A Vision-based System for Breathing Disorder Identification: A Deep Learning Perspective

Manuel Martinez[1], David Ahmedt-Aristizabal[1,2], Tilman Väth[1],
Clinton Fookes[2], Andreas Benz[3], Rainer Stiefelhagen[1]

***Abstract*— Recent breakthroughs in computer vision offer an exciting avenue to develop new remote, and non-intrusive patient monitoring techniques. A very challenging topic to address is the automated recognition of breathing disorders during sleep. Due to its complexity, this task has rarely been explored in the literature on real patients using such marker-free approaches. Here, we propose an approach based on deep learning architectures capable of classifying breathing disorders. The classification is performed on depth maps recorded with 3D cameras from 76 patients referred to a sleep laboratory that present a range of breathing disorders. Our system is capable of classifying individual breathing events as normal or abnormal with an accuracy of 61.8%, hence our results show that computer vision and deep learning are viable tools for assessing locally or remotely breathing quality during sleep.**

I. INTRODUCTION

Sleep apnea has a complex nature, as it can be caused by a variety of underlying problems and remains often undetected when the person is sleeping alone. As complex as sleep apnea is, some cases (*i.e.*, obstructive sleep apnea) can be treated by losing weight or wearing a Continuous Positive Airway Pressure (CPAP) device during sleep, which has shown a positive effect on prognosis [1]. For many of the sufferers, sleep apnea remains undetected for all their lives, due to the complex diagnosis procedure [2].

There is the common misconception that an apnea episode corresponds simply to an interruption in the regular breathing pattern, and thus it can be easily detected by monitoring the chest expansion and contraction patterns. This misconception has lead to many experiments where algorithms that are developed to recognize apnea are tested on healthy volunteers that simulate apnea events by simply holding their breath for a few seconds. Instead, when a real patient is suffering an obstructive apnea event (one of the most common types of apnea events), it is the airflow between the lungs and the atmosphere that is obstructed, and is characterized by significant chest movements as the patient, while sleeping, is trying to clear the obstruction. The intensity of those movements increases until the pressure exerted is sufficient to clear the obstruction and finally the patient is able to breathe, finalizing the obstructive apnea event.

Motivated by recent advances in healthcare and patient monitoring based on computer vision and deep learning [3], we aim to use deep learning techniques to identify breathing disorders reducing the need of feature engineering, one of the most time-consuming phases of traditional machine learning practice. To address this problem, following the preliminary studies in [4], we propose a robust deep learning approach which uses an end-to-end architecture based on convolutional neural networks (CNNs) and long short-term memory (LSTM) network. This model classifies breathing disorders exploiting the discriminative information captured from depth cameras without using hand-crafted features. The contributions of our work are summarized as follows:

1) We introduce the first of its kind application of deep learning for vision-based sleep apnea identification.
2) We propose a robust non-obtrusive monitoring system to capture motions in natural healthcare conditions which include factors such as variable illumination conditions, self-occlusion and changing viewpoints.

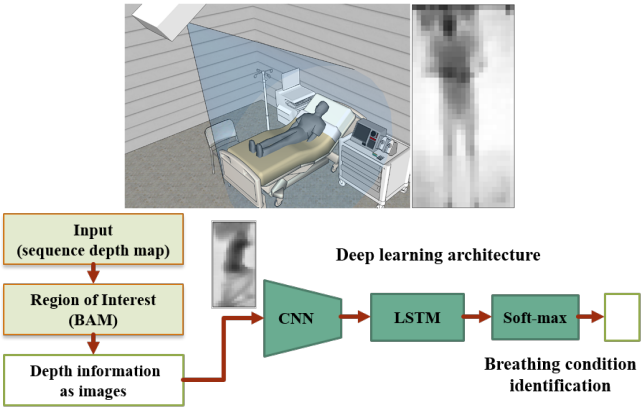

Fig. 1. Top left: our proposed system uses a custom camera system placed in the ceiling above the bed. Top right: we generate a description based on the BAM algorithm. Bottom: Our proposed deep learning framework based on CNNs and LSTM architectures. The output of the system is represented by the classification accuracy of each breathing condition.

II. BACKGROUND

Most research on apnea event recognition is based on contact sensors [5], *e.g.*, pressure mattress [6]. On the other hand, computer vision systems aim to provide non-invasive sleep monitoring methodologies that address the limitations inherent in invasive sensors and marker-based systems, including the need of maintenance, *e.g.*, disinfection, calibration, keeping batteries charged, etc.

