Audio- and video-based estimation of the sleep stages of newborns in Neonatal Intensive Care Unit

Sandie Cabon, Fabienne Porée, Antoine Simon, Bertille Met-Montot, Patrick Pladys, Olivier Rosec, Nicolas Nardi, Guy Carrault

Audio- and Video-based estimation of the sleep stages of newborns in Neonatal Intensive Care Unit

S. Cabon[a,b], F.Porée[a,*] , A. Simon[a,], B. Met-Montot[a], P. Pladys[a], O.Rosec[b], N. Nardi[a] and G. Carrault[a]

Addresses:

[a]Univ Rennes, CHU Rennes, Inserm, LTSI - UMR 1099, F-35000 Rennes, France

[b]Voxygen, F-35000 Rennes, France

Corresponding author:

Fabienne Porée

Laboratoire Traitement du Signal et de l'Image (LTSI), Université de Rennes 1, Campus de Beaulieu, 35042 Cedex, Rennes, France

Tel.: +33-2-23-23-62-24; Fax: +33-2-23-23-69-17

E-mail address: fabienne.poree@univ-rennes1.fr

Abstract

Objective

Premature babies have several immature functions and begin their life under high medical supervision. Since the sleep organization differs across postmenstrual age, its analysis may give a good indication of the degree of brain maturation. However, sleep analysis (polysomnography or behavioral observation) is difficult to install, time consuming and cannot systematically be used. In this context, development of new ways to automatically monitor the neonates, using contactless modalities, is necessary. Therefore, this study presents an innovative non-invasive approach to semi-automatize the classification of infant behavioral sleep states.

Methods

First, three descriptors were extracted from audio and video recordings: vocalizations, motion and eye state of the baby. For this purpose, an original semi-automatic algorithm for the estimation of the eye state was proposed. Secondly, the three descriptors were used in order to obtain an estimation of the behavioral sleep states. Five classifiers (K-Nearest Neighbors, Linear Discriminant Analysis, Support Vector Machine, Random Forest and Multi-Layer Perceptron) were compared to an expert annotation.

Results

Firstly, the comparison of the semi-automatic eye state estimation to manual annotations of 10 videos led to a mean accuracy of 99.4%. Secondly, sleep stage classification was performed. Best results were obtained with Random Forest, for Quiet Alert and Active Alert stages, with 93.5% and 99.0% of accuracy respectively.

Conclusion

The proposed method provides a high capacity to identify alert sleep stages but the differentiation between Quiet Sleep and Active Sleep only by behavioral observations still remains a difficult task to achieve.

Significance

Results presented in this paper are new since no similar approach was proposed in the literature in the context of neonatal intensive care unit. They augur well for the automatic sleep organization assessment to improve newborn care.

Keywords

Premature newborn; sleep; monitoring; video processing; audio processing; neonatal intensive care unit.

1. Introduction

Preterm birth, defined as birth before 37 weeks of gestation, is concerning 15 million babies per year or 11% of all live births worldwide and this number tends to increase every year [1].

Premature babies have several immature functions such as digestive, immunological, cardio-respiratory or neurological functions and begin their life in a Neonatal Intensive Care Unit (NICU), generally in an incubator, under high medical supervision. Their health status is monitored, as well as their maturation, in order to program the incubator exit and then the discharge home. This monitoring relies on several vital signs (cardiac activity, breathing, blood pressure...), and may be extended by a sleep analysis, leading to a sleep stage sequence as a function of time, also called hypnogram [2, 3].

Since the sleep behavior differs across PostMenstrual Age (PMA), its analysis may give a good indication of the degree of brain maturation. In particular, the duration of sleep/wake cycles is supposed to increase with PMA and the sleep organization to evolve with more time spent in quiet sleep [4]. For neonates, two sleep scoring techniques exist: the polysomnography based on the analysis of the ElectroEncephaloGram (EEG) and the direct behavioral observation, which is most commonly used. Based on the rules of Prechtl [5], the sleep scoring is performed in the presence of the baby, by observing body activity levels, eye state (open or closed), respiration regularity, vocalizations... This technique, contactless and without constraint for the baby, is particularly recommended in the NIDCAP (Newborn Individualized Developmental Care and Assessment Program) [6], centered on the comfort of the baby. However, sleep analysis (polysomnography or behavioral observation) is difficult to install, time consuming and cannot systematically be used. In this context, development of new ways to automatically monitor the neonates, using contactless modalities, is necessary.

This work is a part of the Digi-NewB project, funded by the European Union programme for Research and Innovation Horizon2020. Its objective is to reduce mortality, morbidity and health costs of hospitalized newborns by assisting clinicians in their decision-making related to sepsis risk and neurobehavioral maturation. For this purpose, the project aims to develop a new generation of monitoring systems in NICU, using clinical and signal data from different sources (electrophysiological, audio and video).

First studies using video as a support of sleep analysis appeared in 1969, when the Association of the Psychophysiological Study of Sleep (APSS) highlighted the importance to develop a guide for assessing infant sleep. In fact, contemporary criteria (Rechtschaffen and Kales [7]) were not applicable to the infants, who present unique behavioral features of development. Two years later, this manual was proposed by Anders et al. and recommended to supply polysomnographic recordings by behavioral observations [8]. From there, Anders et al. proposed to study infants using only behavioral observations, sometimes supplemented by time-lapse video recordings, an alternative method for long-term recordings. In [9], a study was performed with full-term infants, at two and eight weeks of age. Behavioral states were scored from video considering

eye state, vocalizations and movements. The polygraphic scoring was based on EEG, electrooculogram, electromyogram and respiratory signals. A correlation of 0.79 was obtained between both scorings of the three states (Active REM Sleep, Quiet REM Sleep, Wakefulness). Fuller et al. proposed a similar approach with premature newborns, but only focused on the eye state and the body movements. Furthermore, only sleep stages were considered and vocalizations were not included [10].

The automatic sleep stage classification has been much less addressed in newborns, full-term or preterm, than in adults. However, several modalities were studied including EEG [11–16], cardiorespiratory signals [17] and facial expressions [18]. Though, these methods offer a sleep stage classification more or less specific regarding the PMA. In fact, EEG can only be investigated to distinguish quiet sleep from all other sleep stage under the age of 32 weeks PMA, whereas facial expression assessment can provide a more specific sleep stage classification since 26 weeks PMA. For their part, cardiorespiratory analyses can be reliable for particular sleep stage qualification before 32 weeks PMA [3].

