Automatic sleep stage classification based on easy to register signals as a validation tool for ergonomic steering in smart bedding systems

Tim Willemen[a,*], Dorien Van Deun[a], Vincent Verhaert[a], Sandra Pirrera[b], Vasileios Exadaktylos[b], Johan Verbraecken[b], Bart Haex[b] and Jos Vander Sloten[b]

[a]*Division of Biomechanics, Katholieke Universiteit Leuven, Celestijnenlaan 300C, 3001 Heverlee, Belgium*
[b]*Biological Psychology, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium*
[c]*M3-BIORES, Katholieke Universiteit Leuven, Kasteelpark Arenberg 30, 3001 Heverlee, Belgium*
[d]*Multidisciplinary Sleep Disorders Center, Antwerp University Hospital and University of Antwerp, Wilrijkstraat 10, 2650 Edegem, Belgium*

Abstract. Ergonomic sleep studies benefit from long-term monitoring in the home environment to cope with daily variations and habituation effects. Polysomnography allows to asses sleep accurately, but is costly, time-consuming and possibly disturbing for the sleeper. Actigraphy is cheap and user friendly, but for many studies lacks accuracy and detailed information. This proof-of-concept study investigates Least-Squares Support Vector Machines as a tool for automatic sleep stage classification (Wake-N1-Rem to N2-N3 separation), using automatic trainingset-specific filtered features as derived from three easy to register signals, namely heart rate, breathing rate and movement. The algorithms are trained and validated using 20 nights out of a 600 night database from over 100 different healthy persons. Different training and test set strategies were analyzed leading to different results. The more person-specific the training nights to the test nights, the better the classification accuracy as validated against the hypnograms scored by experts from the full polysomnograms. In the limit of complete person-specific training, the accuracy of the algorithm on the test set reached 94%. This means that this algorithm could serve its use in long-term monitoring sleep studies in the home environment, especially when prior person-specific polysomnographic training is performed.

Keywords: LS-SVM, actigraphy, polysomnography

1. Introduction

Although sleep research has been around for almost a hundred years [1], there is still a huge need for studies looking at the influence of environmental effects on sleep, both from a physiological as a psychological point of view. In general, these studies would benefit from long-term monitoring.

Polysomnography [2,3] allows to asses sleep accurately, but is costly and time-consuming to apply in long-term studies. Due to the extensive amount of sensors attached to the sleeper's body, it could also disturb the normal sleep pattern, both before and during the night [4]. Although habituation takes place after consecutive nights, letting the sleeper return to its normal sleep patterns, in most cases only the first night is disregarded from experimental analysis (due to cost and time). Habituation might however take longer than this so-called 'first-night effect' [5].

Actigraphy [6] on the other hand is both cheap and user-friendly for long-term studies, but for many studies lacks not only detailed information due to its ability to only distinguish between sleep and wake states, but also accuracy due its low specificity in detecting wake states [7]. No other methods reached the point beyond prototype or have been sufficiently validated to fill the gap in between [8].

---

*Corresponding author. E-mail: tim.willemen@mech.kuleuven.be.

This study investigates the use of Least-Squares Support Vector Machines (LS-SVM) as a tool for automatic sleep stage classification, validating their ability to classify sleep with gold standard polysomnography.

SVM theory [9] is a supervised learning method that analyzes data and recognizes patterns. Per data point in time or space, called a vector, a set of predefined variables is collected, called features. Based on these so-called feature vectors, the algorithm predicts which of the two possible classes the input is a member of. It does this with the help of a set of training examples, each belonging to one of two categories, upon which the algorithm builds the binary classification model. With the help of kernels and weights, it automatically transforms the multi-dimensional feature space, making it a nonlinear-classification tool.

The strength of LS-SVMs [10] is their quadratic computational dependence on the amount of features, compared to their cubic dependence on the amount of vectors. As long as the amount of features is kept smaller than the amount of vectors, extra features can easily be added without increasing the computational cost. This makes the method easy to expand and improve, because the more relevant features present, the better the classification will be.

In order to have a user-friendly tool in long-term monitoring sleep studies, easy to register signals should be used. This study uses features calculated from three easy to register signals. Firstly heart and breathing rate, two physiological signals coupled to the autonomic nervous system, of which variations in sympathetic and parasympathetic activation are known to correlate well with variations in sleep stages [16,17,18]. And secondly movement, of which its link with sleep and wake states is extensively researched in the domain of actigraphy [6,19].

