TOWARDS UNOBTRUSIVE AUTOMATED SLEEP STAGE CLASSIFICATION

Polysomnography using Electrodes on the Face

Igor J. Berezhnoy, Gert-Jan de Vries, Tim Weysen, Jonce Dimov and Gary Garcia-Molina

Philips Research Laboratories, High Tech Campus 34, 5656 AE Eindhoven, The Netherlands

Keywords: Sleep, Automated Scoring, Comfortable Polysomnography.

Abstract: Although sleep stage annotation (SSA) is historically known from clinical practice and typically performed by a certified expert on the basis of visual examination of polysomnography (PSG) signals. Automatic SSA has emerged as a tool to assist sleep experts and to accelerate the analysis of PSG data. New advances in signal processing and sensor technology start to enable the application of SSA in home solutions as well. In today's busy lives, sleep plays a central role and good quality sleep helps us to deal with the stress of everyday life. Being able to enhance sleep quality thus is a major opportunity to help people in reducing the influence of stress on their live, health and wellbeing. The advent of consumer products aimed at enhancing the sleep experience has propelled the need for home sleep monitoring and inducing solutions which can i) provide automatic SSA using sensors that interfere minimally with the sleep process and ii) provide sleep stage information in real-time in order to be suitable for closed-loop sleep inducing solutions. In this paper, we examine two possible alternatives for unobtrusive sleep monitoring. The first one uses respiratory, cardiac and wrist actigraphy signals while the second one relies on Facial PSG electrodes positioned on the facial area which allow for unobtrusive and comfortable sensors arrangements.

1 INTRODUCTION

Sleep stage scoring in clinical practice results from manual visual scoring on the basis of Polysomnography (PSG) data. Typical PSG setups require affixing electrodes at various locations of the patient's body to capture a number of physiological signals including electroencephalogram (EEG), cardiac activity, respiratory effort, ocular activity (EOG), and chin myographic activity (EMG). Consumer applications relying on sleep monitoring do not require the same level of accuracy as clinical practice. In addition, the level of obtrusiveness of traditional clinical PSG setups is unacceptable in the consumer domain.

In this paper, we consider two minimally obtrusive sleep monitoring approaches. The first approach, which we refer to as RHA, is based on the monitoring of respiratory effort, cardiac signals, and wrist actigraphy. The second approach, which we refer to as Facial PSG, is based on the measurement of signals from two electrodes positioned on a single side of the subjects's face.

To establish a reference for our study, in Table 1 we present average agreement scores between two sleep stage scoring approaches i) clinical sleep scoring (performed by trained human experts and scoring according to criteria of (Rechtschaffen and A.Kales, 1968), and ii) an automated scoring techniques. Both of their approaches use data of the full PSG setup (EEG, EOG and EMG).

Table 1 shows the level of agreement between raters in terms of i) Cohen's Kappa statistics (second column) and ii) percentage of agreement (third column). It is important to note that the reported values in Table 1 were not obtained from the same dataset. They originate from various studies: i) average agreement between two independent sleep experts (row two) observing the same PSG data has been reported in (Virkkala, 2005), ii) comparison of automated techniques has been performed on the data recorded in this study.

Table 1: Sleep stages classification agreement figures.

	Cohen's Kappa	Agreement(%)
Human expert(PSG) vs. Human expert (PSG)	0.80	86.00
Automated technique A (PSG) vs. Automated technique B (PSG)	0.69	85.48

J. Berezhnoy I., de Vries G., Weysen T., Dimov J. and Garcia-Molina G..
TOWARDS UNOBTRUSIVE AUTOMATED SLEEP STAGE CLASSIFICATION - Polysomnography using Electrodes on the Face.
DOI: 10.5220/0003890104870492
In *Proceedings of the International Conference on Health Informatics* (BSSS-2012), pages 487-492
ISBN: 978-989-8425-88-1

From the agreement figures shown in the Table 1 we can conclude that agreement between human raters is significantly higher than that of automated techniques with expert annotation despite operating on the same kind of input data (PSG).

In this paper we compare performances of the RHA and Facial PSG approaches on the task of automated epoch bases sleep state annotation. This paper is organized as follows. Section 2 describes the experimental setup. In Section 3 we present the signal processing and machine learning methods that were used in this study. The results are discussed in Section 4. Finally Section 5 summarizes the main conclusions of this work and proposes future research directions.