This paper proposes to estimate behavioral sleep states from audio and video acquisitions, which is suitable with a contactless and non invasive monitoring, an approach never envisaged before in the context of NICU [19]. It is based on two main steps (Figure 1):

- Audio and video processing, leading to the characterization of information of three types: i) vocalizations, ii) motion and iii) eye state.
- Combination of the three signals in order to obtain an estimation of the behavioral sleep states.

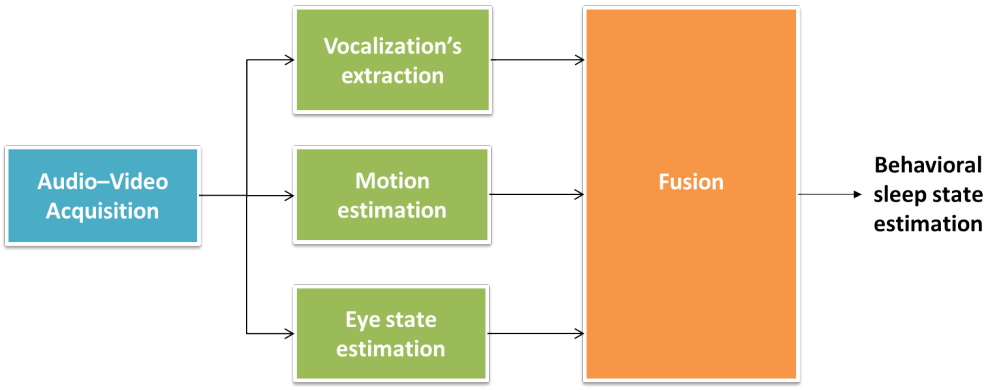

Figure 1: Workflow of the methodological steps for the behavioral sleep state estimation.

Paper is organized as follows. In section 2, methods for the extraction of vocalizations, motion and eye state and the automatic estimation of sleep stages are described. Section 3 is devoted to the results concerning the eye state estimation algorithm and the classification of sleep stages.

2. Methods

In this section, methodology for the extraction of information from audio and video is first presented. A larger part is devoted to the description of the method we developed for the eye state estimation. Then, we propose to estimate sleep stages from these extracted signals by the use of different classifiers.

2.1. Database

Videos were acquired during a project conducted at the University Hospital of Rennes, approved by the Committee on Protection of Individuals (CPP Ouest 6-598) and complying with standards established by the Declaration of Helsinki. Ten newborns were included in this study and the signing of an informed consent, was obtained from parents for each of them.

During the experiments, a camera was set up in the room of the babies to record the scene. It was installed near the bed in order to observe the major part of the body. Recordings were performed in moderate obscurity. The camera had a resolution of 720x756 pixels and recorded 25 frames per second. Sound was acquired by a microphone integrated in the camera with a frequency sampling of 8 kHz. The choice of this low sampling rate has been motivated by our objective to simply detect periods with sound activity while keeping a fast computation time. To consider conditions compatible with a monitoring context, no specific setup was imposed.

Recordings were performed between the 7th and the 11th day of life of premature newborns having a GA ranged from 26 to 32 weeks and, consequently, a PMA comprised between 28 and 33 weeks (see Table 2). Each video duration was between 10 and 32 minutes, leading to a total duration of more than 4 hours (242 minutes and 14 seconds).

For each video recording, a frame was randomly selected and reported on Figure 2 to illustrate the complexity of the database. One can observe that all infants lay on one side and are most of the time covered by a blanket. Four of them are equipped with a ventilatory support and six of them are intubated. Differences can also be noticed in luminosity conditions and camera distances. In audio recordings, only background noises of low energy, for example emitted by the ventilator, were reported.

A scoring of sleep stages, based on a direct behavioral observation [5], was synchronously carried out by a NIDCAP expert during the recording, considering five stages: Quiet Sleep (QS), Active Sleep (AS), Drowsiness (D), Quiet Alert (QA) and Active Alert (AA).

2.2. Vocalizations' extraction

The development of automated sound processing began in the 1960s (see [20] for an historical review). In the context of monitoring, the main issue is to extract newborns vocalizations, also called Voiced/UnVoiced (V/UV) detection. Several strategies were recently proposed to perform this detection automatically. A V/UV detection procedure was proposed in [21], where an interval was selected as voiced if the maximum of

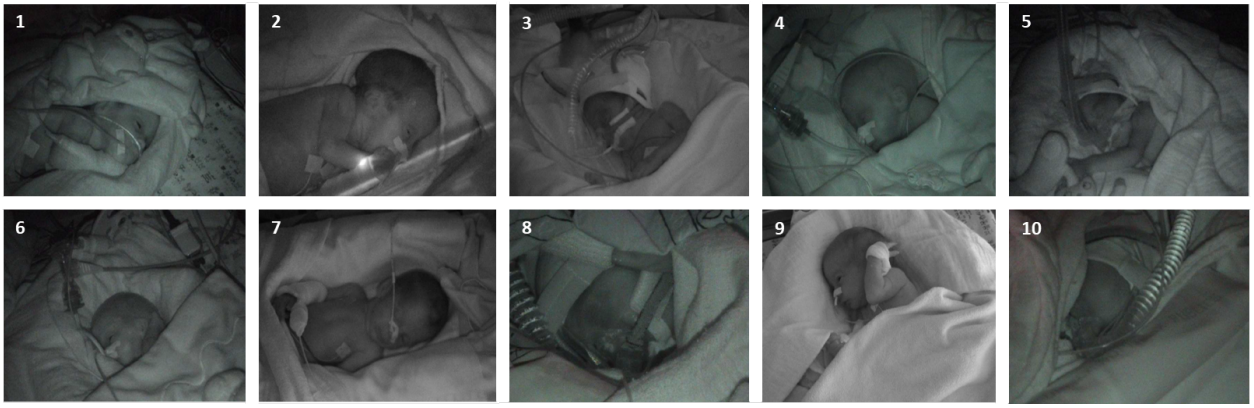

Figure 2: Overview of 10 videos of premature newborns with a PMA comprised between 28 and 33 weeks. All infants lay on one side and are most of the time covered by a blanket. Four of them are equipped with a ventilatory support and six of them are intubated.

the autocorrelation function is greater than a fixed threshold. Several techniques based on the thresholding of the Short-Term Energy (STE) were also investigated [22–27].

Here, baby vocalization detection is performed by applying the methodology proposed in [24]. It is based on the computation of the STE in 20 ms length windows, with 50% overlap between adjacent windows. Then, the highest STE intervals, corresponding to baby vocalizations, are detected using two thresholds, automatically selected using Otsu method [28]. Finally, to construct $V(t)$ signal, values of unvoiced frames are set to 0. An example of a raw sound signal and the resulting vocalization signal $V(t)$ is given in Figure 3.