Nakajima et al. [7], used a near-infrared camera to monitor posture changes and estimate the respiratory rate of a single subject that wears clothing with a specific mosaic pattern.

This research was supported by the German Federal Ministry of Education and Research within the SPHERE project.

[1] CV-HCI Laboratory, Karlsruhe Institute of Technology, Germany.
[2] Image and Video Research Laboratory, SAIVT, Queensland University of Technology, Australia.
[3] Heidelberg University Hospital, Germany.

TABLE I

BENCHMARK OF MARKER-FREE SYSTEMS.

Patients	Breathing analysis
<5	Aoki et al. [12], [13], [15]; Reyes et al. [9]; Zhu et al. [11]
<20	Chase et al. [8]; Takemura et al. [14]; Liao and Yang [10]; Yu et al. [17]; Martinez and Stiefelhagen [16]; Al-Naji et al. [18]
⩾ 20	Martinez and Stiefelhagen [19]^R; Martinez and Stiefelhagen [4]^R

Patients: Number of participants in the user study. *R*: Real world scenarios.

Chase et al. [8] and Reyes et al. [9] used the difference between consecutive images to measure agitation in intensive care units using color cameras. Similarly, Liao and Yang [10] used the same technique but were able to work during the night thanks to the use of infrared cameras. Thermal cameras have also been used to monitor sleep. However, those are comparatively expensive and the images are difficult to process due to the presence of afterimages that are created by the hot regions of the bed [11].

The use of RGB-D cameras (color and depth streams, *e.g.*, Microsoft Kinect) have shown sufficient accuracy when compared to the established and precise optical multi-camera motion capture system [12], [13], [14], [15], [16]. The limitations of RGB-D were analyzed in [16], which suggests that the signal-to-noise ratio of the respiration signal decreases with the 4^{th} power of the distance, and thus analyzing respiration patterns from close distances is easy, but it is extremely hard to do more than two meters away. The approach is further evaluated in a real-world setting in [4].

Other relevant works based on the RGB-D camera include Yu et al. [17] that estimates sleep position and breathing rate, and Al-Naji et al. [18] who recognizes apnea in infants and young children. However, these approaches used simplified techniques that are not suitable for real scenarios (*e.g.*, Al-Naji et al. [18] require the children to sleep without a blanket). Martinez and Stiefelhagen [19] used a compact representation of the depth map to provide summaries for nursing home residents.

III. MATERIALS AND METHODS

In this paper, we propose a non-invasive computer vision approach to capture and classify breathing conditions using a depth camera as a sensor. A block diagram of the proposed system is displayed in Fig. 1.

We collected a dataset from sleep laboratory patients using a 3D camera installed above the bed and a polysomnogram, which was analyzed according to the Manual of the American Academy of Sleep Medicine (AASM) by the sleep laboratory doctors from Thoraxklinik Heidelberg.

We capture depth information as images based on the Bed Aligned Map (BAM) algorithm [20], that generates a depth field aligned to the bed. A depth field can be obtained in the dark, and is robust to changes in camera location, scene illumination, and clothing, thus providing robust data.

Then, we adopt a deep learning architecture to classify breathing events between normal and abnormal breathing, as classified by the clinical experts. Details of the model architectures and strategy for each phase are described in the following subsections.

A. Data collection and specifications

A total of 76 patients were observed in a sleep laboratory, providing 94 records, each representing a night session of 8 hours. We use a custom recording device equipped with an ASUS Xtion 3D camera, *i.e.*, similar to the Kinect v1 sensor. The camera is installed in the ceiling, aimed at the patient. The sensor configuration is illustrated in Fig. 1.

In order to provide a reference for the experiments, we have also collected the polysomnography signal used in the sleep laboratory, taking care of having both synchronized in time. This polysomnogram information was labeled by the sleep laboratory doctors in ThoraxKlinik Heidelberg, as is normally performed in sleep studies (AASM 2.5). Those manually annotated labels are also available to us and synchronized to the camera data.

Our dataset includes 8 categories for abnormal breathing conditions: central apnea, obstructive apnea, mixed apnea, undefined hypopnea, obstructive hypopnea, central hypopnea, Cheyne-Stokes respiration, and respiratory event related arousal. However, Cheyne-Stokes respiration and respiratory event related arousal events comprise less than 0.01% of all recorded events, thus those categories were not considered in this study, and will need to be targeted by specifically designed studies.

