Transfer Learning for Clinical Sleep Pose Detection Using a Single 2D IR Camera

Sara Mahvash Mohammadi, Shirin Enshaeifar, *Member, IEEE*, Adrian Hilton, *Member, IEEE*, Derk-Jan Dijk, and Kevin Wells

***Abstract*— Sleep quality is an important determinant of human health and wellbeing. Novel technologies that can quantify sleep quality at scale are required to enable the diagnosis and epidemiology of poor sleep. One important indicator of sleep quality is body posture. In this paper, we present the design and implementation of a non-contact sleep monitoring system that analyses body posture and movement. Supervised machine learning strategies applied to noncontact vision-based infrared camera data using a transfer learning approach, successfully quantified sleep poses of participants covered by a blanket. This represents the first occasion that such a machine learning approach has been used to successfully detect four predefined poses and the empty bed state during 8-10 hour overnight sleep episodes representing a realistic domestic sleep situation. The methodology was evaluated against manually scored sleep poses and poses estimated using clinical polysomnography measurement technology. In a cohort of 12 healthy participants, we find that a ResNet-152 pre-trained network achieved the best performance compared with the standard *de novo* CNN network and other pre-trained networks. The performance of our approach was better than other video-based methods for sleep pose estimation and produced higher performance compared to the clinical standard for pose estimation using a polysomnography position sensor. It can be concluded that infrared video capture coupled with deep learning AI can be successfully used to quantify sleep poses as well as the transitions between poses in realistic nocturnal conditions, and that this non-contact approach provides superior pose estimation compared to currently accepted clinical methods.**

***Index Terms*— Pose detection, convolutional neural networks (CNN), sleep, transfer learning, polysomnography (PSG).**

I. Introduction

SLEEP plays an important role in physical and mental health and is a major determinant of well-being [1]. Sleep disorders are prevalent across the life-span, increase with aging, and contribute to neurodegeneration [2]. Sleep disorders are frequently under-diagnosed. Methodologies to quantify sleep can be classified into two broad categories: in the first, physiological variables are monitored during sleep and used for sleep stage classification and sleep quality assessments [3], [4]; the second category of methodologies examines individuals' external body characteristics during sleep, by pose and movement detection [5], [6]. During a sleep episode, periods of immobility are interspersed with movements that may or may not lead to a change in body position or sleep pose and may also be associated with brief awakenings [7]. Body movements during sleep and brief awakenings are directly related to the perceived quality and depth of sleep [8].

Some sleep disorders, such as periodic limb movement disorder or rapid eye movement (REM) sleep behavior disorder are characterized by major or minor-movements [9]. Sleep-disordered breathing is modulated by body position such that it is more severe in the supine position [10].

Recent research has focused on non-contact methodologies to measure external body characteristics to determine body positions and detect movement during sleep [5], [6], [11].

Polysomnography (PSG) is the gold standard method for assessing sleep and encompasses electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), and other measurements such as breathing-related variables and a body position signal. Although PSG offers vital information, it is an expensive, inconvenient, and time-consuming approach that requires many sensors to be attached to the participant which may, in themselves, impact on sleep quality of the participant under investigation. PSG recordings are usually obtained in a dedicated sleep laboratory which is a high-cost facility that may not be widely accessible. The unfamiliar sleep laboratory environment may also reduce the quality of sleep. These characteristics prevent PSG from being used for long-term monitoring, at large scale, or in populations that do not tolerate the attachment of sensors.

Body position is recorded in standard PSG by a position sensor that is placed on the participants' abdomen using a belt. However, the current standard manual or automated scoring of PSG does not include a comprehensive quantification of body position during sleep [12].

Quantitative non-contact sleep monitoring methods represent a potential approach to address these issues and could eventually be used to diagnose sleep disorders, improve sleep-interventions, and thereby increase quality of life.

Manuscript received September 12, 2020; revised December 17, 2020; accepted December 27, 2020. Date of publication December 30, 2020; date of current version March 1, 2021. This work was supported in part by the U.K. Dementia Research Institute (DRI) which receives its funding from DRI Ltd., through the U.K. Medical Research Council, the Alzheimer's Society, and the Alzheimer's Research U.K. *(Corresponding author: Sara Mahvash Mohammadi.)*

Sara Mahvash Mohammadi, Shirin Enshaeifar, Adrian Hilton, and Kevin Wells are with the Centre for Vision, Speech and Signal Processing, Faculty of Engineering and Physical Science, University of Surrey, Guildford GU2 7XH, U.K., and also with the UK Dementia Research Institute, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: s.mahvash@surrey.ac.uk; k.wells@surrey.ac.uk).

Derk-Jan Dijk is with the Surrey Sleep Research Centre, Faculty of Health and Medical Sciences, University of Surrey, Guildford GU2 7XH, U.K., and also with the UK Dementia Research Institute, University of Surrey, Guildford GU2 7XH, U.K.

Digital Object Identifier 10.1109/TNSRE.2020.3048121

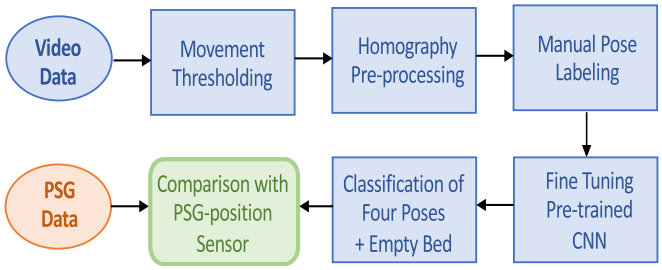

Fig. 1. Schematic diagram of the proposed methodology. Starting with 2D IR video camera data a block matching algorithm is used to determine major and minor-movements, representing pose changes, and intra-pose motion, respectively. The data are then transformed using homography to a common viewpoint to assist machine learning across different participants. Video data are hand-labeled for one of four poses plus the empty bed state representing ground truth prior to machine learning using fine-tuning of a pre-trained CNN network. Classification is then assessed and compared to manual labels using a leave-one-out cross-validation strategy. Performance is assessed by comparison to manual labeling and pose estimation data obtained from standard clinical PSG equipment using various metrics and a Markov Chain transition analysis.

Non-invasive methods such as camera-based techniques represent a convenient, low-cost approach to monitoring sleep behavior. Therefore, in this study, a camera-based method using a single infrared (IR) camera is used to record sleep behavior in a cohort of 12 healthy participants during 8-10 hour overnight sleep periods alongside PSG. The overall strategy of this work is illustrated in Fig. 1. Such video data can be analyzed using a supervised deep learning (DL) methodology in specific convolutional neural networks (CNNs) to classify images of sleep behavior in a set of accepted predefined sleep poses: supine, left, right, prone as well as the null pose herein referred to as the 'empty bed' state. Applications of CNNs conventionally encompass image processing as well as techniques drawn from the computer vision domains. For example, CNNs have successfully demonstrated superior performance within the realm of image classification when compared to humans [13], [14]. In this paper, we focus on CNN algorithms for the classification of five discrete states that represent four pre-defined sleep poses and the empty bed state. Instead of developing the CNN model *de novo*, a pre-trained CNN model can significantly reduce the training time and requires significantly less training data. In this paper, various CNN architectures pre-trained on the Imagenet dataset [15] are used for the classification task of sleep poses and compared to *de novo* performance of a 4-layer dedicated CNN architecture. The output of the best pre-trained networks for classifying the five states was compared with the states detected using a standard position sensor provided by a clinical PSG belt sensor. Post-acquisition manual annotation of body position by visual inspection of the video was considered as ground truth. In addition to the detection of sleep poses we also compared the performance of the various methodologies in quantifying the transitions between poses by applying Markov chain analysis.

