes assess performance from different perspectives. They all range from 0 to 1, and a higher value means better performance. In order to make an overall assessment of classification performance for the two proposed methods, the values of all the indexes in Table 9 are averaged and the results are presented in Table 10.

Table 9: Assessment indexes for recognition of each sleep stage with threshold and 5km2 methods

Evaluation indexes	Method	Awake	REM	Light	Deep
$\text{Sensitivity} = \frac{TP}{TP + FN}$	Threshold	0.83	0	0.25	0.39
	5km2	0.42	0	0.46	0.77
$\text{Specificity} = \frac{TN}{FP + TN}$	Threshold	0.55	0.96	0.83	1.00
	5km2	0.76	0.93	0.49	0.99
$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$	Threshold	0.78	0.96	0.75	0.97
	5km2	0.48	0.93	0.49	0.98
$\text{Precision} = \frac{TP}{TP + FP}$	Threshold	0.89	0	0.19	1.00
	5km2	0.89	0	0.12	0.75
$\text{Balanced accuracy} = \left(\frac{TP}{TP + FP} + \frac{TN}{TN + FN}\right) \Big/ 2$	Threshold	0.66	0.50	0.53	0.98
	5km2	0.56	0.50	0.49	0.87
$\text{F1 score} = \frac{2 \cdot \text{precision} \cdot \text{sensitivity}}{\text{precision} + \text{sensitivity}}$	Threshold	0.86	0	0.22	0.56
	5km2	0.57	0	0.19	0.76

Table 10: The average of all the performance evaluation indexes in Table 8.

Method	Awake	REM	Light	Deep
Threshold	0.76	0.40	0.46	0.82
5km2	0.61	0.39	0.37	0.85

According to Table 9, the classification for deep sleep shows good performance in both of the proposed methods. The performance of the classification for awake is acceptable in both proposed methods, but the Threshold method is better than the 5km2 method. The performance of the classification for REM and light sleep is relatively lower in the two proposed methods. For the REM classification performance, the two proposed methods are very close, but the Threshold method is also better than 5km2 method for the light sleep classification performance.

7 Conclusion

In this paper, we propose sleep stages classification algorithms based only on wrist movements acquired by a worn accelerometer. The proposed algorithms include the “Threshold method” and the “5km2 method”. The “Threshold method” uses three thresholds to achieve falling asleep/ waking up detection and sleep stages (“awake”, “light sleep”, “deep sleep” and “REM”) classification. The “5km2 method” achieves sleep stages (“awake”, “light sleep”, “deep sleep” and “REM”) classification by performing k-means clustering (k=2) 5 times. We enrolled 5 volunteers (2 males, 3 females) who carried out validation tests for 16 full nights. Among the 16 nights, 10 nights show that the “5km2” method is better than the “Fitbit” and the “Threshold” methods, 4 nights show a close performance, only 2 nights show that the “5km2” method is worse. However, the Fitbit is not the gold standard for sleep monitoring, just as subjective feedback on sleep is not sufficiently reliable as a reference either. Moreover, we have defined a sleep score calculation method to assess the sleep quality of a full night. With tests conducted over 15 nights, the sleep score obtained by the method we propose shows promising performance in determining the sleep is good or not. As a preliminary validation of the two methods proposed for the sleep stages classification, one volunteer done a full night's sleep monitoring with the PSG at the hospital. Based on the confusion matrix analysis, the results show that the proposed 5km2 method and the Threshold method has a slight and fair agreement with the PSG respectively. Both methods are particularly efficient in the detection of awake and deep sleep. In the future, we plan to adopt the PSG as a reference device for testing the proposed methods by recruiting more subjects and organizing more trials for each subject. In addition, some classical machine learning methods such as multilayer perceptron, support vector machines and random forests (Maior et al., 2020; Tsekoura and Foka, 2020); time-frequency transform methods such as Fourier transform, wavelet transform and statistical feature extraction methods (Mohammed et al., 2019) as well as the popular neural networks (Arefnezhad et al., 2020) have performed well in similar research areas. In our future work, we also plan to try to adopt these methods to improve the performance of our proposed algorithms.