B. Capturing depth information as images

To extract features related to the patients' motion, we first define the Region of Interest (RoI) that contains the patient. This ensures that the majority of depth maps used in the kinematic analysis are consistent and come from the patient and not from other objects also visible in the videos. We perform object boundary detection using the BAM algorithm, which is robust to self-occlusion and articulation of the bed. This process converts each depth map into a BAM representation, which is a normalized depth map, bordered by the length and width of the bed. The map is divided into 10cm $\times$ 10cm tiles, where each tile is assigned to the average height of its local region minus the height of the bed.

Through the subdivision into tiles, we achieve a heavy feature size reduction. Hence, while the depth map obtained by our camera has a resolution of 640×480 pixels, the BAMs in our dataset have a resolution of only 40×26 cells. Furthermore, we use a sampling rate to 5 frames per second.

We preprocess further the BAMs before feeding it into our deep learning model with the following specifications: (1) All BAMs are normalized through mean subtraction, and division by standard deviation. (2) Depending on the bed size, the size of the original BAMs can differ between different records. To ensure a consistent spatial input size, we zero-pad all BAMs to the maximum size of 40×26. (3) We augment the data by applying 5 different crops on the zero-padded BAMs: central, upper left, upper right, lower left, and lower right. Each crop leads to a size reduction from 40×26 to 38×24. Additionally, for each crop, the BAMs are flipped around their vertical axis. In total, we increase the number of input samples by a factor of 10.

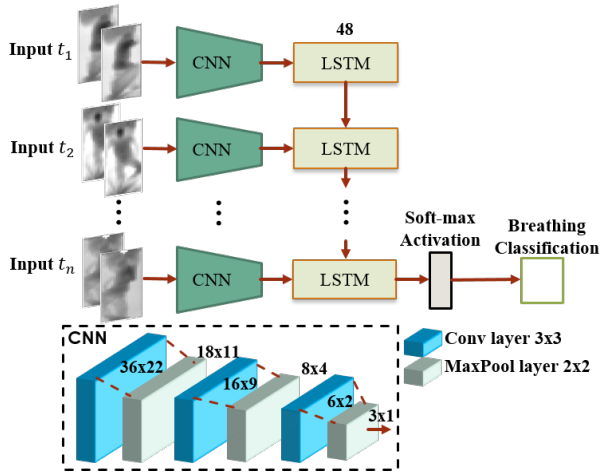

Fig. 2. An end-to-end CNN-LSTM architecture is designed and trained for the identification of breathing conditions using sequences of depth images. The CNN architecture is used to extract spatial features which is followed by an LSTM architecture to extract temporal features. Classification is performed using a densely connected layer with a soft-max activation.

C. Deep learning architecture

Human respiration is directly connected to the movement of the chest when the air is inhaled and exhaled. When capturing depth images, this movement is encoded through a change in depth over time. Consequently, we quantify a sequence of subsequent frames to detect specific breathing patterns that let us classify different breathing conditions.

We adopt a cascade network to first extract discriminative representations from static images using CNNs, and then input these features to sequential networks such as an LSTM for the computation of temporal features [21]. Through extensive experiments, we explore different design choices for our model. The network architecture that shows the best performance is displayed in Fig. 2.

The CNN architecture contains three convolutional layers all with 16 filter kernels of size 3×3 with a stride of 1 and ReLU non-linearity activation functions, each followed by a batch normalization layer. After each convolutional layer, the dimension is reduced through 2×2 max pooling with a stride of 2. The CNN output is subsequently fed to an LSTM with a single layer of 48 units. NormStabilizer layers are placed behind every LSTM to stabilize activations [22]. Finally, the output of the recurrent layer is fed into a densely connected layer with a soft-max activation function to identify abnormal or normal breathing condition.

IV. EVALUATION

A. Experimental setup

We split the dataset patient-wise into a training set (60 patients) and a test set (16 patients). Each patient was monitored over one or two nights. To balance the data per patient, we first set the number of samples per patient to a task-specific value. When there exists more than one record for a patient, the drawing of samples is equally distributed to all available records. After the balancing process, the reminder class imbalance can be addressed by applying class weights on the gradient update. Common values for the class weights are the inverted frequency of the class appearance in training. The frequency can be estimated by averaging the appearance frequency of the class samples in some trial runs.