The major contributions of this paper are as follows: 1. First demonstration of using a single 2D IR video camera and CNN-DL for authentic clinical standard sleep-pose classification: supine, left, right, and prone. This approach is also capable of monitoring minor within-pose movements, empty bed states, and other anomalous states. Moreover, this is the first time that different pre-trained CNN networks have been used for standard sleep-pose estimation, demonstrating superior performance compared to a dedicated 4-layer *de novo* network. 2. First demonstration of using a single 2D IR video camera and CNN-DL for clinical standard sleep pose estimation during realistic sleep conditions, robust to body occlusion by blankets, variable illumination, and camera viewpoint. 3. With an accuracy of 95.1% for pose classification during sleep, this represents state-of-the-art performance for video-based non-contact sleep pose estimation and this performance is better than the clinical standard for pose estimation using a PSG system (88.2%).

II. Prior Work

There are several reports on non-contact methodologies for estimating pose during sleep. These methodologies fall into two main categories: instrumented beds (covering instrumentation applied to the mattress, the bed frame, and the pillow) and video-based monitoring approaches.

Within the first category, Hoque *et al.* suggested using RFID (radio-frequency identification)-based 3-D accelerometers tagged to the bed legs to monitor four different sleep poses: supine, left, right, prone and, empty bed [6]. An average accuracy of 90.0% was achieved for one subject over six nights. Another study classified six common body poses using a pressure-sensitive bedsheet utilizing high-resolution textile pressure sensors [16]. Adami *et al.* [17] presented an approach using load cells under the bed for unobtrusive continuous monitoring of sleep patterns. The system was capable of classifying sleep poses (supine, left, & right) with a correct classification rate of 90.82 % and to analyse in-bed and out-of-bed events. The study in [18] claimed that the morphology of the human QRS (Q wave, R wave, and S wave of Electro-Cardio-Gram (ECG)) ECG complex changes with different body poses. Four body poses therefore using the ECG signal extracted from the sensors attached to the conductive textile sheet and an accuracy of 98.4% was obtained for simulated study over 13 participants. Instrumented mattress-based methods are effective at localising areas of increased pressure and can automatically classify sleep poses, but the relatively high cost of the pressure sensor array has prevented this solution from achieving large-scale uptake. These sensors when attached to either the mattress or pillow may also lead to discomfort and thereby affect sleep quality [19].

Visual sensing through video cameras is one of the most popular approaches for human pose estimation due to the low cost of the technology and ease of maintenance. Such approaches have usually harnessed the power of machine learning to analyze the acquired video data. A neural-network approach was employed [19] to recognize simulated body movement and body poses (supine, left, right, & prone) using the features extracted from the image sequences captured with IR camera. In this study the participants were not covered by a bed sheet and therefore has limited application in real-world scenarios. Furthermore, the study did not provide any

quantitative assessment, thus it is difficult to assess the performance. In another approach, a sleep monitoring system was developed to identify movement and six body poses during sleep [5] using x, y and, z positions of 25 body joint skeleton information from a Kinect depth camera. Whilst this is an interesting approach to detect poses, no performance matrix was provided. Furthermore, the system was not able to detect skeleton information when participants were occluded by a blanket and the method was unable to detect the prone position. Recently, frequency-based feature selection was used to extract sleep pose information from depth data in a simulated sleep study in 14 participants [20]. The extracted features were used to train a support vector machine (SVM) for the two-class problem of supine and side-lying with and without the presence of a bed covering such as a blanket. Another study [21] also solved a two-class problem of supine and side-lying based on depth camera technology analysis. This study used a cross-section method to localise participant's head and torso from the sequence of depth images. The algorithm was evaluated in eight participants in a simulated sleep experiment and achieved an accuracy of 97.0% for classifying two poses. Liu and Ostadabbas [22] proposed a computer vision-based method for predicting the sleep positions of hospital patients using a standard video camera in a simulated scenario. The histograms of oriented gradients (HOG) were used to extract features from the images and were fed into the SVM classification to detect three sleep poses: supine, left, and right. One of the major challenges reported by the authors was that their method was ineffective for subjects covered by a sheet or blanket. Also, regular video cameras cannot be applied in a low light sleep environment. To address the problem of variations in lighting, a more recent solution used a near-infrared modality [23]. The researchers fine-tuned the convolutional pose machine (CPM) [24] as a pre-trained network for classification of supine, left, and right. Although the CPM obtained good performance for pose tracking (accuracy = 86.7%), the study was performed on a mannequin which could not be considered as representative of real human pose variation and also the method assumes that there is no occlusion of the participant. Recently, simultaneous analysis of respiration, head posture, and body posture was presented [25]. In this study, a Kinect motion sensor was utilized to obtain a skeleton description of seven individuals simulating three body poses. Machine learning techniques were then used for classification of supine, left, and right sleep poses. However, the proposed study was unable to distinguish between the prone and supine sleeping pose. Also, it was not possible to use it for real sleep scenarios when poses are occluded by a blanket. The current state-of-the-art in performance for video-based pose classification is described in [26]. The authors used a dimensional reduction technique known as bed aligned maps from depth images. They employed a CNN network for classifying sleep poses including supine, left, right, and empty bed for 78 patients and achieved an accuracy of 94.0%. This was a mixed dataset with some patients with and some without occlusion which will improve the classification accuracy as it is much easier to classify any pose without a blanket. Moreover, the test data were down-sampled from 65 million images to 1880 and the prone position was also removed from the dataset for the analysis.

Among these technologies, non-contact video-based methods have been demonstrated some of the greatest promise. One of the limitations of prior work in this area is that they are not validated against the gold standard or clinical methods. Moreover, many of these prior approaches have either used simulated sleep rather than real sleep or require subjects to avoid any occlusion from bedding. Therefore, it is difficult to evaluate how well such methods can operate in a routine sleep situation as they do not represent the true behaviour of human sleep patterns. These aspects constitute significant limitations for the routine implementation of video monitoring to measure body position and movements during sleep. This paper addresses these limitations by using data captured from actual sleep, with participants using a blanket bed covering and comparing the approach to clinical standard methods and gold-standard manual annotation of the video data.

III. Materials and Methods

The study aimed to develop a system that can identify sleep poses, and that is robust to bed covering and camera orientation, and performs as well as the current clinical standard. Data were collected during an overnight 8-10 hour sleep study at the Surrey Sleep Research Centre (SSRC). The study was designed to demonstrate the proof of concept of measuring sleep poses using IR camera data as input to a pose estimation method that is robust to bed covering and orientation. The study was completed in 12 participants and the performance of the new system was compared to body position as detected by a body position sensor which is part of the PSG set up. The performance was also compared to manual scoring of body position by visual inspection of the video-PSG. Manual scoring of the poses observed on video playback was considered the ground truth.