Not only can all of these three signals be registered using a strap similar to the ones in classical heart rate monitors used for sports, they can even potentially be measured off-body in the bedding system itself, e.g. using upper body pressure mats [11,12], force sensors in the bed legs [13] or optical fibers in the bed cover [14]. Firstly, this would avoid disturbing the person during or preceding his or her sleep, both physically and mentally. Secondly, it would allow for autonomous working, so that it can easily be used in long-term studies, even at home. Measurements like these were already used in attempts to analyze sleep [11,12].

2. Methods

2.1. Subjects

During a four year project, about 100 subjects participated in different sleep studies, leading to a database of 605 different nights. All were volunteers free from medical problems that would interfere with normal sleep. All studies were approved by the Ethics Committee of the Vrije Universiteit Brussel. A subset of 20 nights was selected for this proof-of-concept study, from a total of 10 subjects between the age of 19 and 30 years old (22.95 ± 2.74 y).

2.2. Procedures

From the subset of 20 nights, three different training and test set combinations were constructed to analyze the effect of more and more person-specific training information. Table 1 lists an overview of the amount of different nights and different subjects used in each combination.

Table 1
Overview of different training and test set combinations.

	A	B	C
# training nights	8	5	3
# test nights	4	3	1
# subjects training nights	5	3	1
# subjects test nights	4	3	1

In combination A, the training set contained nights from different subjects compared to those from the test set. This was done to evaluate zero person-specific training. In combination B, from every night from a subject in the test set, there was at least one night from that same subject in the training set. Finally in combination C, all nights in the training set were from the same subject as the nights in the test set resulting in complete person-specific training.

2.3. Measurements

For all nights in the projects database, complete polysomnographic recordings were performed using the Medatec Dream System (Medatec nv., Brussels, Belgium), after which sleep stages were classified by experts according to AASM rules [3]. Additionally, heart rate was registered through ECG, breathing rate using two respiratory belts (one at the chest area and one at the abdomen) and movement by taking differences in mattress indentation as measured with one

second resolution by the Custom8 DynaSleep system (Custom8, Leuven, Belgium) [15].

2.4. Data Analysis

Heart rate (HR), breathing rate (BR) and movement (MOV) were scaled to unit mean to ensure a better inter-person working of the algorithm.

HR and BR were manually synchronized with the MOV signal based on the differences in amplitude of the raw breathing signal, as measured by the respiratory belts when the person was moving. Since sleep stages in a hypnogram, as scored according to AASM rules, only have a resolution of an epoch of length 30 seconds, this manual synchronization is more than precise enough.

Different variables, averaged over different sizes of epoch intervals, were defined based on these three signals: mean, variance, different percentiles and approximate entropy of HR and BR, frequency intervals and ratios between frequency intervals of HR and BR, time to previous and to future movements using different movement thresholds to distinguish between different movement sizes.

Using a significant clustering algorithm, a subset of semi-uncorrelated features is extracted from the complete feature set based specifically on the data from the training set and the targeted binary classification. This is done by calculating the mean and variance across all epochs for every feature, separately for each of the two classes. This allows to estimate the significance by which those two means differ from each other using statistics. This simple feature filtering not only suppresses noise on the future classification results, but also enables to learn which features are more relevant than others.

In this study, a simple Wake-N1-Rem versus N2-N3 classification is targeted. A set of 1000 epochs from each of the two classes was used to train the LS-SVM algorithm's parameters using grid search and 10-fold cross validation. Epochs were selected based on having maximum entropy in the complete set of 2000 feature vectors, in order to ensure a complete representation of the total training set. The same was done for the test sets.

Sensitivity, specificity and accuracy parameters were calculated for all three combinations using an epoch-by-epoch agreement analysis between the AASM-scored epochs and the resulting classification by the trained LS-SVM algorithm. Accuracy was calculated as the percentage total agreement, sensitivity as the percentage correctly identified Wake-N1-Rem epochs and specificity as the percentage correctly identified N2-N3 epochs. For extra information, accuracies on all 20 complete nights were also calculated using each of the three trained LS-SVM algorithms.

3. Results

Table 2, 3 and 4 list the results from training and test set combination A, B and C respectively. Firstly, the 10-fold cross-validated accuracy on the 2000 feature vectors of the training set. Secondly, the accuracy, sensitivity and specificity on the complete set of 2000 feature vectors from the test set. Thirdly, the accuracies on the complete nights of training and test set. Finally for extra information, the accuracy on the set of nights not present in training or test set for that specific combination. For combination A, these are nights from subjects also present in the training night selection; while for combination B and C, they are from subjects not present in the training night selection.