We train the CNN-LSTM network by optimizing the stochastic gradient descent (SGD) and a learning rate of 0.001 on mini-batches of size 30 and 300 samples per patient. To diminish the classification towards one class, we try out class weights on the gradient updates with weights as high as the inverted frequency of the classes in the training set. The described framework is implemented with the machine learning framework Torch v7 [23]. The weight initialization scheme from LeCun et al. [24] is used for all layers, which is the default one in Torch.

B. Experimental results

The framework takes a sequence of 101 BAMs (approximately 20 seconds) as input and classifies it as a normal or abnormal breathing condition. The model reached an overall accuracy of 61.87%. Fig. 3 shows the resulting confusion matrix. To provide another deep learning baseline for the breathing analysis, we build up a 3D-CNN architecture based on widely used architectures for action recognition [25]. The designed architecture with a stack of 5 convolutional and 5 max-pooling layers reduces the input size from $101 \times 38 \times 24$ to a final output size of $2 \times 2 \times 2$. Throughout the network, we compute 64 feature maps. However, experimentally, we prove that this architecture shows slightly worse performance and a steady increase of the validation loss as depicted in Fig. 4. Therefore, we adopt the CNN-LSTM approach over the 3D-CNN approach.

C. Discussion

A more fine-grained differentiation between the breathing patterns has not shown to be manageable by our proposed model. Nevertheless, the results are promising by eliminating the need for feature engineering and managing highly complex clinical situations. Other marker-free studies that record their data from some distance, require a frontal view on the chest to recognize breathing patterns, and then use hand-crafted features to make a prediction. Instead, we deal with arbitrary sleep positions, the usage of blankets, and only rely on automatically learned features, extracted from the depth maps (BAMs). A general problem in the area of sleep-related breathing disorders is a consistent definition of the disorders. The most common guideline [26], published by the American Academy of Sleep Medicine (AASM), states a minimum breathing pause of 10 seconds for an apnea event.

The main problem in training has been overfitting. Throughout our experiments, common regularization strategies, like dropout or weight decay, have not shown to work. In the best case, they slow down convergence, and regularly lead the training to collapse. Instead, batch normalization [27] has shown better performance which is what we adopted in the proposed model. An additional problem of our automated approach is related to the manual correction of the annotation in the data, *e.g.* when the patient is not actually in the bed because of visits to the toilet, thereby triggering

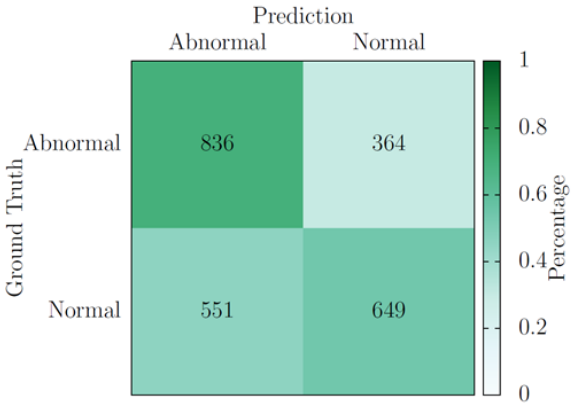

Fig. 3. Confusion matrix when we differentiate between normal and abnormal events. It states a classification accuracy of 61.87%.

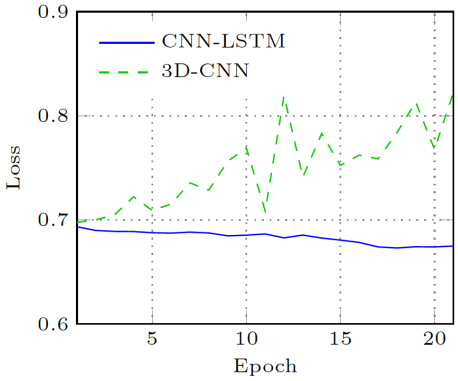

Fig. 4. Loss progress on the validation set for the CNN-LSTM and 3D-CNN architecture.

the need to incorporate human detection techniques in the system.

We argue that the automatic identification of breathing disorders enables more objective information to support the evaluation of these conditions. Our system is a novel approach based on computer vision and deep learning to take on the complex nature of breathing monitoring. Reasonable performance on heavily compressed input data is reached without the use of hand-crafted feature engineering. This opens up new opportunities in this research direction, such as evaluating the computational cost and performance of using higher imaging resolution as input, and by capturing depth information only from the automatic detection of the chest area.