A. Data Collection

Participants were recruited from the University of Surrey and the general public.[1] In total, 12 healthy participants (five females and six males, aged between 18 and 65 years) were screened for eligibility[2] and were enrolled in the study at SSRC. All participants provided written informed consent before participation. Participants were fitted with the electrodes and sensors of the SomnoHD PSG system (Somnomedics Gmbh, Germany). The SomnoHD is a wireless PSG system approved and validated for sleep medicine and stores data locally to a memory card and data are also sent to a remote monitoring station. Somnomedics uses a 3-axis micro electro mechanical system (MEMS) accelerometer to record body poses during the sleep period with a sampling rate of once every 30-seconds. The position sensor was attached to the participants' abdomen using a belt. The position sensor captures four basic sleep poses: supine, left, right, & prone.

[1]This study, EGA application no 'UEC 2018 051 FEPS' was submitted to the University of Surrey Ethics Committee for ethical review on 30/05/2018 and granted a favourable ethical opinion on 29/06/2018.

[2]Inclusion: Healthy male and female; Aged over 18 years Exclusion: Diagnosed with a previous sleep disorder; Known pregnant woman

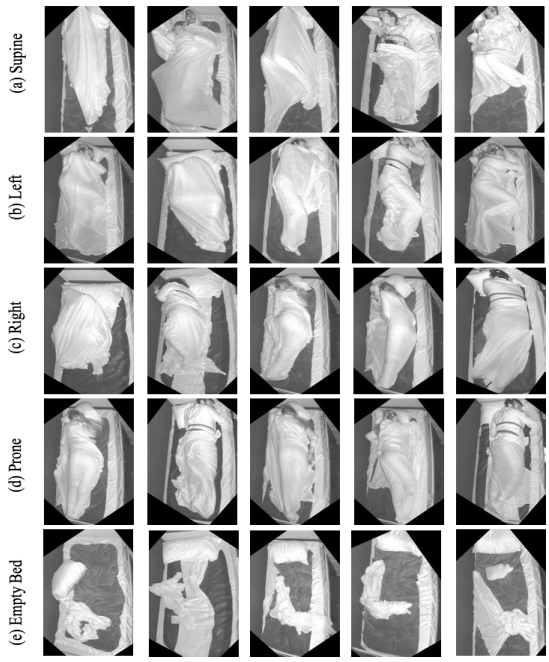

Fig. 2. Examples of poses in different participants: (a) supine, (b) left, (c) right, (d) prone, & (e) empty bed. Note that on some occasions some of the participants were observed sleeping only partly, and very occasionally not at all, covered by the blanket.

In cases where the participants left the bedroom, for example for a toilet break, the sensor scores the pose as an artifact due to losing connection with the recording unit which was located at the back of the bedroom. We assign these events to the 'empty bed' state. Other sensors such as EEG, ECG, and other physiological signals were also acquired for subsequent analysis. In the standard PSG acquisition, IR video recording is also included. The IR video was captured with a Somnomedics active IR camera system. The camera was set to take images at 25 frames per second (FPS) thus providing approximately 900,000 frames over the 8-10 hour recording period for each participant. Participants slept in individual, sound attenuated, temperature-controlled, windowless bedrooms at the dedicated Sleep Laboratory within the University of Surrey's Clinical Research Facility. We recorded sleep for up to 10 hours because longer than habitual sleep periods are characterized by more awakenings and movements and thereby provide a good model to test systems for monitoring of undisturbed and disturbed sleep. Fig. 2 presents examples of each of the five pre-defined states (four poses + empty bed) from different participants. This illustrates the diversity of pose appearance across the dataset during actual sleep, in contrast to the limited variation seen in simulated approaches with mannequins or compliant volunteers [27].

B. Frame Extraction & Movement Detection

For movement detection and position tracking, frames were extracted from the raw video data. To make the analysis more tractable and to reduce the computational cost of the algorithm downsampling was undertaken. The temporal resolution of the extracted frames was reduced by a factor two. Next frames were classified into either no-movement, minor-movement, or major-movement using a motion estimation and thresholding algorithm based on block-matching and decision tree [28]. In this study body movements were grouped into two main classes:

Class 1 (Major movement): Motion in the body's torso such as changing body pose; Getting in and out of the bed.

Class 2 (Minor movement): Any minor movement of a body part such as re-positioning the arm or head but which did not result in a change of overall pose.

C. Block Matching Algorithm

A block-matching algorithm was used as a motion estimation algorithm to detect movement. Its ease of use along with minimum computational costs made this approach very effective. After applying the block-matching algorithm, we employed a decision tree classifier to automatically predict the threshold for static and dynamic states. To generate the training samples of the decision tree, we selected 800 frames from each participant which included all the aforementioned motion classes, i.e. no-movements, minor-movement, and major-movement. Two sets of thresholds were defined based on visual inspection of the block-matching output for each participant. One for distinguishing the static and dynamic episodes. Another threshold was also set to divide movement episodes into classes 1 and 2 i.e., minor-movement and major-movement. Thus, using the two sets of thresholds, class labels were derived. Any movement below $threshold_1$ is a no-movement and any movement between the two sets of thresholds is a minor-movement and finally, any movement greater than $threshold_2$ considered to be major-movement. The corresponding block-matching vector was used as input variables for constructing the decision tree using the class labels generated. A leave-one-subject-out cross-validation was conducted to examine the misclassification rate (MCR) of the decision tree classifier. The decision tree predicts the thresholds with an averaged MCR error of 0.0046. Fig. 3 represents an example of a block-matching output and the predicted thresholds through the decision tree algorithm for one of the participants. In this study, the block-matching algorithm combined with decision tree was used to automatically detect movement. This resulted in every raw frame of video data being tagged with a timestamp and labeled in one of three classes: no-movement, minor-movement, and major-movement. Therefore episodes of no-movement and minor-movement carried the same state and those frames involving major-movement represent transitions between sleep posed.

D. Manual Scoring

Video recordings and sensor recordings were synchronised so that timestamps for each pose could be identified and scored in conjunction with the standard PSG-position sensor. Ground truth labels for sleep pose, lighting and motion were annotated for each frame using manual expert observation (Table I). Label #6 was given to frames with major movement, indicating a change of pose. Label #7 was assigned to frames where the lights were on. Note, labels #6 and #7 were removed from the dataset for the purposes of training and test. All other image

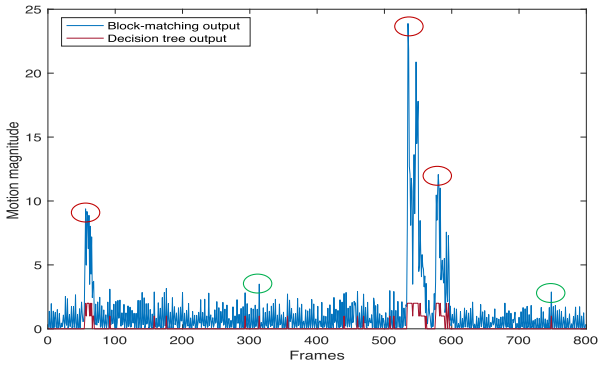

Fig. 3. The blue graph represents the motion magnitude output of the block-matching algorithm during a period which contains all three events: no-movements, minor-movement, and major-movement. The red graph shows the predicted thresholds using the decision tree algorithm where 0 label corresponds to no-movement, 1 to minor-movement, and 2 to major-movement frames. Examples of major-movement and minor-movement are represented by red and green circles respectively.