Acknowledgement

This work was partially supported by the China Scholarship Council (CSC201701810147). Special thanks are due to Dr Debs from Toulouse-Purpan Hospital (CHU-Sleep Unit) for her support in interpreting the PSG data.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Reference

Acebo, C., & LeBourgeois, M. K. (2006). Actigraphy. *Respiratory care clinics of North America*, *12*(1), 23-30.

Arefnezhad, S., Samiee, S., Eichberger, A., Frühwirth, M., Kaufmann, C., & Klotz, E. (2020). Applying deep neural networks for multi-level classification of driver drowsiness using Vehicle-based measures. Expert Systems with Applications, 162, 113778. https://doi.org/10.1016/j.eswa.2020.113778

Beattie, Z., Oyang, Y., Statan, A., Ghoreyshi, A., Pantelopoulos, A., Russell, A., & Heneghan, C. J. P. M. (2017). Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals. *Physiological measurement*, *38*(11), 1968-1979. https://doi.org. 10.1088/1361-6579/aa9047

Brand, S., Gerber, M., Kalak, N., Kirov, R., Lemola, S., Clough, P. J., ... & Holsboer-Trachsler, E. (2014). Adolescents with greater mental toughness show higher sleep efficiency, more deep sleep and fewer awakenings after sleep onset. Journal of Adolescent Health, 54(1), 109-113. https://doi.org. 10.1016/j.jadohealth.2013.07.017

Buboltz Jr, W., Jenkins, S. M., Soper, B., Woller, K., Johnson, P., & Faes, T. (2009). Sleep habits and patterns of college students: an expanded study. *Journal of College Counseling*, *12*(2), 113-124. https://doi.org/10.1002/j.2161-1882.2009.tb00109.x

Carskadon, M. A., & Dement, W. C. (2005). Normal human sleep: an overview. *Principles and practice of sleep medicine*, *4*, 13-23. https://doi.org/10.1016/B0-72-160797-7/50009-4

Chambon, S., Galtier, M. N., Arnal, P. J., Wainrib, G., & Gramfort, A. (2018). A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *26*(4), 758-769. https://doi.org/10.1109/TNSRE.2018.2813138

Daskalova, N., Lee, B., Huang, J., Ni, C., & Lundin, J. (2018). Investigating the effectiveness of cohort-based sleep recommendations. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2(3), 1-19. https://doi.org/10.1145/3264911

Diykh, M., Li, Y., & Wen, P. (2016). EEG sleep stages classification based on time domain features and structural graph similarity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *24*(11), 1159-1168. https://doi.org/10.1109/TNSRE.2016.2552539

Diykh, M., Miften, F. S., Abdulla, S., Saleh, K., & Green, J. H. (2019). Robust approach to depth of anaesthesia assessment based on hybrid transform and statistical features. IET Science, Measurement & Technology, 14(1), 128-136. https://doi.org/10.1049/iet-smt.2018.5393

Dregan, A., & Armstrong, D. (2011). Cross-country variation in sleep disturbance among working and older age groups: an analysis based on the European Social Survey. *International psychogeriatrics*, *23*(9), 1413-1420. https://doi.org/10.1017/S1041610211000664

Gu, W., Shangguan, L., Yang, Z., & Liu, Y. (2015). Sleep hunter: Towards fine grained sleep stage tracking with smartphones. *IEEE Transactions on Mobile Computing*, *15*(6), 1514-1527. https://doi.org/10.1109/TMC.2015.2462812

Guettari, T., Istrate, D., Boudy, J., Benkelfat, B. E., Fumel, B., & Daviet, J. C. (2016). Design and first evaluation of a sleep characterization monitoring system using a remote contactless sensor. *IEEE journal of biomedical and health informatics*, *21*(6), 1511-1523. https://doi.org/10.1109/JBHI.2016.2639823

Güneş, S., Polat, K., & Yosunkaya, Ş. (2010). Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting. *Expert Systems with Applications*, *37*(12), 7922-7928. https://doi.org/10.1016/j.eswa.2010.04.043

Hirshkowitz, M., Whiton, K., Albert, S. M., Alessi, C., Bruni, O., DonCarlos, L., ... & Neubauer, D. N. (2015). National Sleep Foundation's sleep time duration recommendations: methodology and results summary. *Sleep health*, *1*(1), 40-43. https://doi.org/10.1016/j.sleh.2014.12.010

Kalkbrenner, C., Brucher, R., Kesztyüs, T., Eichenlaub, M., Rottbauer, W., & Scharnbeck, D. (2019). Automated sleep stage classification based on tracheal body sound and actigraphy. *GMS German Medical Science*, *17*. https://doi.org/10.3205/000268

Krieger, A. C. (2017). Social and Economic Dimensions of Sleep Disorders, An Issue of Sleep Medicine Clinics, E-Book (Vol. 12, No. 1). Elsevier Health Sciences.