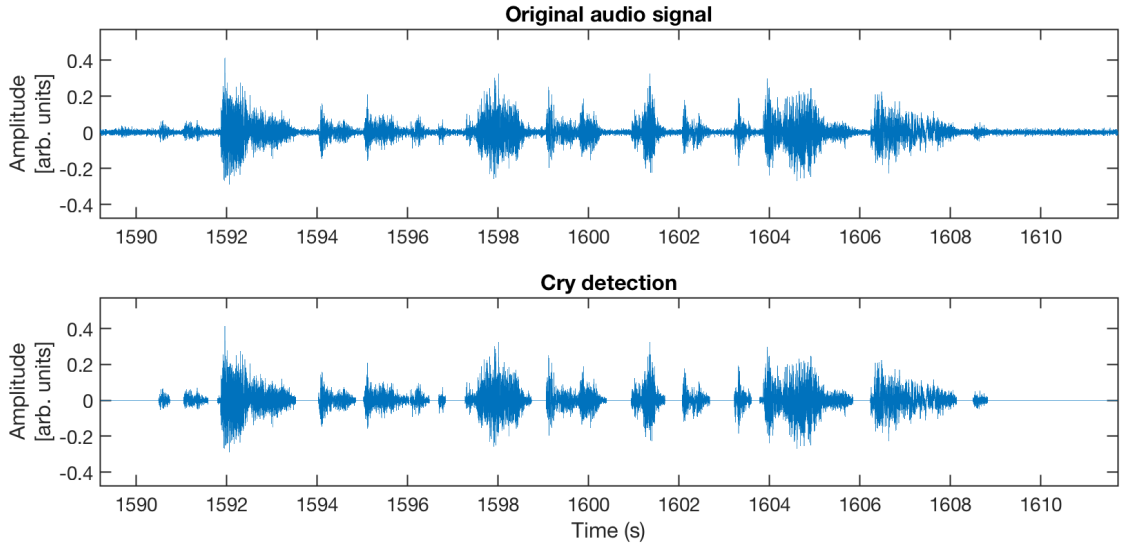

Figure 3: Example of a soundtrack processing (video 9): Top: Raw sound signal - Bottom: Vocalization signal $V(t)$.

2.3. Motion estimation

Many methods have been proposed in the literature to estimate and characterize motion in videos [29]. In paediatrics, specific topics such as general movement assessment or detection of neonatal seizures were addressed [30–32]. In this work, the goal is not to estimate the local motion of the baby, which would be very challenging because of the unconstrained acquisition setup, but to globally characterize its activity. For this purpose, we consider the modifications between two successive frames by computing their difference [33].

In order to limit the influence of noise, the resulting difference image is thresholded with a value T_M (typically low). The amount of activity at a time t, or "motion signal" $M(t)$, is obtained by counting the number of pixels above the threshold. An example is given in Figure 4.

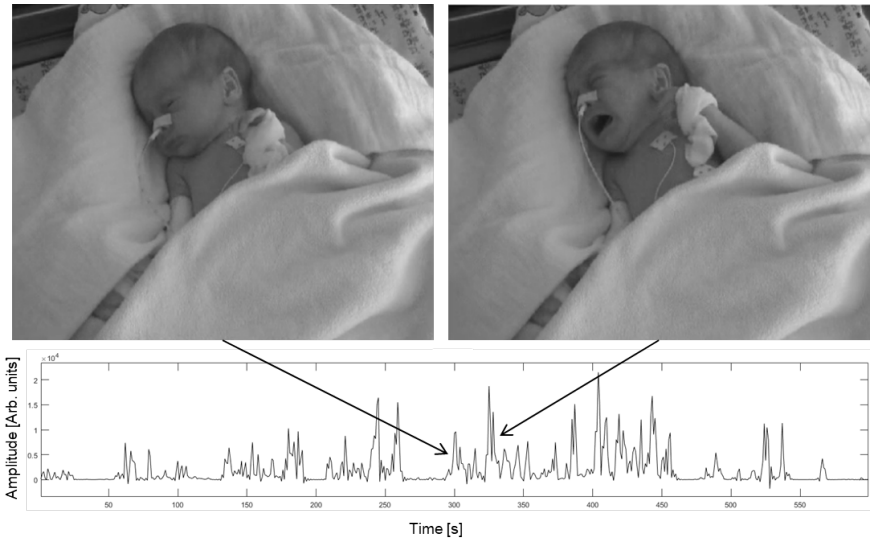

Figure 4: Example of a motion signal $M(t)$ (video 9). Frames acquired before and during movements are presented.

2.4. Eye state estimation

As for adults, where the conditions are usually controlled (full face, front view and open eyes, good luminosity, no occlusion...) [34], eye tracking of infants has been only addressed in a few specific studies. For instance, they were always located in front of the camera and seated either on a parent's lap [35], in an infant chair [36] or in a baby car seat [37, 38].

These conditions being not fulfilled, a specific algorithm was developed for the estimation of the eye state $E(t)$. As the baby may move during the recordings, the algorithm relies on a tracking step of the region of an eye associated with a detection step in this area. It is a semi-automatic procedure, with a limited number of user interactions.

2.4.1. Initialization

The region of the eye has to be tracked since the baby moves during the video acquisition. However, since the state of the eye is changing (open, closed, or in-between), its appearance is often modified. Thus, we decided to perform the tracking of another region of the face, supposed to keep the same appearance and to be at a constant distance from the eye. This region is called the "reference" region of interest (R_{Ref}), and may for example include the nose or an ear. Depending on the acquisition characteristics, the choice is left to the user and performed during the initialization of the processing, i.e. on the first frame ($R_{Ref}(0)$). The user has also to select the region of the eye ($R_{Eye}(0)$). The link between both regions is defined by the relative position between the regions' centers, called δ_{ROI} (Figure 5(a)).

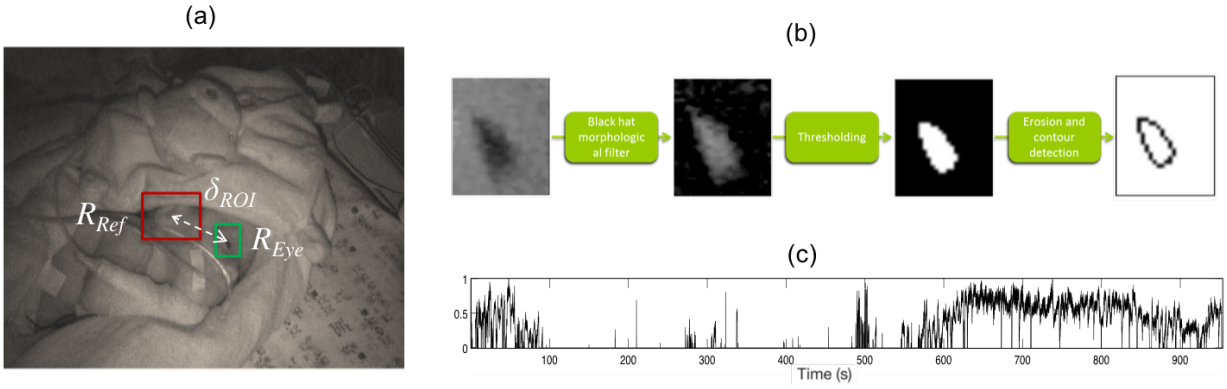

Figure 5: Illustration of the eye state estimation algorithm (video 1): (a) Initialization; (b) Processing steps for the eye segmentation; (c) Example of an eye state signal $E(t)$ with normalized values.