TABLE I
GROUND TRUTH LABELS NUMBER (#) AND THEIR CORRESPONDING STATUS & DETAILS USED FOR MANUAL ANNOTATION

Label	Status	Details
#1	Supine	Still image
#2	Left	Still image
#3	Right	Still image
#4	Prone	Still image
#5	Empty Bed	Empty bed
#6	Change Position	Major-movement
#7	Lights	Lights on
#8	Supine & Movement	In supine with minor-movement
#9	Left & Movement	In left with minor-movement
#10	Right & Movement	In right with minor-movement
#11	Prone & Movement	In prone with minor-movement

frames containing no-movement or minor movement were then manually labelled by selecting a small number of frames from the period of consistent pose and to determine a label for this consistent pose interval. No-movement frames were assigned labels #1 to #5 (supine, left, right, prone, empty bed) based on the specifications in Table II. Two dimensions were considered in defining poses including the upper and lower body. Regarding the frames containing minor movement, they were labelled against #8, #9, #10, & #11 in case participants were in supine, left, right, & prone respectively.

E. Train and Test

In order to prepare the data for training and testing of the CNN, the raw data needed to be prepared. Most of the frames consist of still images without any detectable changes in pose. To train a robust CNN, we need to provide significant variation in the image content for a given class label. Therefore, only a small selection of no-movement frames was included in the training set, alongside any frame labeled with minor-movement to maximise the variance across the training data. Across all participants, states supine, left, right, & prone were the most common positions. To have a balance in the training set, data downsampling was applied to each state based on its proportion. To this end, only the first 2, 3, 5, & 8 frames of label #1, #2, #3, & #4 respectively were considered. All frames with label #8, #9, #10, & #11 were

TABLE II
DEFINITIONS USED FOR MANUAL ANNOTATIONS OF THE FOUR BODY POSITIONS (SUPINE (S), LEFT (L), RIGHT (R), PRONE (P)) AND EMPTY BED (E) OF MANUAL ANNOTATION

(S)	Upper	Lying on back (the angle between the bed surface and the shoulder to shoulder line is less than 30°) with both arms straightened, bent, or one arm straightened and one bent.
	Lower	Both legs straightened, bent, one leg straightened and one bent, both legs towards left, or right.
(L)	Upper	Lying on the left hand with both arms straightened, bent, or one arm straightened and one bent.
	Lower	Lying on the left legs with both legs straightened, bent, or one leg straightened and one bent.
(R)	Upper	Lying on the right hand with both arms straightened, bent, or one arm straightened and one bent.
	Lower	Lying on the right legs with both legs straightened, bent, or one leg straightened and one bent.
(P)	Upper	Lying on stomach (the angle between the bed surface and the shoulder to shoulder line is less than 30°) with both arms straightened, bent, or one arm straightened and one bent.
	Lower	Both legs straightened, or one leg straightened and one bent.
(E)	–	When participants were out of bed for the toilet break.

assigned to state supine, left, right, & prone, respectively. For the frames in the category of empty bed state, only the first 10 frames of label #5 were selected in order to sample different configuration of the pillow and bed. The leave-one-subject-out cross-validation approach was used to generate training and testing sets. Therefore, for training the network, in this case, eleven (i.e. n-1) cases were used to train the network and the remaining subject dataset was then used for the test. For each participant on the test, all data, which consist of approximately 90,000 frames for the entire night of sleep, were included in the test data. The process was repeated 12 times so that the network had been trained, using transfer learning, on 12 separate occasions with each fold of the training and validation data removing one participant that was kept as a unique test case for that fold. At each fold 20% of the training data was used as validation data to set the network's parameters.

F. Image Transformation Using Homography

The camera position and orientation in each of the six bedrooms was different and created a different view of the bed. To prevent over-training and ensure the CNN trained on the pose variations in image content rather than background geometry, we aligned all images with different camera views to the same reference using homography [29]. Each IR pixel from the captured camera view was mapped into a virtual camera view using homography as illustrated in Fig. 4. Homography is a linear geometric transformation that connects two images of the same plane. The relationship between two planes can be represented by a 2D projective transformation, abbreviated as H [29]:

$$X_t = H X_s \tag{1}$$

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \tag{2}$$

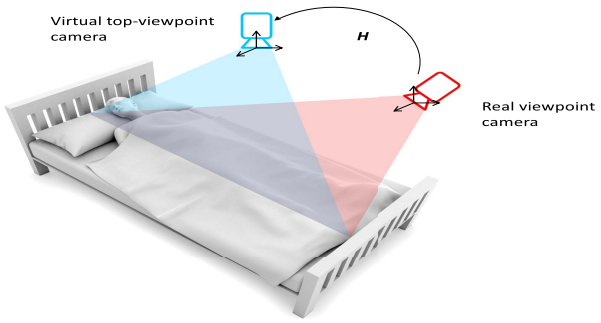

Fig. 4. The installed camera position at the sleep lab and the transformed virtual camera view from the top generated through homography.

where $X_t = [u_t, v_t, 1]^T$ and $X_s = [x_s, y_s, 1]$ are the homogeneous coordinates of a pair of corresponding points in the target and source images, respectively.

G. Data Augmentation

Due to the small size of our training set with frames that illustrated the prone and empty bed state, a data augmentation method was applied to boost the multi-classification performance and resolve the imbalanced class problem. Augmentation was applied by randomly rotating input images in the range $[0, \pi/6]$ and randomly translating them horizontally and vertically in the range $[0, 45]$ pixels, to reduce sensitivity to intra-pose orientation. This was applied to prone and empty bed states for all participants increasing the training set from 85083 to 89731 frames representing an increase of 6%. Particularly, it boosted the number of prone and empty bed states from 2068 to 6204 frames and from 256 to 512 frames respectively. To avoid border effects, all the images were cropped and scaled to 400×700 pixels after performing the augmentation so that a bounding box was shaped around the body shape and bed.

H. Supervised Classification Approach

1) De novo CNN: A 4-layer CNN network encompassing four convolutional layers (C1-C4) with rectification (ReLU), 2×2 max pooling (MP1-MP3), and batch normalization, followed by two fully connected layers and a softmax layer was trained from scratch. For training the network a stochastic gradient descent was used as an optimiser with hyperparameters initially sampled from a Gaussian with zero bias and the learning rate was set to 0.0001. The *de novo* CNN was trained using a leave-one-subject-out methodology with 12-way cross-validation.