2 EXPERIMENT DESCRIPTION

This study considers two data sets: First and Second, acquired in two separate studies. The First dataset includes respiratory effort, cardiac, and acti-graphy signals. The RHA approach was tested on the First data set.

The Second dataset was used for validation of the Facial PSG approach which had as a major objective to show that sleep state estimation can be performed with electrodes applied on a single side of the face (for which we used EOG-Left placed above the eye and mastoid reference). Further details on the datasets are provided in the next sections.

2.1 FirstDataset

The First data set contains overnight PSG recordings of six young healthy volunteers (mean age 27 y.o., 4 males and 2 females). In a screening phase, selection of participants was based on absence of subjective sleep complaints and regular sleep/wake patterns.

Participants entered the sleep laboratory at 21.00h and were prepared for PSG measurements. Lights were turned off at around 23.00h. The waking up signal was given around 7 o'clock.

Sleep recordings and analysis Polysomnographic sleep recordings were obtained during all sleep episodes with a digital recorder, and included EEG (F3/A2, F4/A1, C3/A2, C4/A1, O1/A2, O2/A1) sampled at 100Hz, electrooculogram (EOG) sampled at 100Hz, electrocardiogram (ECG) sampled at 500Hz and chin electromyogram (EMG) sampled at 200Hz. Respiratory effort was measured with chest and abdominal respiratory effort belts at 10Hz. Obtained PSG recording were scored into sleep stages using 30s epochs according to standard criteria (Rechtschaffen and A.Kales, 1968) by the Alice Philips Resperonics system and further proof checked by human experts.

2.2 SecondDataset

The Second data set contains overnight PSG recordings of six young healthy volunteers (mean age 27, 4 males and 2 females). They signed a consent form. In a screening phase, selection of participants was based on absence of subjective sleep complaints and regularity of sleep/wake patterns. Screening was based on two questionnaires: the Sleep Disorders Questionnaire (SDQ) (Douglass et al., 1994) and the Pittsburgh Sleep Quality Index (PSQI) (Buysse et al., 1989). All selected participants scored within the normal range of the PSQI. Moreover, the participant should not score higher than the cutoff scores on the subscales narcolepsy, apnea, restless legs, and psychiatry of the SDQ (Douglass et al., 1994).

Participants entered the sleep laboratory at 21.00h and were prepared for Polysomnographic measurements. Lights were turned off at around 0.00h. The waking up signal was given around 7 o'clock.

Polysomnographic sleep recordings were obtained during all sleep episodes with a digital recorder (Vitaport-3, TEMEC Instruments, Kerkrade, Netherlands), and included EEG (F3/A2, F4/A1, C3/A2, C4/A1, O1/A2, O2/A1) obtained with the Sleep BraiNet system (Jordan NeuroScience, San Bernardino, CA), electrooculogram (EOG), electrocardiogram (ECG) and chin electromyogram (EMG). Respiratory effort was measured with Pro-Tech chest and abdominal respiratory effort belts (Pittsburgh, USA). The signals were recorded at a sampling frequency of 256Hz. Obtained PSG recording were scored by the Siesta Group's software system - ”Somnolyzer” and further proof checked by Siesta's experts.

3 METHODS

3.1 Feature Extraction

In the following two subsections we describe the data preprocessing and feature extraction methodology used for both the RHA and the Facial PSG approaches.

3.1.1 RHA Features

The raw respiration signal was first lowpass filtered (cut-off 0.5Hz) and then analyzed for individual breaths. Based on a localized min/max filter, lo-

Table 2: List of features in the RHA approach, extracted from respiration, heart and wrist actigraphy signals.