2.4.2. Tracking of the reference region

For each new frame, the reference region is firstly tracked using the template matching approach. It is based on the comparison between the template and each possible position in the new frame. Since the motion amplitude is limited between two successive frames, the search is restricted to a region centered on the previous position of the reference. For each position in the search region, a metric is computed to evaluate the correspondence between the template and the new frame's region centered on this position. The reference region being supposed to keep its appearance, the considered metric was the Sum of Squared Differences (SSD) between the pixels' intensities.

For each new frame at time t, the considered template is, firstly, the initial reference region ($R_{Ref}(0)$). Then, the minimal value of the metric $SSD(t)$ obtained by the template matching is compared to two threshold values D_1 and D_2, as follows:

- $SSD(t) < D_1$: the reference region is considered to have been found in the frame t; $ROI_{Ref}(t)$ is defined as the corresponding position;

- $D_1 \leq SSD(t) \leq D_2$: the resulting position has to be refined; a new template matching is applied, using $R_{Ref}(t-1)$ as the template;
- $SSD(t) > D_2$: the reference region has been lost, the tracking is stopped and the system waits for user's interactions.

At each frame t, after the estimation of $R_{Ref}(t)$, the eye region $R_{Eye}(t)$ is retrieved thanks to the relative position δ_{ROI}.

2.4.3. Eye detection

Once the eye region $R_{Eye}(t)$ has been found in the frame t, a segmentation process is used to extract the eye contour. It includes the following steps (Figure 5(b)):

- A "black hat" morphological transformation using a structuring element of size 15x15 to enhance the contrast between the darkest regions and their neighborhood;
- A thresholding of the resulting image by a value T_E;
- A morphological erosion using a structuring element of size 5x5 to remove small regions;
- An edge detection of the extreme outer contours, using Green's theorem.

Then, the eye state $E(t)$ is defined by its surface, depending on the number k of detected contours, as follows:

- $k = 0$: the eye is considered closed and $E(t) = 0$;
- $k = 1$ or $k = 2$: the eye is considered as open, $E(t)$ is the sum of the surfaces of the k detected edges and δ_{ROI} is updated with the center of the eye area. The case $k = 2$ occurs when the contour is divided in two areas separated by the pupil;
- $k > 2$, this result corresponds to noisy detections. The eye is taken for not detected, the tracking is stopped and the system waits for a user interaction.

2.4.4. User interactions

In the case of uncertainty concerning the tracking ($SSD(t) > D_2$) or concerning the detection ($k > 2$), the algorithm stops and the user has to select again both regions of interest (as in the first frame), possibly after forwarding the video if one occlusion occurs.

2.4.5. Smoothing of the eye state signal

Once the video recording has been processed, a sliding median filter is applied on the eye opening values to limit the brief incoherent changes. Since an eye blanking has been observed to last less than five successive frames, the median filter window length has been set up at 5.

An example of a eye state signal is given in Figure 5(c).

2.5. Sleep stage estimation

In this section, we propose a strategy to characterize newborn sleep organization based on the fusion of the extracted descriptors.

In this objective, data are first standardized by applying a set of post-processing to the three signals $V(t)$, $M(t)$ and $E(t)$:

- A Hilbert transform is applied to the vocalizations signal $V(t)$ to recover the signal envelope and proceed next with a positive signal;
- It is also downsampled to 25 Hz, the sampling frequency of motion and eye state signals;
- The three signals are smoothed using a median filter on 1-second length windows;
- The three signals are normalized to the range $[0, 1]$, relatively to the global maximum of the database (separately for each type of signal).

The resulting signals are called $\bar{V}(t)$, $\bar{M}(t)$ and $\bar{E}(t)$.

Then, a model to estimate sleep stages on the whole population, based on machine learning, can be built. For this purpose, each t is considered as a sample defined by three values $\bar{V}(t)$, $\bar{M}(t)$ and $\bar{E}(t)$, associated with its sleep stage label (QS, AS, D, QA or AA). We selected five commonly known approaches that cover a large scope of classification hypotheses: K-Nearest Neighbors (KNN) [39], Linear Discriminant Analysis (LDA) [40], Support Vector Machine (SVM) [41], Random Forest (RF) [42] and Multi-Layer Perceptron (MLP) [43]. Since those methods need balanced dataset, a random under-sampling method was first applied to equalize the number of elements of each sleep stage class. Then, the dataset is randomly split into a training and a testing part containing respectively 60% and 40% of the balanced dataset. These operations are repeated 30 times.

3. Results

This section is dedicated to the validation of our approach. First, software and platforms that were used to produce these results are reported. As sound segmentation and motion estimation have been already

evaluated by their authors, their conformity was only confirmed by visual assessment. Nevertheless, an original strategy was defined for the evaluation of our eye state estimation method.

Then, performances concerning sleep stage estimation are given by comparing the results of the five different machine learning approaches to an expert annotation.

3.1. Software and Platforms

Several software and platforms have been involved in this project. Video processing (motion and eye state estimation) was developed in C++ with OpenCV 3.0 library whereas vocalization extraction and statistical analyses were performed with Matlab R2018a. Machine learning approaches were implemented in Python 3.6 using scikit-learn 0.20.0 [44].

3.2. Tuning of the parameters

Eye state detection. Some parameters in motion as well as in eye state detection algorithms had to be tuned to fit the database properties.

Regarding motion analysis, the threshold T_M has been defined by studying the cumulative histogram of video sequences with empty rooms (without baby or adult), resulting to a low value of 10, initial intensity ranging in [0, 255].

Three thresholds were used in the eye state estimation algorithm. Thresholds D_1 and D_2 (intensity ratios), that defined the accepted appearance modifications of the reference region, were empirically set to 0.02 and 0.04, respectively. The threshold T_E was manually selected between 15 and 25, depending on the video luminosity. In fact, the contrast is less pronounced in low luminosity videos and consequently, a lower threshold is necessary.