2) Transfer Learning: Transfer learning is a machine learning method that can transfer the knowledge learned from one task in one field to a different task in another field. By the use of transfer learning in the CNN model, we propose that knowledge learned in identifying everyday objects within ImageNet database [15] can be used in our task of sleep pose classification. This also addresses the issue of limited training images to train the CNN model. The CNN model is usually divided into two parts: the feature extractor (representing shallower layers: convolutional, pooling and rectification layers) and the fully connected classifier layer. The shallower layers of the pre-trained model learn the basic low-level general features of the images which may be applicable to a variety of different vision tasks including pose estimation during sleep while features from deeper layers (e.g. the fully connected layer) are more abstract and task-specific. Therefore, in this study, we retained all weights of convolutional layers and reinitialise the weights of fully connected layers.

TABLE III
PRE-TRAINED NETWORK SPECIFICATIONS

Models	# of layers	Input size	Model description	Replaced layers
AlexNet	8	227x227	5conv+3fc	fc8, prob
VGG-16	16	224x224	13conv+3fc	fc8, prob
VGG-19	19	224x224	16conv+3fc	fc8, prob
GoogLeNet	22	224x224	21conv+1fc	softmax, cls1-fc2, cls2-fc2, cls3-fc
ResNet-50	50	224x224	49conv+1fc	fc1000, prob
ResNet-101	101	224x224	100conv+1fc	fc1000, prob
ResNet-152	152	224x224	151conv+1fc	fc1000, prob

In this study AlexNet [30], VGG-16, VGG-19 [31], GoogLeNet [32], ResNet-50, ResNet-101, & ResNet-152 [33] have been explored to examine the variation in performance of different CNN architectures for the transfer learning task of sleep pose estimation. A summary of the specification and description of these networks are presented in Table III. These convolutional networks are pre-trained on the ImageNet dataset (with more than 1.2 million RGB images) of natural objects and the knowledge learned from ImageNet can be used in our task. Fine-tuning begins with transferring the weights from a pre-trained network to the network we wish to train. For all the pre-trained networks, we have frozen the weights of the all convolutional layers and fine-tuned the weights of the last layers. The last layers including fully connected, softmax, and classification layers of the network were removed and replaced with new layers that are relevant to the current sleep pose classification task. For example, a new fully connected layer was added to the networks but with fewer parameters, since we have fewer classes to classify, and initialized with random weights, whereas the rest of the layers were initialized using the weights of the pre-trained networks. During the training process, we only trained the newly added layers with our training data and set the gradient of other layers in the back-propagation process to zero. The networks were trained using a stochastic gradient descent optimisation using a batch size of 32. The learning rate was also empirically set and optimised to 0.0001

I. Quantifying Sleep Behavior Using State Transitions From a Markov Chain

The transitions between sleep poses can be analyzed through a Markov chain probability-based model which provides a view of sleep dynamics. The Markov chain is a stochastic process $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$ that undergoes the transition from one to another state wherein for this work the states are presented as sleep poses and empty bed: $\mathbf{S} = \{s_s, s_l, s_r, s_p, s_e\}$ (where $x_i \in \mathbf{S}$) [34], [35]. Within this regime, we assume that

$$\mathbf{T_x}= \begin{bmatrix} P_{s,s} & P_{s,l} & P_{s,r} & P_{s,p} & P_{s,e} \\ P_{l,s} & P_{l,l} & P_{l,r} & P_{l,r} & P_{l,e} \\ P_{r,s} & P_{r,l} & P_{r,r} & P_{r,p} & P_{r,e} \\ P_{p,s} & P_{p,l} & P_{p,r} & P_{p,p} & P_{p,e} \\ P_{e,s} & P_{e,l} & P_{e,r} & P_{e,p} & P_{e,e} \end{bmatrix}$$

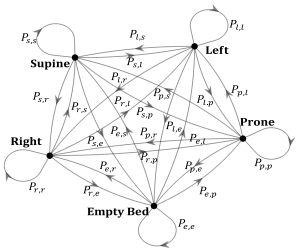

Fig. 5. Generic five state Markov chain model and the corresponding generic transition matrix.

the probability distribution of each x_i depends only on the value of the previous sample and not on the sequence of events that preceded it x_{i-2}, x_{i-3}, etc. State transition probabilities describe the probability of going from x_{i-1} to x_i:

$$P(x_i|x_{i-1}, \cdots, x_1) = P(x_i|x_{i-1}) = P_{i-1,i} \tag{3}$$

Fig. 5 represents the five states Markov chain with a corresponding generic transition matrix. $\mathbf{T}_x$ illustrates the transition probabilities between these five states. Each element of the matrix represents a transition probability, for example, $P_{s,r}$: the probability of transition from supine to right. While the matrix diagonal shows the state maintenance probabilities, where $P_{l,l}$ describes the probability of remaining in the left state. This approach has been utilized here as a way of compactly characterizing individual sleep behavior, which can then be used to globally compare the performance of the DL video scoring sleep pose classification with that of standard pose estimation, using manual scoring of the video as ground truth.

IV. Results & Discussion

The evaluation of supervised classification of sleep poses in our cohort of 12 participants during sleep is divided into three sections: the first section includes analysis of the influence of various CNN architectures for classification of five states using data obtained from a simple 2D IR camera system; the second section compares the performance of DL video scoring and standard PSG-position sensor for body pose and empty bed detection. This represents a direct comparison between pose classification and ground truth in our cohort under the realistic conditions of actual sleep and using a bed covering; the third section evaluates these two approaches to pose classification by comparing their performance in describing individual sleep dynamics, using a Markov-based analysis. This is used to consider the statistical significance of each particular pose transition compared to the ground truth. In this study, body position was manually annotated by visually inspecting the video that was recorded as part of the PSG recording and thereby considered ground truth for our analyses.

For the evaluation of the influence of various CNN architectures on performance, the leave-one-subject-out cross-validation was used to validate different pre-trained DL architectures using a transfer learning paradigm, as well as a 4-layer *de novo* CNN network. To evaluate the performance of the classifier five statistical measures were obtained: accuracy, precision, recall, F1-score, and Cohen's kappa.

Table IV illustrates the performance of AlexNet, VGG-16, VGG-19, GoogLeNet, ResNet-50, ResNet-101, ResNet-152, and a 4-layer *de novo* CNN network trained from scratch, using the afore-mentioned statistical measures. The results in this table show the average performance of 12 participants over five sleep poses. Fine-tuning of all pre-trained networks yields relatively better performance than the standard dedicated *de novo* CNN network and other pre-trained networks. Fig. 6 presents a different, but more informative view of the accuracy values for different pre-trained networks, because it also visualises the depth and the number of network's parameters. The first thing that is very obvious in this graph is that VGG-16 & VGG-19, although widely used in various applications, are by far the most expensive architecture regarding the number of parameters. ResNet-50, ResNet-101, and ResNet-152 architectures in terms of the number of layers are isolated from all other networks. However, the ResNet architecture starts to outperform in accuracy by increasing the number of layers. In Fig. 7 we illustrate these results by selecting accuracy and Cohen's Kappa statistics across the different networks used for analysis. This shows that performance across all networks, apart from GoogLeNet, is fairly similar, but that with ResNet-152 there is a reduced spread in the results and an incrementally better mean accuracy and Cohen's kappa compared to other networks. We, therefore, take ResNet-152 as our preferred network architecture on which further analyses are based.