Table 2
Results for **combination A**.

Accuracy training set	94.20%
Accuracy test set	75.96%
Sensitivity test set	68.38%
Specificity test set	83.55%
Accuracy complete training nights	94.47 ± 2.31 %
Accuracy complete test nights	78.56 ± 10.57 %
Accuracy non-selected nights	89.39 ± 3.00 %

Table 3
Results for **combination B**.

Accuracy training set	94.90%
Accuracy test set	82.00%
Sensitivity test set	77.30%
Specificity test set	86.70%
Accuracy complete training nights	95.58 ± 1.26 %
Accuracy complete test nights	85.00 ± 3.61 %
Accuracy non-selected nights	79.87 ± 5.16 %

Table 4
Results for **combination C**.

Accuracy training set	94.11%
Accuracy test set	94.25%
Sensitivity test set	90.80%
Specificity test set	97.70%
Accuracy complete training nights	93.15 ± 1.53 %
Accuracy complete test nights	91.48 ± 0.00 %
Accuracy non-selected nights	81.09 ± 4.89 %

4. Discussion

4.1. Combination A

The results of table 2 show a very high accuracy on the training set (94.20%) as on the complete training nights (94.47%). This is not surprising since epochs from the same nights are used for training the classifier, thus for every epoch in a training night, there will always be an epoch close by that was used in the training set itself. Feature values from that epoch won't be too much different from the ones nearby, thus giving rise to the good results. However, this is still a good indicator that the feature variables defined are able to separate the multi-dimensional feature space between Wake-N1-Rem and N2-N3 well, thus confirming that HR, BR and MOV are not just randomly changing over sleep stages but have a certain pattern towards different sleep stages.

A reasonable accuracy was also found for the test set (75.96%) as on the complete test nights (78.56%), mainly due to a good specificity (83.55%), but lacking sensitivity (68.38%). This is however a better result than having a high sensitivity but lacking specificity, because once the classifier indicates an epoch as Wake-N1-Rem, one can at least be reasonably sure that this is correct. Sensitivity can possibly be increased afterwards using window-filters, filling in gaps between close-by-scored Wake-N1-Rem epochs, which hasn't been tried nor implemented yet. The high standard deviation on the accuracy of the full test nights (10.57%) indicates that some nights or some subjects are better or worse for automatic classifications than others.

Compared to other classification algorithms found in literature, converting their listed results to reflect the same binary classification, the results described here are on the same level. A heart rate based classification algorithm led to an accuracy of 73.27% [20], whereas an algorithm using peripheral arterial tone and actigraphy reached an accuracy of 80.56% [21]. Important to also consider are inter- and intra-scorer variations during golden standard manual scoring of polysomnographic measurements. One publication e.g. gives a compared accuracy between two expert scorers of only 85.93% on the same binary classification [22], thus lowering the threshold for having good classification results. Automatic algorithms also have reproducible results and are consistent in mistakes they make.

4.2. Combination B

The results of table 3 show again the high accuracy on the training set (94.90%) and complete training nights (95.58%). For the test set the same trend is present of having a better specificity (86.70%) than sensitivity (77.30%), leading to accuracies of 82.00% on test set and 85.00% on complete test nights. Values on all test set parameters are higher than in combination A, indicating that using more person-specific information leads to better classification possibilities.

This time, smaller standard deviations are present compared to combination A. This could either mean that no difficult-to-score nights or subjects were present in the test set, or that due to information from all subjects being present in the training set, classification was better overall.

4.3. Combination C

The results of table 4 clearly indicate that training person-specific leads to better classification. The accuracy percentage on the test set (94.25%) even becomes equal to the training set value (94.11%), indicating that not much difference in feature values must have been present between training and test nights for the same sleep stages.

4.4. General

In the current algorithm both information from past, present and future epochs is used, since some feature variables are defined having averaged values over different symmetric interval lengths around the current epoch. This is of no problem when only evaluation of sleep quality is required.

However, limiting the algorithm to only use past and present epoch information, or maybe just information from a maximum amount of epochs towards the future, will possibly enable it to steer different devices depending on the sleep state of the person. Future work will investigate how much this would deteriorate current classification results, and with which time resolution transitions between different sleep stages can be detected.

Further investigation will also consist of extending the data set for better validation and extending the sleep classification towards a full hypnogram.

5. Conclusion

Overall, results obtained look promising towards use in long-term studies, especially when person-specific training on the basis of polysomnography is possible.

Acknowledgements

References

[1] H. Piéron, Le problème physiologique du sommeil, Masson, Paris, 1913.