Feature number	Heart rate features:	Description
1	Mean	Average heart rate
2	Median	Median heart rate
3	Standard deviation	Standard deviation of the heart rate
4	Gradient	Inclanation coefficient, shows whether heart rate goes up or down within the epoch
5	Variability	Heart rate variability
	Actigraphy features:	
6	Amount of motion over time	Scalar with values between 0 and 1, shows relative amount of motion detected
	Respiration frequency features:	
7	Mean	Average respiration rate
8	Median	Median respiration rate
9	Standard deviation	Standard deviation of the respiration rate
10	Gradient	Inclanation coeficiant, shows whether respiration rate goes up or down within the epoch
	Respiration amplitude features:	
11	Mean	Average of amplitudes of respiration cycles
12	Median	Median of amplitudes of respiration cycles
13	Standard deviation	Standard deviation of amplitudes of respiration cycles
14	Gradient	Inclination coefficient, shows whether amplitude goes up or down within the epoch

cal minima and maxima were detected. When found in the right order, they characterize a single breath. Based on the distribution of identified breath amplitudes in a signal, breaths which were too small or too large (outliers) were removed.

In a similar manner, the ECG signal is lowpass filtered (cut-off 5 Hz) and de-trended. Individual heart beats are detected using pattern matching. Again, outlier removal is applied and the resulting signal is a sequence of inter beat intervals (IBIs), which has been transformed into (instantaneous) heart rate (in bpm) by taking its reciprocal and multiplying by 60. The wrist actigraphy signal has been low-passed using a running average filter (5x30 seconds epochs in size) and further normalized on a unit interval.

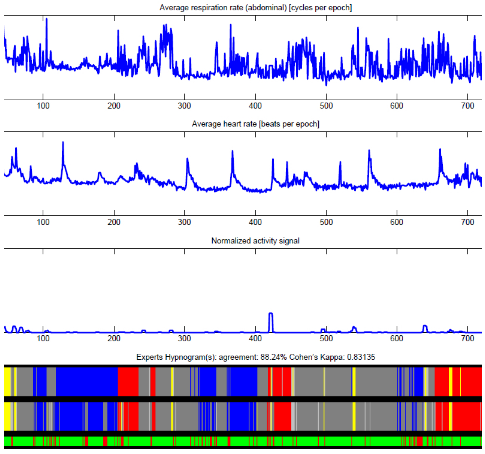

Figure 1: Respiration rate, heart rate and actigraphy signals as they change over the night, plotted along the two hypnograms from two experts. Colors in the lower graph represent sleep states.

Sleep is normally scored (Rechtschaffen and A.Kales, 1968) in non-overlapping 30-second long intervals (*epochs*). Thus, features of respiration, heart and actigraphy signals are calculated on a per-epoch basis. Table 2 lists the complete set of features extracted from the signals.

3.1.2 Features in the Facial PSG Approach

The raw signal used for feature extraction in the Facial PSG approach was recorded by electrodes placed at the following three standardized locations:(1) upper left eye (EOG_Left), (2) left mastoid bone (reference A1) and (3) ground electrode at the neck of the participant. Given this setup for signal extraction the signal recorded at A1 channel was subtracted from the signal of the EOG_L channel. Furthermore, to estimate the power spectral density of each epoch, we applied Welch's method (Welch, 1967). Figure 2 shows results of the Welch's method where the color represents the power at a certain frequency (top plot).

To facilitate the visual interpretation of the relation between the Welch's power plot(features) and the reference scoring (labels), the bottom plot in the figure shows corresponding hypnogram and the middle plot shows power plot but specifically for low frequencies which correspond to deeper sleep (a.k.a. slow wave sleep, SWS). As it can be easily seen, peaks of power in SWS plot correspond to N3 sleep stages of the hypnogram.

For the machine learning part of the Facial PSG approach input-output pairs were constructed in the following manner: for each epoch, a power spectrum vector was computed and coupled with a sleep stage label. This resulted in about 800 input-output pairs per subject (corresponding to 7 hours of sleep).

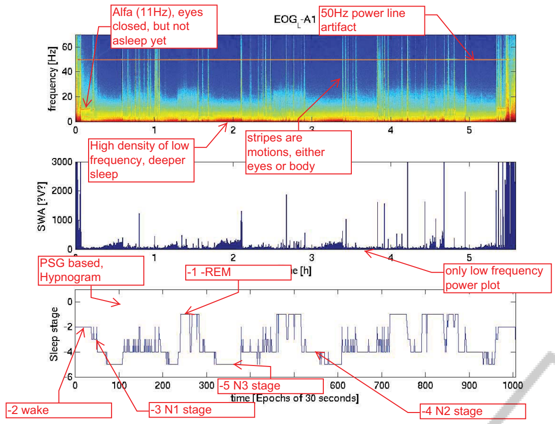

Figure 2: Facial PSG approach, features and labels (from top-to bottom): (1) Signal power vs. frequency over time, (2) Low frequency(deeper sleep) power over time, (3) hypnogram plot.