In the second part, we focus on comparing the standard PSG-position sensor with DL video scoring. For this, the temporal resolution of the body position sensor and the video data need to be identical. The standard PSG-position sensor scores the body position for every 30-second epoch (i.e. two frames per minute (FPM)), while the manual video scoring and camera-based algorithm annotated pose at 2 FPS. Therefore, we downsampled the frames from 2 FPS to 2 FPM.

TABLE IV
Statistical Evaluation of Different Pre-Trained Networks and 4-Layer *de novo* CNN Network Using Accuracy, Precision, Recall, F1score, and Cohen's Kappa (%) as Well as Their Standard Deviations (SD) Over 12 Participants. Gold Standard is the Manually Score Video Data. The Performance of the Standard PSG-Position Sensor Currently Routinely Used in Clinical Practice is Also Provided

Networks	Accuracy	F1score	Precision	Recall	kappa
4-layer CNN	65.5±0.18	64.1±0.18	50.9±0.17	70.3±0.19	59.0±0.25
AlexNet	80.5±0.17	80.8±0.17	77.0±0.16	76.6±0.19	71.2±0.25
VGG-16	89.9±0.11	90.1±0.10	82.5±0.18	82.7±0.17	89.1±0.15
VGG-19	92.2±0.08	92.2±0.08	82.3±0.14	83.9±0.14	87.9±0.12
GoogLeNet	81.1±0.10	80.9±0.11	72.3±0.13	72.3±0.13	68.2±0.19
ResNet-50	91.6±0.11	91.4±0.11	86.5±0.15	86.0±0.14	87.2±0.15
ResNet-101	93.5±0.08	93.3±0.08	87.4±0.13	85.3±0.13	89.9±0.12
ResNet-152	**95.1±0.07**	**94.9±0.08**	**88.2±0.14**	**90.0±0.12**	**92.2±0.10**
PSG-position sensor	88.2±0.10	89.3±0.11	70.0±0.13	70.3±0.13	81.8±0.19

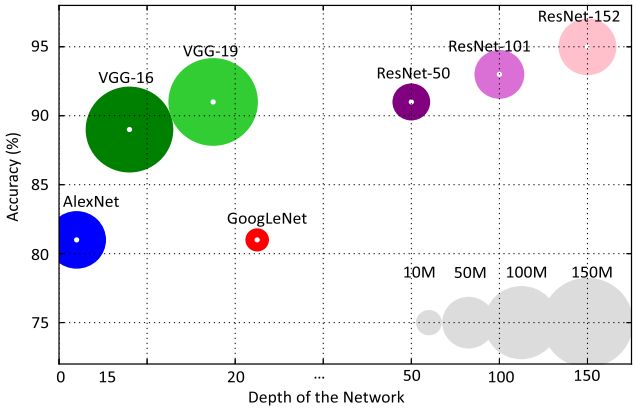

Fig. 6. Accuracy of different pre-trained networks vs. number of layer and number of parameters of each network. The size of the blobs is proportional to the number of network parameters; a legend is reported in the bottom right corner.

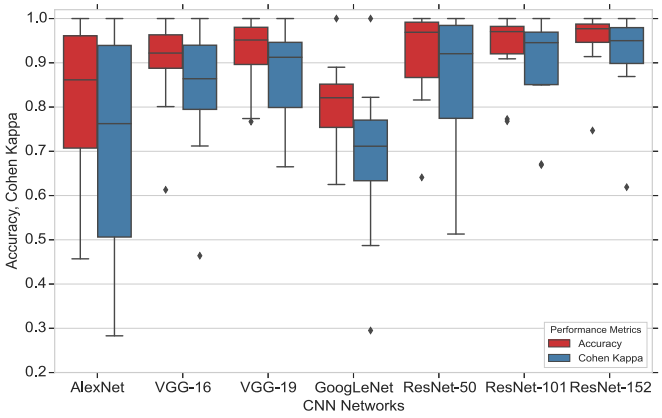

Fig. 7. Boxplot representing the performance of various pre-trained networks: AlexNet, VGG-16, VGG-19, GoogLeNet, ResNet-50, ResNet-101, & ResNet-152. The central line shows the median, the edges of the box represent the 25^{th} and 75^{th} percentiles, the error bars represent 95% confidence intervals, and the additional black markers represent statistical outliers for accuracy (red) and Cohen's kappa (blue). The ResNet-152 outperforms other networks in both comparisons.

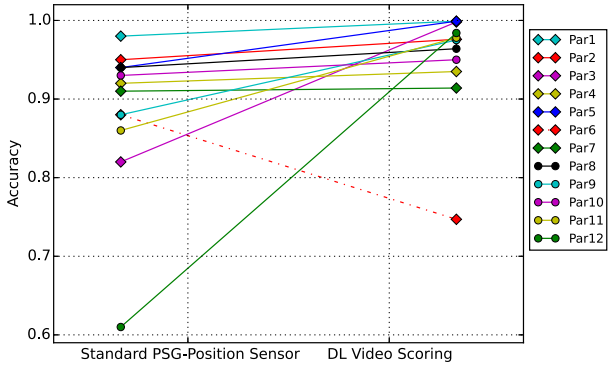

Fig. 8. Comparison of Accuracy between standard PSG-position sensor and DL video scoring (ResNet-152) for classification of sleep poses in 12 participants.

Where there was any variation in pose estimation during the downsampled period, a majority vote was used to describe the pose during this interval. When compared to pose estimation

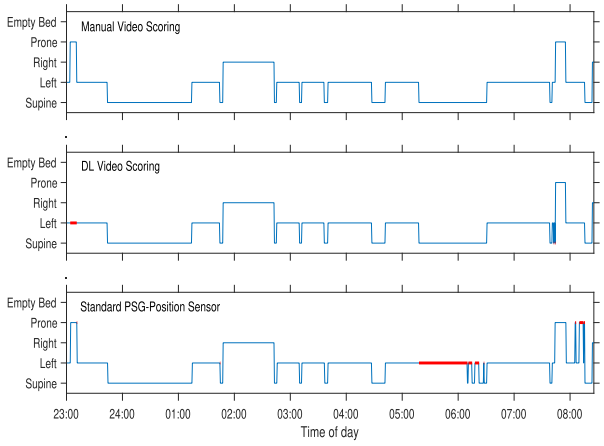

Fig. 9. Time course of poses as detected by Manual video scoring (top), DL video scoring (middle) and standard PSG-position sensor (bottom) during a nocturnal sleep episode in participant # 11. Red lines indicate discrepancies with manual video scoring.

via standard PSG-position sensor, DL video scoring achieved the highest performance on all metrics (accuracy = 95.1%, F1score = 94.9%, kappa = 92.2%) compared with standard PSG-position sensor (accuracy = 88.2%, F1score = 89.3%, kappa = 81.8%).