[2] A. Rechtschaffen and A. Kales, A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects, National Institutes of Health Publications No. 204, U.S. Government Printing Office, Washington D.C., 1968.

[3] C. Iber, S. Ancoli-Israel, A. L. Chesson and S. F. Quan, The AASM Manual for the Scoring of Sleep and Associated Events, American Academy of Sleep Medicine, Westchester, IL, 2007.

[4] M. L. Metersky and R. J. Castriotta, The effect of polysomnography on sleep position: possible implications on the diagnosis of positional obstructive sleep apnea, Respiration 63 (1996), 283-287.

[5] O. Le Bon, L. Staner, G. Hoffmann, M. Dramaix, I. San Sebastian, J. R. Murphy, M. Kentos, I. Pelc and P. Linkowski, The first-night effect may last more than one night, Journal of Psychiatric Research 35 (2001), 165-172.

[6] T. Morgenthaler, C. Alessi, L. Friedman, J. Owens, V. Kapur, B. Boehlecke, T. Brown, A. C. Junior, J. Coleman, T. Lee-Chiong, J. Pancer and T. J. Swick, Practice Parameters for the Use of Actigraphy in the Assessment of Sleep and Sleep Disorders: An Update for 2007, Sleep 30 (2007), 519-529.

[7] J. Paquet, A. Kwainska and J. Carrier, Wake detection capacity of actigraphy during sleep, Sleep 30 (2007), 1362-1369.

[8] A. T. Van de Water, A. Holmes and D. A. Hurley, Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography: a systematic review, Journal of Sleep Resarch 20 (2011), 183-200.

[9] N. Christiani and J. Shawe-Taylor, An introduction to support vector machines, Cambridge University Press, Cambridge, 2000.

[10] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle, Least-Squares Support Vector Machines, World Scientific, Singapore, 2002.

[11] J. M. Kortelainen, M. O. Mendez, A. M. Bianchi, M. Matteucci and S. Cerutti, Sleep staging based on signals acquired through bed sensor, Information Technology in Biomedicine 14 (2010), 776-785

[12] G. G. Bader and S. Engdal, The influence of bed firmness on sleep quality, Applied Ergonomics 31 (2000), 487-497.

[13] M. Brink, C. H. Müller and C. Schierz, Contact-free measurement of heart rate, respiration rate, and body movements during sleep, Behavior Research Methods 38 (2006), 511-521.

[14] J. Li, R. K. Kapania and W. B. Jr. Spillman, Placement Optimization of Distributed-Sensing Fiber-Optic Sensors Using Genetic Algorithms, A/AA Journal 46 (2008), 824-836.

[15] V. Verhaert, B. Haex, T. De Wilde, R. Cluydts, D. Berckmans, J. Verbraecken and J. Vander Sloten, Automatic posture recognition during sleep, Proceedings of the 17th world congress on ergonomics in Beijing, International Ergonomics Association (IEA) 2009.

[16] V. K. Somers, M. E. Dyken, A. L. Mark and F. M. Abboud, Sympathetic-Nerve Activity during Sleep in Normal Subjects, New England Journal of Medicine 328 (1993), 303-307.

[17] M. H. Bonnet and D. L. Arand, Heart rate variability: sleep stage, time of night and arousal influences, Electroencephalography and Clinical Neurophysiology 102 (1997), 390-396.

[18] H. J. Burgess, J. Trinder and Y. Kim, Cardiac autonomic nervous system activity during presleep wakefulness and stage 2 NREM sleep, Journal of Sleep Research 8 (1999), 113-122.

[19] A. Sadeh, The role and validity of actigraphy in sleep medicine: an update, Sleep Medicine Reviews 15 (2011), 259-267.

[20] S. Canisius, T. Ploch, T. Penzel, A. Jerrentrup, D. Krefting and K. Kesper, Classifying sleep stages using the heart rate spectrum – comparison of feasibility in healthy subjects and sleep apnea patients, Proceedings of biosignal in Berlin, Germany, 2010.

[21] S. Herscovici, A. Pe'er, S. Papyan and P. Lavie, Detecting REM sleep from the finger: automatic REM sleep algorithm based on peripheral arterial tone (PAT) and actigraphy, Physiologic Measurements 27 (2006), 1-12.

[22] M. Schwaibold, R. Harms, B. Schöller, I. Pinnow, W. Cassel, T. Penzel, H. F. Becker and A. Bolz, Knowledge-based automatic sleep-stage recognition – reduction in the interpretation variability, Somnologie 7 (2003), 59-65.