3.2 Classification

Robust Soft Learning Vector Quantization (RSLVQ) is one of many LVQ variants, originally developed by Kohonen (Kohonen, 1995). This family of machine learning algorithms has been applied to classification problems in many fields (Centre, 2002) and is characterized by its computational efficacy. LVQ is a method of prototype-based, multi-class classification, where each class is represented by one or more prototypes. A prototype is defined as a point in the *N*-dimensional feature space with an accompanying class label, and trained by sequential handling of training data. Each time a training sample is presented, the closest prototypes with correct and incorrect label are pulled towards or pushed away from the training sample, respectively. When training progresses, the prototypes will progressively better represent the classes. When applied to unseen data, classification is performed by returning the label of the closest prototype. Usually Euclidean distance is used as distance measure.

In a recent study (Witoelar et al., 2010), the performance of several LVQ variants in a controlled environment was analyzed. The (relative) robustness and convergent properties (i.e., insensitivity to overtraining) motivated our choice for RSLVQ, as proposed in (Seo and Obermayer, 2003). In this 'soft' version of LVQ the magnitude of displacement of prototypes in each training step is relative to their distance from the training sample. This method makes an assumption on the distribution of data samples around the prototypes, which we chose to be Gaussian with equal variances (for each prototype). The total distribution of data from a single class therefore is assumed to be a mixture of Gaussian distributions.

3.3 Performance Measurement

In order to allow in depth comparisons of the two techniques we present classification results of both experiments in the shape of confusion matrixes. Essentially confusion matrix contains three widely known (in classification tasks assessments) comparison entities: (1) confusion matrix, (2) percentage of agreement and (3) Cohen's Kappa agreement coefficient. The confusion matrix can be used for detailed assessment of classifier's performance in terms of which classes are often mistaken for what other classes. Furthermore they allow calculation of a baseline performance based on just class priors. Since we were mostly interested in overall performance assessment, in section 4 for each cycle of the cross validation scheme we only present it's outcome with two values: (1) percentage of agreement and (2) Cohen's Kappa coefficient.

4 RESULTS AND DISCUSSION

This section presents the results obtained by two sleep monitoring approaches, namely Facial PSG and RHA. Section 4.1 reports the results obtained with Facial PSG while Section 4.2 reports the results obtained with the RHA approach. Both subsections contain tables presenting percentages of agreement and Cohen's Kappa coefficients per cross validation run, as well as overall agreement matrices.

4.1 Facial PSG Results

Table 3 shows Cophen's Kappa and percentage of agreement figures per run of the cross-validation scheme. The last column contains average values.

Table 3: Facial PSG approach results, per cross validation run (per subject).

	Subjects (Second study)						
	1	**2**	**3**	**4**	**5**	**6**	**mean**
Agreement %	76.20	71.43	71.46	64.71	80.91	82.07	74.46
Cohen's Kappa	0.66	0.58	0.62	0.50	0.69	0.74	0.63

Table 4 shows the overall agreement matrix that contains: confusion matrix (in bold), percentage of agreement, Cohen's Kappa coefficient, positive predictive values (PPV) and sensitivity of the classifier per class.

From table 4 it can be seen that the overall performance significantly exceeds random guessing, which is $1989/6292 = 31.61\%$. Furthermore it can be seen the largest number of confusions is for actual wake epochs being (falsely) recognized as light sleep. Actually, the classifier is falsely biased towards light

Table 4: Facial PSG approach overall results. Confusion matrix (in bold), percentage of agreement, Cohen's Kappa coefficient, positive predictive values (PPV) and sensitivity of the classifier per class.

Overall	Wake	Light	Deep	Rem	Sum	Sensitivity
Wake	**1206**	**42**	**8**	**42**	1298	92.91%
Light	**320**	**1931**	**456**	**430**	3137	61.56%
Deep	**22**	**102**	**735**	**15**	874	84.10%
Rem	**26**	**118**	**16**	**796**	956	83.26%
Sum	1574	2193	1215	1283	6265	
PPV	76.62%	88.05%	60.49%	62.04%		
Agreement						74.51%
Cohen's Kappa						0.64317

sleep, as it classifies half of the total number of epochs as light sleep (i.e, $3173/6292 = 50.43\%$), resulting in a low sensitivity (52.66%) for that class.