In Fig. 8 we compare the performance of DL video scoring using ResNet-152 against standard PSG-position sensor across the different participants, using manual scoring as our ground truth. It illustrates higher accuracy for all participants through DL video scoring except for participant #6. By visual inspection, we realised that this participant covered their whole body and face with a blanket during periods of the night which may explain this deviation from the general trend. To illustrate the comparison that is seen across the participant cohort in terms of the temporal pose classification behavior time series, sleep poses for a single typical participant quantified by manual video scoring, DL video scoring, and standard PSG-position sensor are shown in Fig. 9.

Markov chain transition matrices can also be used to quantify the transitions between poses and was used here to provide a visual global description of participants' sleep behavior. The five state Markov chain model for sleep poses is presented in Fig. 10 for three participants with the highest, average, and lowest number of state transitions, respectively. For this section, manual video scoring is again considered as ground truth against which performance of the two other methods has been evaluated. Regarding participant #6, although the standard PSG-position sensor performs better in predicting the states than DL video scoring, it has missed the transition between the left and right state. Similar to participant #6, the standard PSG-position sensor could not detect some of the state transitions for participant #4 as well. DL video scoring performs similarly to the manual video scoring for the state as well as transition detection for participant #3, while PSG recognizes a false activation of prone state which led to additional false transition.

Table V-VII summarises the average sleep pose transition probability matrix derived from Markov analysis over 12 participants for manual video scoring, DL video scoring, and the

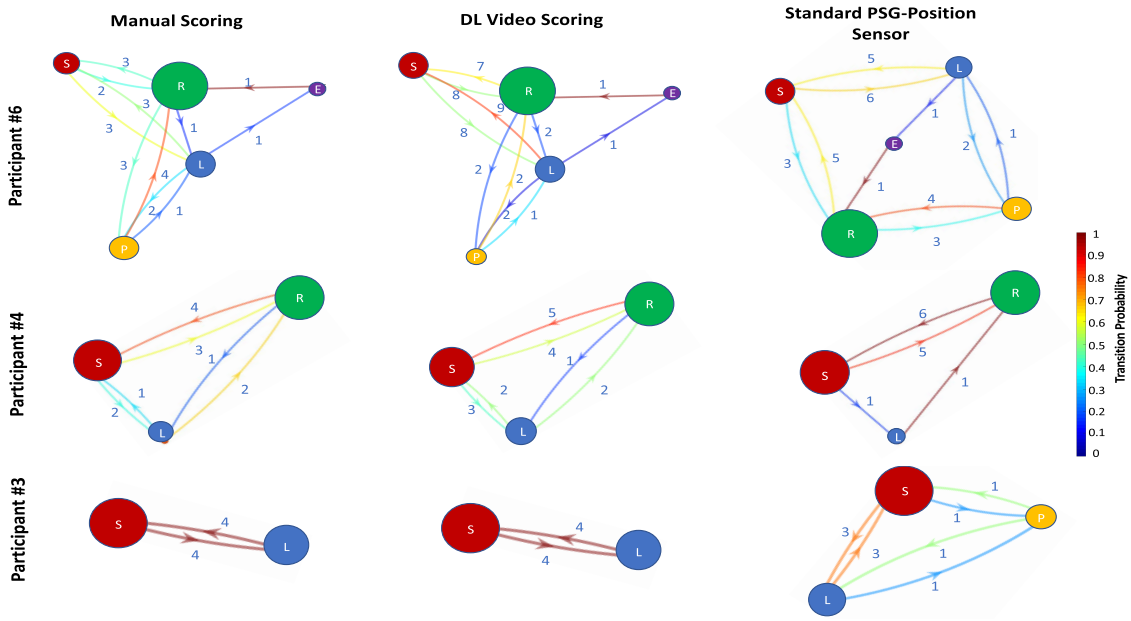

Fig. 10. Sleep pose transition diagram for three participants with the highest (participant 6), average (participant 4), and the lowest number of transitions (participant 3). The left column corresponds to the models created by manual video scoring, the middle column corresponds to DL video scoring, and the right column represents the models from the standard PSG-position sensor. The size of circles correspond to the probability of maintaining a state (i.e., the time spent in the corresponding pose) and each pose is represented with a different colour (supine (S): red, left (L): blue, right (R): green, prone (P): yellow, empty bed (E): purple); circle size is proportional to the state duration. Straight arrows correspond to transitions between states; arrows colour are proportional to the transition probability colourmap, and the numbers on arrows represent the number of transitions between two states.

TABLE V
AVERAGE PROBABILITY MATRIX & SD OF TRANSITIONS BETWEEN STATES AND OF REMAINING IN A STATE FOR MANUAL SCORING

	S	L	R	P	E
S	**0.974±0.01**	0.012±0.01	0.010±0.01	0.000±0.00	0.001±0.00
L	0.008±0.00	**0.984±0.00**	0.004±0.00	0.001±0.00	0.001±0.00
R	0.006±0.29	0.001±0.29	**0.809±0.40**	0.000±0.00	0.000±0.00
P	0.000±0.00	0.006±0.01	0.002±0.00	**0.174±0.38**	0.000±0.00
E	0.031±0.07	0.026±0.00	0.022±0.00	0.000±0.00	**0.374±0.43**

TABLE VI
AVERAGE PROBABILITY MATRIX & SD OF TRANSITIONS BETWEEN STATES AND OF REMAINING IN A STATE FOR DL VIDEO SCORING

	S	L	R	P	E
S	**0.962±0.03**	0.019±0.02	0.016±0.01	0.000±0.00	0.001±0.00
L	0.014±0.01	**0.981±0.01**	0.003±0.00	0.000±0.00	0.001±0.00
R	0.014±0.29	0.002±0.29	**0.884±0.29**	0.007±0.01	0.000±0.00
P	0.012±0.04	0.006±0.02	0.015±0.05	**0.238±0.40**	0.000±0.00
E	0.031±0.07	0.019±0.04	0.033±0.05	0.000±0.00	**0.369±0.42**

standard PSG-position sensor respectively. It revealed high probabilities for staying in one pose (matrix diagonal) and much lower probabilities of transition between different poses. To evaluate the general performance of the transition matrix for DL video scoring and standard PSG-position sensor against

TABLE VII
AVERAGE PROBABILITY MATRIX & SD OF TRANSITIONS BETWEEN STATES AND OF REMAINING IN A STATE FOR STANDARD PSG-POSITION SENSOR

	S	L	R	P	E
S	**0.967±0.02**	0.018±0.01	0.012±0.01	0.000±0.00	0.001±0.00
L	0.011±0.00	**0.977±0.02**	0.008±0.02	0.002±0.00	0.001±0.00
R	0.005±0.29	0.000±0.29	**0.719±0.46**	0.001±0.00	0.000±0.00
P	0.000±0.24	0.022±0.24	0.013±0.23	**0.415±0.48**	0.001±0.00
E	0.054±0.10	0.030±0.07	0.112±0.29	0.000±0.00	**0.437±0.42**

ground truth, we take Root Mean Square Error (RMSE) as evaluation indicators. The mean and SD of RMSE for DL video scoring and standard PSG-position sensor amongst all participants were 0.061±0.093 and 0.115±0.112 respectively demonstrating the effectiveness of this approach compared to standard clinical PSG methods.