Classifier parameters. In section 3.4.1, five classifiers are compared. For each of them, the set of parameters resulting to the highest performances was first identified. For that purpose, several parameters and hyper-parameters have been tuned. A summary of the tests is reported in Table 1. From there, the best set of parameters (in bold in Table 1) was retained for each method. It is important to note that some parameters had little or no influence on the performances. When several values were suitable, we chose to keep the one with the lowest computational time.

3.3. Performances of the eye state estimation method

The eye state estimation algorithm was evaluated by comparing its results with a manual analysis of the videos. On the one hand, video durations and sampling rate implied that a visual scoring frame by frame was not possible. A scoring with another resolution (for example one value per second) was also rejected because it could avoid some short events (as eye blinking). For these reasons, we chose to perform a scoring of 5% of randomly selected frames. On the other hand, intermediate states being difficult to objectively

Table 1: Parameters testing summary. Final selecting sets of parameters are marked in bold.

Method	Parameters
KNN	Number of neighbors ∈ [**1**, 3, 5, 10, 20]
	Minkowski distance: Manhattan or **Euclidean**
LDA	Solver ∈ [singular value decomposition, **least squares solution**, eigenvalue decomposition]
SVM	Kernel ∈ [linear, **Gaussian**, polynomial]
	Hyper-parameters depending on the kernel:
	→ linear: no additional parameter
	→ Gaussian: margin ∈ [0.01, 0.1, 1, 10, 100, 10^3, 10^4, **10^5**, 10^6, 10^7]
	gamma ∈ [0.01, 0.1, 1, 10, 100, 10^3, **10^4**, 10^5, 10^6, 10^7]
	→ polynomial: degree ∈ [1, 2, **3**, 4]
RF	Number of trees ∈ [5, 10, **50**, 100, 200]
	Quality split criterion: gini or **entropy**
MLP	Hidden layer size ∈ [1, **2**, 5, 10, 20]
	Activation function ∈ [identity, logistic sigmoid, **hyperbolic tan**, rectified linear unit]

determine, the user decided if the eyes were 'Open' (=1) or 'Closed' (=0). In this context, the values of the surfaces provided by the algorithm (Figure 5(c)) were binarized i.e. all the non-zero values were set to 1 (i.e. "Open").

Considering the manual scoring as the reference, the Sensitivity (Se), Specificity (Sp) and Accuracy (Acc) of the proposed method were computed for the 10 videos and are reported in Table 2. Results show that accuracies range from 96.56% to 100% (99.4% on average). More precisely, sensitivity and specificity are always greater than 95% and 97% respectively, except in the case of the video 7 where Se is equal to 78.57%. In this video, the baby had very rapid motions that led to tracking errors.

The total number of user interactions (Table 2) has also been quantified and is supplemented by the number of them not due to hidden eye. We can also notice that most of time only few interactions were required (often none and up to two) except for three videos (5, 7 and 9). For video 5, the total number of interactions is consistent since it appeared mostly in case of hidden eye. However, processing of videos 7 and 9 requested irrelevant manual interactions due to the change of appearance of the ROI_{Ref}, for example after a rotary motion of the head. Nevertheless, performances on video 7 and 9 are high with a global concordance of 96.56% and 99.25%, respectively.

Computational time of our approach is attractive since the algorithm takes 0.047 second to process one frame, in its current version. As an example, the video 8 (duration: 41'02), that required no interaction

Table 2: Newborn data (number, GA and PMA in weeks+days). Video data (duration in min'sec, number of frames visually scored and their repartition 'Open'/'Closed'). Algorithm's performance (sensitivity, specificity, accuracy in %). Number of user interactions (total number and regardless hidden eye).

Newborns			Videos				Performances			Nb of Interactions	
N°	GA (w+d)	PMA (d)	Duration (min'sec)	Nb of frames visually scored	Nb of frames 'Open'	Nb of frames 'Closed'	Se (%)	Sp (%)	Acc (%)	Total number	Regardless hidden eye
1	28+4	29+6	17'27	1384	40	1344	98.21	99.68	99.64	2	2
2	28+4	29+4	31'50	2367	343	2024	99.71	99.75	99.75	1	1
3	28+4	29+4	30'58	2357	132	2225	97.73	99.87	99.75	2	2
4	28+6	30+0	27'10	2105	56	2049	98.21	100.00	99.11	0	0
5	27+0	28+4	10'11	831	0	831	-	100.00	100.00	25	2
6	28+6	29+6	24'04	1634	146	1478	99.32	100.00	99.66	0	0
7	30+6	32+0	15'07	1198	28	1170	78.57	99.74	99.25	43	37
8	26+0	27+1	41'02	3064	4	3060	100.00	100.00	100.00	0	0
9	31+3	32+4	16'03	1191	584	607	95.38	97.69	96.56	54	25
10	29+5	30+6	28'22	2156	4	2152	100.00	100.00	100.00	0	0

(except the initialization step), was processed in 48 minutes and 12 seconds. In case of interaction, time to resume the algorithm is equivalent to the initialization step duration, meaning a few seconds. In addition, several videos can be processed simultaneously optimizing considerably the time required to assess the sleep of different babies.

3.4. Results of sleep stage classification from extracted descriptors

3.4.1. Descriptor analysis

As a first step, the distribution of the values (mean ± std) of signals $\bar{V}(t)$, $\bar{M}(t)$ and $\bar{E}(t)$, obtained on the whole database, as functions of the sleep stages provided by the expert, are reported in Figure 6.