V. CONCLUSIONS

We have presented a computer vision approach to capture motion in order to evaluate breathing conditions. This paper presents the first quantitative representation of the evolution of breathing following a deep learning approach under challenging natural clinical settings. The simplicity of our methodology is a promising baseline for assistive medical diagnosis based on the monitoring of patients' behavior. This work is a completely novel method, unreported in the literature, to attempt to tackle a highly complex area through further research.

ETHICS STATEMENT

The experimental procedures involving human subjects described in this paper were approved by the Ethical Commission of the Medical Faculty of Heidelberg.

REFERENCES

[1] J. M. Marin, S. J. Carrizo, E. Vicente, and A. G. Agusti, "Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study," *The Lancet*, 2005.

[2] V. Kapur, K. P. Strohl, S. Redline *et al.*, "Underdiagnosis of sleep apnea syndrome in us communities," *Sleep and Breathing*, 2002.

[3] S. Srivastava, S. Soman *et al.*, "Deep learning for health informatics: Recent trends and future directions," in *ICACCI*, 2017.

[4] M. Martinez and R. Stiefelhagen, "Breathing rate monitoring during sleep from a depth camera under real-life conditions," in *WACV*, 2017.

[5] H. Alshaer, "New technologies for the diagnosis of sleep apnea," *Current hypertension reviews*, 2016.

[6] L. Samy, M.-C. Huang, J. J. Liu, W. Xu, and M. Sarrafzadeh, "Unobtrusive sleep stage identification using a pressure-sensitive bed sheet," *Sensors*, 2014.

[7] K. Nakajima, Y. Matsumoto, and T. Tamura, "Development of real-time image sequence analysis for evaluating posture change and respiratory rate of a subject in bed," *Physiological Measurement*, 2001.

[8] J. G. Chase, F. Agogue *et al.*, "Quantifying agitation in sedated icu patients using digital imaging," *Computer methods and programs in biomedicine*, 2004.

[9] M. Reyes, J. Vitria, P. Radeva, and S. Escalera, "Real-time activity monitoring of inpatients," in *MICCAT*, 2010.

[10] W.-H. Liao and C.-M. Yang, "Video-based activity and movement pattern analysis in overnight sleep studies," in *ICPR*, 2008.

[11] Z. Zhu, J. Fei, and I. Pavlidis, "Tracking human breath in infrared imaging," in *BIBE*, 2005.

[12] H. Aoki, Y. Takemura *et al.*, "Development of non-restrictive sensing system for sleeping person using fiber grating vision sensor," in *Micromechatronics and Human Science*, 2001.

[13] ——, "A non-contact and non-restricting respiration monitoring method for a sleeping person with a fiber-grating optical sensor," *Sleep and Biological Rhythms*, 2003.

[14] Y. Takemura, J. Sato, and M. Nakajima, "A respiratory movement monitoring system using fiber-grating vision sensor for diagnosing sleep apnea syndrome," *Optical review*, 2005.

[15] H. Aoki, M. Miyazaki *et al.*, "Non-contact respiration measurement using structured light 3-d sensor," in *SICE*, 2012.

[16] M. Martinez and R. Stiefelhagen, "Breath rate monitoring during sleep using near-ir imagery and pca," in *ICPR*, 2012.

[17] M.-C. Yu, H. Wu, J.-L. Liou *et al.*, "Multiparameter sleep monitoring using a depth camera," in *Biomedical Engineering Systems and Technologies*, 2012.

[18] A. Al-Naji, K. Gibson, S.-H. Lee, and J. Chahl, "Real time apnoea monitoring of children using the microsoft kinect sensor: a pilot study," *Sensors*, 2017.

[19] M. Martinez and R. Stiefelhagen, "The sphere project: Sleep monitoring using computer vision," in *Forum Bildverarbeitung 2016*, 2016.

[20] M. Martinez, B. Schauerte, and R. Stiefelhagen, "BAM! depth-based body analysis in critical care," in *International Conference on Computer Analysis of Images and Patterns*, 2013.

[21] J. Donahue, L. Anne Hendricks, S. Guadarrama *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015.

[22] D. Krueger and R. Memisevic, "Regularizing rnns by stabilizing activations," *arXiv preprint arXiv:1511.08400*, 2015.

[23] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, NIPS*, 2011.

[24] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*. Springer, 2012.

[25] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *CVPR*, 2015.

[26] R. B. Berry, R. Brooks, C. E. Gamaldo *et al.*, "The AASM manual for the scoring of sleep and associated events," *Rules, Terminology and Technical Specifications*, 2012.

[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.