In addition to numerical representation of the classification figure 3 shows both input data (Facial PSG spectrum; middle plot) and graphical representation of the hypnograms both target and estimated (bottom plot). The top plot of the figure shows the power spectrum of C4-A1 PSG channel, which served as an input for an additional experiment we conducted. The essence of the experiment was in substituting Facial PSG signal with C4-A1 signal. Given the fact that C4 electrode of the PSG setup is mounted close to the brain and subsequently has a stronger signal, our assumption was to observe gain in classification performance. However, this experiment proved an opposite effect. Despite better signal to noise ratio, a significant drop in performance of the classifier appeared. This may indicate that when it comes to a single channel PSG, electrode-positions in the Facial PSG are better suited for sleep stage estimation.

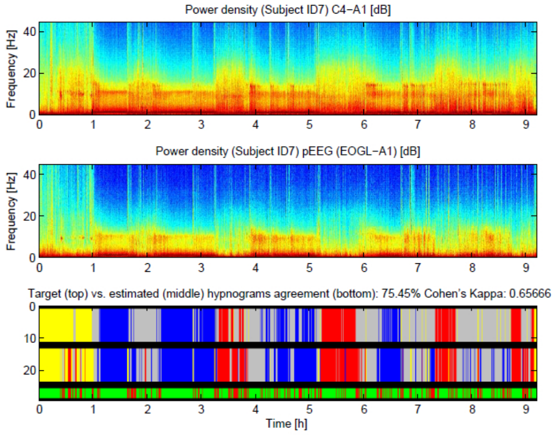

Figure 3: Results and input data (spectrum) visualization for subject ID7.

Table 5 shows overall performance matrix for C4-A1 channel. From this table we may see (when compared to Table 4) that Cohen's Kappa statistics lowered by 0.0662 and percentage of agreement by 6.55%.

Table 5: Facial PSG approach overall results on C4-A1 channel. Confusion matrix (in bold), percentage of agreement, Cohen's Kappa coefficient, positive predictive values (PPV) and sensitivity of the classifier per class.

Overall	Wake	Light	Deep	Rem	Sum	Sensitivity
Wake	**1199**	**18**	**15**	**66**	1298	92.37%
Light	**649**	**1263**	**799**	**426**	3137	40.26%
Deep	**44**	**97**	**729**	**4**	874	83.41%
Rem	**36**	**10**	**241**	**669**	956	69.98%
Sum	1928	1388	1784	1165	6265	
PPV	62.19%	90.99%	40.86%	57.42%		
Agreement						61.61%
Cohen's Kappa						0.49303

4.2 RHA , - Respiration, Heart and Actigraphy Signals for Hypnogram Estimation

Table 6 shows Cohen's Kappa and percentage of agreement figures per run of the cross-validation scheme. The last column contains average values.

Table 6: RHA , - respiration, heart and actigraphy features approach results, per cross validation run (per subject).

	Subjects (Boston data set)						
	1	**2**	**3**	**4**	**5**	**6**	**mean**
Agreement %	48.37	33.62	27.87	32.33	43.74	13.55	**33.25**
Cohen's Kappa	0.26	0.09	0.06	0.09	0.25	0.32	**0.18**

Table 7 shows the overall agreement matrix that contains: confusion matrix (in bold), percentage of agreement, Cohen's Kappa coefficient, positive predictive values (PPV) and sensitivity of the classifier per class.

Table 7: RHA , - respiration, heart and actigraphy features approach overall results. Confusion matrix(in bold), percentage of agreement, Cohen's Kappa coefficient, positive predictive values (PPV) and sensitivity of the classifier per class.