V. CONCLUSION

In this work, we developed a non-contact video-based algorithm to automatically monitor body poses and movement during sleep. Nocturnal sleep was quantified in 12 healthy participants by polysomnography. For the classification of five sleep states (four sleep poses + empty bed) transfer learning was introduced to fine-tune the pre-trained deep networks to improve learning efficiency. Our transfer learning strategy relied on keeping and freezing the convolutional layers and updating the fully connected layers to recognize sleep poses.

The performance of seven well-known pre-trained networks including AlexNet, VGG-16, VGG-19, GoogLeNet, ResNet-50, ResNet-101, &ResNet-152 has been explored. ResNet-152 yielded the highest accuracy of 95.1% which was better than all other pre-trained networks as well as a 4-layer *de novo* CNN network. To benchmark the quality of the algorithms performance was compared with the clinical standard PSG-position sensor and validated against manual annotation. We achieved superior performance using the proposed algorithm compared with the standard PSG-position sensor for the detection of sleep poses. Finally, the Markov-based transition matrix was employed to evaluate the performance of the algorithm in describing individual sleep dynamics. It can be concluded that the developed method could be used to monitor sleep positions overnight to assess sleep quality and irregular sleeping patterns. In this study, we have not investigated the impact of thicker blankets, but this is being investigated as part of our on-going research. Future avenues of investigation include investigating an intelligent non-contact monitoring system using the same set-up for the home-environment. It also include exploring the application of this approach across the lifespan.

ACKNOWLEDGMENT

The authors would like to thank Giuseppe Atzori, Ciro Della Monica, and Nayantara Santhi for their help with the data acquisition. They would also like to thank the U.K. Dementia Research Institute (DRI).

REFERENCES

[1] F. S. Luyster, P. J. Strollo, P. C. Zee, and J. K. Walsh, "Sleep: A health imperative," *Sleep*, vol. 35, no. 6, pp. 727–734, Jun. 2012.

[2] R. Winsky-Sommerer, P. de Oliveira, S. Loomis, K. Wafford, D.-J. Dijk, and G. Gilmour, "Disturbances of sleep quality, timing and structure and their relationship with other neuropsychiatric symptoms in Alzheimer's disease and schizophrenia: Insights from studies in patient populations and animal models," *Neurosci. Biobehav. Rev.*, vol. 97, pp. 112–137, Feb. 2019.

[3] S. M. Mohammadi, S. Kouchaki, M. Ghavami, and S. Sanei, "Improving time–frequency domain sleep EEG classification via singular spectrum analysis," *J. Neurosci. Methods*, vol. 273, pp. 96–106, Nov. 2016.

[4] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 758–769, Apr. 2018.

[5] J. Lee, M. Hong, and S. Ryu, "Sleep monitoring system using kinect sensor," *Int. J. Distrib. Sensor Netw.*, vol. 11, no. 10, 2015, Art. no. 875371.

[6] E. Hoque, R. F. Dickerson, and J. A. Stankovic, "Monitoring body positions and movements during sleep using WISPs," in *Proc. Wireless Health*, 2010, pp. 44–53.

[7] J. Wilde-Frenz and H. Schulz, "Rate and distribution of body movements during sleep in humans," *Perceptual Motor Skills*, vol. 56, no. 1, pp. 275–283, Feb. 1983.

[8] C. D. Monica, S. Johnsen, G. Atzori, J. A. Groeger, and D.-J. Dijk, "Rapid eye movement sleep, sleep continuity and slow wave sleep as predictors of cognition, mood, and subjective sleep quality in healthy men and women, aged 20–84 years," *Frontiers Psychiatry*, vol. 9, p. 255, Jun. 2018.

[9] A. Stefani and B. Högl, "Diagnostic criteria, differential diagnosis, and treatment of minor motor activity and less well-known movement disorders of sleep," *Current Treat. Options Neurol.*, vol. 21, no. 1, p. 1, Jan. 2019.

[10] O. Omobomi and S. F. Quan, "Positional therapy in the management of positional obstructive sleep apnea—A review of the current literature," *Sleep Breathing*, vol. 22, no. 2, pp. 297–304, 2018.

[11] M. Gall *et al.*, "A novel approach to assess sleep-related rhythmic movement disorder in children using automatic 3D analysis," *Frontiers Psychiatry*, vol. 10, p. 709, Oct. 2019.

[12] C. Iber *et al.*, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*, vol. 1. Westchester, IL, USA: American Academy of Sleep Medicine, 2007.

[13] G. E. Hintonand and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[14] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[16] J. J. Liu *et al.*, "A dense pressure sensitive bedsheet design for unobtrusive sleep posture monitoring," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, Mar. 2013, pp. 207–215.

[17] A. M. Adami, T. L. Hayes, and M. Pavel, "Unobtrusive monitoring of sleep patterns," in *Proc. 25th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 2, Sep. 2003, pp. 1360–1363.

[18] H. J. Lee, S. H. Hwang, S. M. Lee, Y. G. Lim, and K. S. Park, "Estimation of body postures on bed using unconstrained ECG measurements," *IEEE J. Biomed. Health Inform.*, vol. 17, no. 6, pp. 985–993, Nov. 2013.

[19] F.-C. Yang, C.-H. Kuo, M.-Y. Tsai, and S.-C. Huang, "Image-based sleep motion recognition using artificial neural networks," in *Proc. Int. Conf. Mach. Learn. Cybern.*, vol. 5, Nov. 2003, pp. 2775–2780.

[20] M. S. Rasouli and S. Payandeh, "A novel depth image analysis for sleep posture estimation," *J. Ambient Intell. Hum. Comput.*, vol. 10, no. 5, pp. 1999–2014, 2019.

[21] M.-C. Yu, H. Wu, J.-L. Liou, M.-S. Lee, and Y.-P. Hung, "Multiparameter sleep monitoring using a depth camera," in *Proc. Int. Joint Conf. Biomed. Eng. Syst. Technol.* Berlin, Germany: Springer, 2012, pp. 311–325.

[22] S. Liu and S. Ostadabbas, "A vision-based system for in-bed posture tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1373–1382.

[23] S. Liu, Y. Yin, and S. Ostadabbas, "In-bed pose estimation: Deep learning with shallow dataset," *IEEE J. Transl. Eng. Health Med.*, vol. 7, pp. 1–12, 2019.

[24] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.

[25] F. Deng *et al.*, "Design and implementation of a noncontact sleep monitoring system using infrared cameras and motion sensor," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 7, pp. 1555–1563, Jul. 2018.

[26] T. Grimm, M. Martinez, A. Benz, and R. Stiefelhagen, "Sleep position classification from a depth camera using bed aligned maps," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 319–324.

[27] S. M. Mohammadi, M. Alnowami, S. Khan, D.-J. Dijk, A. Hilton, and K. Wells, "Sleep posture classification using a convolutional neural network," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 1–4.

[28] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[29] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[32] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] P. A. Gagniuc, *Markov Chains: From Theory to Implementation and Experimentation*. Hoboken, NJ, USA: Wiley, 2017.

[35] B. Kemp and H. A. C. Kamphuisen, "Simulation of human hypnograms using a Markov chain model," *Sleep*, vol. 9, no. 3, pp. 405–414, Sep. 1986.