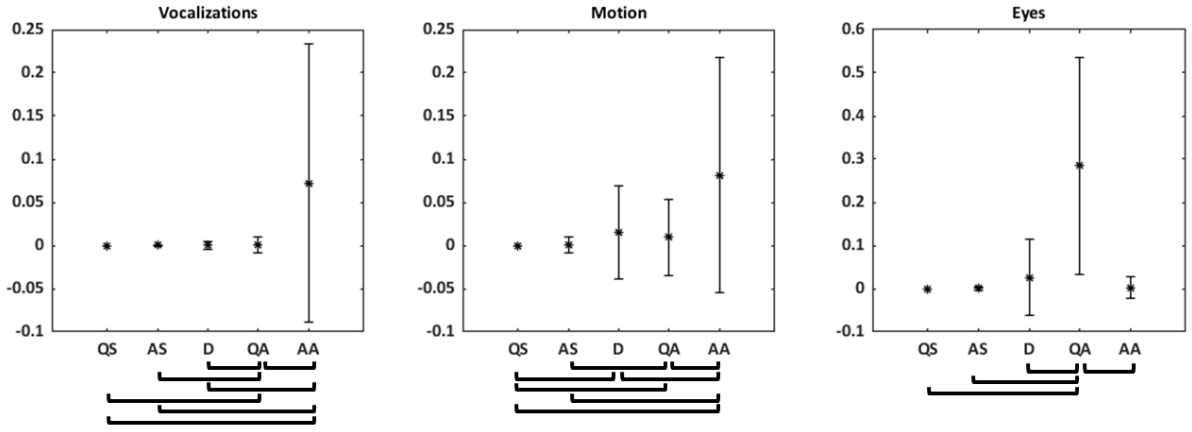

Figure 6: Values (mean ± std) of signals $\bar{V}(t)$, $\bar{M}(t)$ and $\bar{E}(t)$ as functions of the sleep stages: Quiet sleep (QS), Active Sleep (AS), Drowsiness (D), Quiet Alert (QA) and Active Alert (AA). Pairwise comparison among sleep stages that revealed statistically significant differences ($p<0.05$) by Mann-Whitney U test are identified by brackets.

We can observe that there are no vocalization in QS and AS, very low amplitudes in D and QA, and

higher values and dispersion in AA. Motion values are null in QS, very low in AS, moderate in D and QA, and very high in AA.

Results are different for eye state. Eyes are coherently closed in sleep stages QS and AS. In D, values have a low mean amplitude with a moderate standard deviation, corresponding to short openings. Highest values and dispersion of the values are observed in QA since infant eyes are most of the time opened, but can also be closed. In AA, values are low, because a newborn closes most of the time the eyes when he is nervous or while he is crying (see right picture in Figure 4).

Statistical analyses were also conducted by Mann-Whitney U test to discuss pairwise differences between sleep stages for each descriptor. Bootstrap method [45] was applied to minimize the repetitive effect of our dataset. Hence, 100 draws of 24 random samples (representing only 0.2% of the less represented class) have been achieved. The resulting median p-values were studied. Only statistically significant differences ($p<0.05$) are identified by brackets in Figure 6. These results confirm that our set of descriptors is valuable to characterize sleep since most of sleep states can be differentiated from others by at least one descriptor. We can also note that vocalization and motion features are discriminating in more cases (7 over 10) than eyes (4 only). However, no descriptor showed statistical differences for QS vs AS.

Figure 6 shows the complementarity of the three informations, since the value repartition according to the sleep stages is different from a modality to another. Moreover, they are in accordance with the stage definitions for the newborns [5] and give a qualitative validation of the approach. However, they, as of now, augur potential difficulties to differentiate Active Sleep and Quiet Sleep.

3.4.2. Classification results

Performances of the sleep stage classification are evaluated taking as reference the manual scoring performed by the expert. The results of the 30 repeated operations were averaged, which led to a mean accuracy and standard deviation for each sleep stage.

Results presented in Figure 7(a) show first that the five classification methods have greater accuracy values for the alert stages (QA and AA). Results with KNN and LDA are fluctuating with high standard deviations observed for some stages (e.g., D, QS and AS for KNN or QS and AS for LDA). To a lesser extent, the same observation can be made for MLP in QS and AS. Although they are closed to SVM results, best performances are obtained with Random Forest for QA and AA, with 93.5% and 99.0% of accuracy respectively, while the results for calm stages (QS, AS and D) are weaker (under 84.1%).

To complement these results, Cohen's Kappa [46] and Kendall's tau [47] coefficients have been computed and reported in Figure 7(b).

Cohen's Kappa coefficient, that measures a ratio-scaled degree of disagreement between two approaches, shows, with greater values than 0.44 for all methods except LDA, a moderate concordance between sleep stages provided by the expert and the ones automatically estimated [48]. Kendall tau coefficient aims to

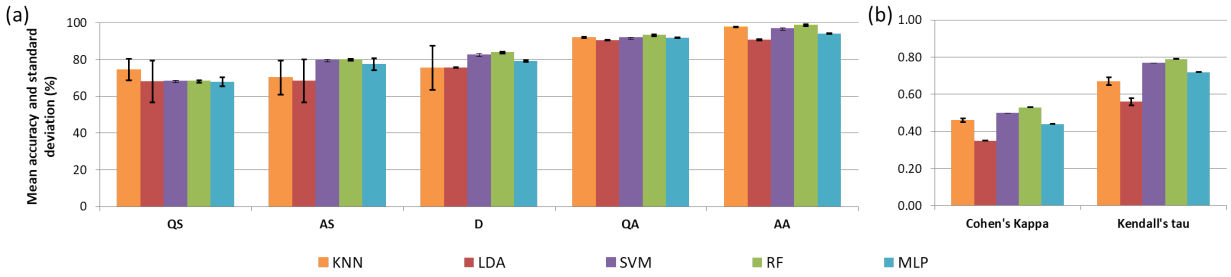

Figure 7: Five machine learning methods are compared: KNN, LDA, SVM, RF and MLP. (a) Classification mean accuracy (%) and standard deviation for each sleep stage: Quiet sleep (QS), Active Sleep (AS), Drowsiness (D), Quiet Alert (QA) and Active Alert (AA); (b) Cohen's Kappa and Kendall's tau coefficients (mean ± std).

measure the association between two quantities, assessing the similarity between ordered data. For this purpose, each stage was affected with a value from 1 (QS) to 5 (AA). For each machine learning approach, excluding LDA, high degrees of concordance are observed (above 0.67), meaning that the estimation errors are mainly made from one stage to another close one (e.g., AS was estimated instead of QS). As proof, a test considering only two classes, QS+AS+D versus QA+AA (in other terms calm vs alert stages), led to a higher accuracy of 94.8% with Random Forest classifier.

In conclusion, all these results suggest that the estimation of alert stages (QA and AA) is correctly performed, but that the differentiation between the three calm stages (QS, AS and D) remains more difficult. These results are not surprising considering that Drowsiness is an intermediate state by definition, and that QS and AS are close in terms of behavior, as shown by Figure 6.

4. Conclusion

In this paper, a whole process was defined to semi-automatically and contactless monitor premature newborns using audio-video acquisitions. It includes different processing to extract baby's vocalizations, motion and eye state. For the last one, a specially developed algorithm based on a two-step approach was evaluated with manual annotations of 10 videos and led to a mean accuracy of 99.4%. Weaknesses have been pointed out in the presence of rapid motion and/or complex transformation of the reference region.

Then, the three descriptors were used in order to obtain an estimation of the behavioral sleep states. Five classifiers (KNN, LDA, SVM, RF and MLP) were compared to a NIDCAP expert annotation. Best results were obtained with Random Forest for the two alert stages QA and AA.