Overall	Wake	Light	Deep	Rem	Sum	Sensitivity
Wake	**247**	**10**	**8**	**73**	338	73.08%
Light	**596**	**761**	**723**	**1020**	3100	24.55%
Deep	**118**	**158**	**370**	**254**	900	41.11%
Rem	**221**	**241**	**53**	**368**	883	41.68%
Sum	1182	1170	1154	1715	5221	
PPV	20.90%	65.04%	32.06%	21.46%		
Agreement						33.44%
Cohen's Kappa						0.12265

Table 8 shows the agreement figures presented earlier in table 1 along with agreement figures achieved by RHA and Facial PSG approaches. From these figures we notice that the Facial PSG approach is superior compared to the RHA one in both percentages of agreement and Cohen's Kappa coefficient numbers. Figures of the RHA approach shows a very low performance of the classifies when based on respiration, heart and actigraphy features. It can be seen that the overall performance is very close to random guessing, which is $1715/5221 = 32.85\%$. Again, the

Table 8: Sleep stages classification agreement figures.

	Cohen's Kappa	Agreement(%)
Human expert(PSG) vs. Human expert (PSG)	0.80	86.00
Automated technique A(PSG) vs. Automated technique B(PSG)	0.69	85.48
Automated technique(PSG) vs. Facial PSG	0.64	74.51
Automated technique(PSG) vs. RHA	0.12	33.44

classifier is falsely biased towards light sleep, as it classifies most of the total number of epochs as light sleep (i.e, $3100/5221 = 59.38\%$), resulting in a very low sensitivity (24.55%) for that class.

5 CONCLUSIONS

In our study we were not able to find significant correspondence at individual level(cross subjects) between PSG based sleep stages estimated by experts and the features we extracted in the RHA approach. This conclusion only holds for this paper's particular combination of features and classifier. A separate study on separability involving at least larger number of subjects is required in order to strengthen this conclusion.

In contrast to the RHA, classification, results obtained on features extracted in the Facial PSG approach look very promising, the good performance of the classifier indicate good separability of different sleep states manifested in Facial PSG features. In the current study, we employed only "simple" (low-capacity, epoch based) classifier that already showed a performance of 74.51% agreement and 0.64 Cohen's Kappa. Performance wise these results position the Facial PSG approach next to full PSG based automated techniques (see Table 1). Further improvement of classification performance can be expected when order and transition probabilities between sleep stages are taken into account.

In addition to very good performance indicators the Facial PSG is also much less obtrusive compared to a full PSG setup. The full PSG setup employs 8+ channels resulting in 12+ electrodes mounted on the skull of a subject, whereas Facial PSG uses only one channel (3 electrodes) mounted on a single side of the face allowing subject to have more natural sleeping position(s) and what is even more important more natural sleep. Given its obvious advantage of being more comfortable the Facial PSG approach makes sleep studies less labor intensive and consequently more accessible, also making multiple night studies more realistic which will enable research to better understand the complex mechanisms involved in humans sleep.

REFERENCES

Buysse, D., Reynolds-III, C., Monk, T., Berman, S., and Kupfer, D. (1989). Pittsburgh sleep quality index: A new instrument for psychiatric practice and research. *Psychiatric Research*, 28(2):193–213.

Centre, N. N. R. (2002). Bibliography on the self-organizing maps (som) and learning vector quantization (lvq).

Douglass, A., Bornstein, R., Ninomurcia, G., Keenan, S., Miles, L., and Zarcone, V. (1994). The Sleep Disorders Questionnaire-I - Creation and Multivariate Structure of SDQ. *Sleep*, 17(2):160–167.

Kohonen, T. (1995). *Learning vector quantization, The handbook of brain theory and neural networks*. MIT Press, Cambridge, MA.

Rechtschaffen, A. and A.Kales (1968). *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. U. S. National Institute of Neurological Diseases and Blindness, Neurological Information Network, Bethesda, Md.,.

Seo, S. and Obermayer, K. (2003). Soft learning vector quantizatio. *Neural Computation*, 15:1589–1604.

Virkkala, J. (2005). *Automatic Sleep Stage Classification Using Electro-oculography*. PhD thesis, Faculty of Computing and Electrical Engineering. Tampere University of Technology.

Welch, P. (1967). The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio Electroacoustics*, AU-15:7073.

Witoelar, A. W., Ghosh, A., de Vries, J. J. G., Hammer, B., and Biehl, M. (2010). Window-based example selection in learning vector quantization. *Neural Computation*, 22(11):2924–2961.