Krističević, T., Štefan, L., & Sporiš, G. (2018). The associations between sleep duration and sleep quality with body-mass index in a large sample of young adults. International journal of environmental research and public health, 15(4), 758. https://doi.org/10.3390/ijerph15040758

Kumar, P., Saini, R., Roy, P. P., Sahu, P. K., & Dogra, D. P. (2018). Envisioned speech recognition using EEG sensors. Personal and Ubiquitous Computing, 22(1), 185-199. https://doi.org/10.1007/s00779-017-1083-4

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. biometrics, 159-174. https://doi.org/10.2307/2529310

Leng, M., Yin, H., Zhang, P., Jia, Y., Hu, M., Li, G., ... & Chen, L. (2020). Sleep quality and health-related quality of life in older people with subjective cognitive decline, mild cognitive impairment, and Alzheimer disease. The Journal of nervous and mental disease, 208(5), 387-396. https://doi.org/10.1097/NMD.0000000000001137

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

Maior, C. B. S., das Chagas Moura, M. J., Santana, J. M. M., & Lins, I. D. (2020). Real-time classification for autonomous drowsiness detection using eye aspect ratio. Expert Systems with Applications, 158, 113505.

Ohayon, M. M. (2007). Prevalence and comorbidity of sleep disorders in general population. La Revue du praticien, 57(14), 1521-1528. PMID: 18018450

Pan, Q., Brulin, D., & Campo, E. (2020). Current Status and Future Challenges of Sleep Monitoring Systems: Systematic Review. JMIR Biomedical Engineering, 5(1), e20921. https://doi.org/10.2196/20921

Pollak, C. P., Tryon, W. W., Nagaraja, H., & Dzwonczyk, R. (2001). How accurately does wrist actigraphy identify the states of sleep and wakefulness?. *Sleep*, *24*(8), 957-965. https://doi.org/10.1093/sleep/24.8.957

Rechtschaffen, A., Kales, A. (1968). A manual of standardized terminology, technique and scoring system for sleep stages of human subjects. *Public Health Service*.

Stephansen, J. B., Olesen, A. N., Olsen, M., Ambati, A., Leary, E. B., Moore, H. E., ... & Sun, Y. L. (2018). Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. Nature communications, 9(1), 1-15. https://doi.org/10.1038/s41467-018-07229-3

Thorpy, M. (2017). International classification of sleep disorders. In Sleep disorders medicine (pp. 475-484). Springer, New York, NY. https://doi.org/10.1007/978-1-4939-6578-6_27

Tsekoura, K., & Foka, A. (2020). Classification of EEG signals produced by musical notes as stimuli. Expert Systems with Applications, 159, 113507. https://doi.org/10.1016/j.eswa.2020.113507

Vail-Smith, K., Felts, W. M., & Becker, C. (2009). Relationship between sleep quality and health risk behaviors in undergraduate college students. College Student Journal, 43(3), 924-930.

Van Hese, P., Philips, W., De Koninck, J., Van de Walle, R., & Lemahieu, I. (2001, October). Automatic detection of sleep stages using the EEG. In *2001 conference proceedings of the 23rd annual international conference of the IEEE engineering in medicine and biology society* (Vol. 2, pp. 1944-1947). IEEE.

Velicu, O. R., Madrid, N. M., & Seepold, R. (2016, February). Experimental sleep phases monitoring. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)* (pp. 625-628). IEEE.

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*(2), 165-193. https://doi.org/10.1007/s40745-015-0040-1

Zambotti, M., Goldstone, A., Claudatos, S., Colrain, I. M., & Baker, F. C. (2018). A validation study of Fitbit Charge 2™ compared with polysomnography in adults. *Chronobiology international*, *35*(4), 465-476. https://doi.org/10.1080/07420528.2017.1413578