Results presented in this paper are new since no similar approach was proposed in the literature in the context of NICU. In fact, no automatic classification of Prechtl sleep stages was ever conducted in preterm infants. All studies dealing with early days of preterm newborns were based on EEG analyses and only quiet sleep detection was performed [13–15]. Only two studies dealing with preterm and full-term newborns

proposed to identify four sleep states with EEG, but was focused on newborns at 38 to 42 weeks PMA [11, 16]. Alternatively, an automatic classification of three states (sleep, awake and crying) was proposed from face analysis [18]. However, authors reported that its application in a realistic hospital environment was not directly possible. In addition, to date, no automated video analysis of sleep has been conducted on preterm infants [3]. The same observation can be made concerning audio analyses. Furthermore, regardless of the clinical target, the combination of audio and video descriptors is also innovative. Despite a wide variety of publications about audio or video processing in paediatrics, only one study integrating both automatically, were, to our knowledge, published [25]. Nevertheless, they were investigated separately.

If this preliminary study shows encouraging results, they will have to be confirmed on a larger database. Although, it is worthwhile to remind that the constitution of such a database is difficult, notably because the annotation of sleep stages by an expert is time consuming.

In the present study, algorithms are applied off-line, on video recordings. The manual interactions for eye state estimation are only required when needed (e.g., occlusions). In that case, the processing is paused. Thus, there is no need for the user nor to perform the analysis in newborn rooms nor to continuously supervise the processing. Consequently, in a heavy workload context for nurses, the sleep analysis can be deferred and thus more newborns may benefit from this follow-up. However, refinements can be envisaged to enhance performances and move towards a fully automated solution. For example, the eye state detection algorithm robustness may be improved by tracking several regions of interest. In addition, an automatic selection of region(s) of interest by the use of a deep learning approach could be considered on a larger database.

Moreover, it is important to note that the level of discomfort induced to the baby by such a strategy is lower or equivalent to actual techniques but its impact, in both forms (semi-automatic/deferred or automatic/continuous), on daily care routine will have to be studied.

Additionally, a better differentiation between Quiet Sleep and Active Sleep may be achieved by adding the cardio-respiratory information, since it has been shown to be discriminative in Quiet Sleep [17]. However, rather than using additional sensors, it would be doubtless preferable to pursue a non-invasive strategy. Indeed, heart rate and respiration were recently estimated by automatic video processing in NICU in real conditions [49, 50].

Nonetheless, these results augur well for the automatic sleep organization assessment to improve newborn care, but also infant well-being and development. Indeed, this work shows the relevance of our approach to estimate sleep stages by the means of non-invasive techniques such as audio and video processing. These results are directly linked to Digi-NewB objectives and suggest the possibility to monitor sleep in premature newborns and, thus, to quantify their neuro-behavioral development *ex-utero*.

Acknowledgment

The authors would like to thank the clinicians from the Pôle Médico-Chirurgical de Pédiatrie et de Génétique Clinique of the CHU of Rennes for their large involvement in this study. Results incorporated in this publication received funding from the European Union's Horizon 2020 research and innovation program under grant agreement N^o 689260 (Digi-NewB project).

References

[1] H. Blencowe, S. Cousens, M. Z. Oestergaard, D. Chou, A.-B. Moller, R. Narwal, A. Adler, C. V. Garcia, S. Rohde, L. Say, J. E. Lawn, National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications, The Lancet 379 (2012) 2162–72.

[2] J. Huvanandana, C. Thamrin, M. Tracy, M. Hinder, C. Nguyen, A. McEwan, Advanced analyses of physiological signals in the neonatal intensive care unit, Physiological Measurement 38 (2017) R253.

[3] J. Werth, L. Atallah, P. Andriessen, X. Long, E. Zwartkruis-Pelgrim, R. M. Aarts, Unobtrusive sleep state measurements in preterm infants–a review, Sleep Medicine Reviews 32 (2017) 109–122.

[4] L. Curzi-Dascalova, M. Mirmiran, Manual of methods for recording and analyzing sleep-wakefulness states in preterm and full-term infant, INSERM, Paris, 1996.

[5] H. F. Prechtl, The behavioural states of the newborn infant (a review), Brain research 76 (1974) 185–212.

[6] H. Als, Program guide: Newborn individualized developmental care and assessment program (NIDCAP): An education and training program for health care professionals, Boston, MA: Children's Medical Center Corporation, 2002.

[7] A. Rechtschaffen, A. Kales, A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects, Los Angeles: UCLA Brain Information Service/Brain Research Institute, 1968.

[8] T. F. Anders, R. N. Emde, A. H. Parmelee, A manual of standardized terminology, techniques and criteria for scoring of states of sleep and wakefulness in newborn infants, Los Angeles: UCLA Brain Information Service, NINDS Neurological Information Network, 1971.

[9] T. F. Anders, A. M. Sostek, The use of time lapse video recording of sleep-wake behavior in human infants, Psychophysiology 13 (1976) 155–8.

[10] P. W. Fuller, W. H. Wenner, S. Blackburn, Comparison between time-lapse video recordings of behavior and polygraphic state determinations in premature infants, Psychophysiology 15 (1978) 594–8.

[11] A. Piryatinska, G. Terdik, W. A. Woyczynski, K. A. Loparo, M. S. Scher, A. Zlotnik, Automated detection of neonate EEG sleep stages, Computer Methods and Programs in Biomedicine 95 (2009) 31–46.

[12] L. Fraiwan, K. Lweesy, N. Khasawneh, M. Fraiwan, H. Wenz, H. Dickhaus, Time frequency analysis for automated sleep stage identification in fullterm and preterm neonates, Journal of Medical Systems 35 (2011) 693–702.

[13] A. Dereymaeker, K. Pillay, J. Vervisch, S. Van Huffel, G. Naulaers, K. Jansen, M. De Vos, An automated quiet sleep detection approach in preterm infants as a gateway to assess brain maturation, International Journal of Neural Systems 27 (2017) 1750023.

[14] O. De Wel, M. Lavanga, A. C. Dorado, K. Jansen, A. Dereymaeker, G. Naulaers, S. Van Huffel, Complexity analysis of neonatal EEG using multiscale entropy: applications in brain maturation and sleep stage classification, Entropy 19 (2017) 516.

[15] A. H. Ansari, O. De Wel, M. Lavanga, A. Caicedo, A. Dereymaeker, K. Jansen, J. Vervisch, M. De Vos, G. Naulaers, S. Van Huffel, Quiet sleep detection in preterm infants using deep convolutional neural networks, Journal of Neural Engineering 15 (2018) 066006.

[16] K. Pillay, A. Dereymaeker, K. Jansen, G. Naulaers, S. Van Huffel, M. De Vos, Automated eeg sleep staging in the term-age baby using a generative modelling approach, Journal of neural engineering 15 (2018) 036004.

[17] R. M. Harper, V. L. Schechtman, K. A. Kluge, Machine classification of infant sleep state using cardiorespiratory measures, Electroencephalography and Clinical Neurophysiology 67 (1987) 379–387.

[18] L. Hazelhoff, J. Han, S. Bambang-Oetomo, P. H. N. de With, Behavioral state detection of newborns based on facial expression analysis, in: International Conference on Advanced Concepts for Intelligent Vision Systems, Springer, 2009, pp. 698–709.

[19] S. Cabon, F. Porée, A. Simon, O. Rosec, P. Pladys, G. Carrault, Video and audio processing in paediatrics: a review, Physiological Measurement (2019). https://doi.org/10.1088/1361-6579/ab0096.

[20] O. Wasz-Höckert, K. Michelsson, J. Lind, Twenty-five years of Scandinavian cry research, in: Infant Crying, Springer, 1985, pp. 83–104.

[21] S. Orlandi, C. Manfredi, L. Bocchi, M. Scattoni, Automatic newborn cry analysis: A non-invasive tool to help autism early diagnosis, in: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, IEEE, 2012, pp. 2953–2956.

[22] M. A. R. Díaz, C. A. R. García, L. C. A. Robles, J. E. X. Altamirano, A. V. Mendoza, Automatic infant cry analysis for the identification of qualitative features to help opportune diagnosis, Biomedical Signal Processing and Control 7 (2012) 43–49.

[23] S. Orlandi, L. Bocchi, G. Donzelli, C. Manfredi, Central blood oxygen saturation vs crying in preterm newborns, Biomedical Signal Processing and Control 7 (2012) 88–92.

[24] S. Orlandi, P. H. Dejonckere, J. Schoentgen, J. Lebacq, N. Rruqja, C. Manfredi, Effective pre-processing of long term noisy audio recordings: An aid to clinical monitoring, Biomedical Signal Processing and Control 8 (2013) 799–810.

[25] S. Orlandi, A. Guzzetta, A. Bandini, V. Belmonti, S. D. Barbagallo, G. Tealdi, S. Mazzotti, M. L. Scattoni, C. Manfredi, AVIM - A contactless system for infant data acquisition and analysis: Software architecture and first results, Biomedical Signal Processing and Control 20 (2015) 85–99.

[26] S. Orlandi, C. A. R. Garcia, A. Bandini, G. Donzelli, C. Manfredi, Application of pattern recognition techniques to the classification of full-term and preterm infant cry, Journal of Voice 30 (2016) 656–663.

[27] C. Manfredi, A. Bandini, D. Melino, R. Viellevoye, M. Kalenga, S. Orlandi, Automated detection and classification of basic shapes of newborn cry melody, Biomedical Signal Processing and Control 45 (2018) 174–181.

[28] N. Otsu, A threshold selection method from gray-level histograms, IEEE Transactions on Systems, Man, and Cybernetics 9 (1979) 62–66.

[29] T. B. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding 104 (2006) 90–126.

[30] A. Stahl, C. Schellewald, O. Stavdahl, O. M. Aamo, L. Adde, H. Kirkerod, An optical flow-based method to predict infantile cerebral palsy, IEEE Transactions on Neural Systems and Rehabilitation Engineering 20 (2012) 605–14.

[31] C. Marcroft, A. Khan, N. D. Embleton, M. Trenell, T. Plotz, Movement recognition technology as a method of assessing spontaneous general movements in high risk infants, Frontiers in Neurology 5 (2014) 284.

[32] M. Pediaditis, M. Tsiknakis, N. Leitgeb, Vision-based motion detection, analysis and recognition of epileptic seizures–a systematic review, Computer Methods and Programs in Biomedicine 108 (2012) 1133–48.

[33] S. Okada, Y. Ohno, K. Kato-Nishimura, I. Mohri, M. Taniike, et al., Examination of non-restrictive and non-invasive sleep evaluation technique for children using difference images, in: 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2008, pp. 3483–3487.

[34] A. Al-Rahayfeh, M. Faezipour, Eye tracking and head movement detection: A state-of-art survey, IEEE Journal of Translational Engineering in Health and Medicine 1 (2013).

[35] S. P. Johnson, J. A. Slemmer, D. Amso, Where infants look determines how they see: Eye movements and object perception performance in 3-month-olds, Infancy 6 (2004) 185–201.

[36] S. Hunnius, R. H. Geuze, Developmental changes in visual scanning of dynamic faces and abstract stimuli in infants: A longitudinal study, Infancy 6 (2004) 231–255.

[37] G. Gredeback, C. von Hofsten, Infants' evolving representations of object motion during occlusion: A longitudinal study of 6- to 12-month-old infants, Infancy 6 (2004) 165–184.

[38] A. Franklin, M. Pilling, I. Davies, The nature of infant color categorization: Evidence from eye movements on a target detection task, Journal of Experimental Child Psychology 91 (2005) 227–248.

[39] L. E. Peterson, K-nearest neighbor, Scholarpedia 4 (2009) 1883.

[40] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, K.-R. Mullers, Fisher discriminant analysis with kernels, in: Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop., Ieee, 1999, pp. 41–48.

[41] V. Vapnik, S. Mukherjee, Support vector method for multivariate density estimation, in: Advances in neural information processing systems, 2000, pp. 659–665.

[42] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.

[43] S. K. Pal, S. Mitra, Multilayer perceptron, fuzzy sets, classifiaction (1992).

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[45] A. M. Zoubir, B. Boashash, The bootstrap and its application in signal processing, IEEE signal processing magazine 15 (1998) 56–76.

[46] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1960) 37–46.

[47] M. G. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–93.

[48] M. Scatena, S. Dittoni, R. Maviglia, et al., An integrated video-analysis software system designed for movement detection and sleep analysis. Validation of a tool for the behavioural study of sleep, Clinical Neurophysiology 123 (2012) 318–23.

[49] L. Cattani, D. Alinovi, G. Ferrari, R. Raheli, E. Pavlidis, C. Spagnoli, F. Pisani, Monitoring infants by automatic video processing: A unified approach to motion analysis, Computers in Biology and Medicine 80 (2017) 158–165.

[50] M. van Gastel, B. Balmaekers, S. B. Oetomo, W. Verkruysse, Near-continuous non-contact cardiac pulse monitoring in a neonatal intensive care unit in near darkness, in: Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics, volume 1050114, International Society for Optics and Photonics, 2018, pp. 1–9.