Automatic Annotation to Train ROI Detection Algorithm For Premature Infant Respiration Monitoring in NICU

Ádám Nagy[a,b,*], Péter Földesy[a], Imre Jánoki[a,b], Máté Siket[a], Ákos Zarándy[a]

[a] *Institute for Computer Science and Control, 13-17. Kende street, Budapest, H-1111, Hungary*
[b] *Faculty of Information and Bionics, Pázmány Péter Catholic University, 50/A Práter street, Budapest, H-1083, Hungary*

Abstract

Visual monitoring of vital parameters of premature infants has become a heavily researched topic in recent years. Respiration rate (RR) is one of the most essential vital sign of newborns, therefore non-contact measurement of respiration is also a strongly studied area. Most of the published algorithms are able to provide better results if a suitable "region of interest" (ROI) detection takes place before the estimation of RR. This ROI is typically generated with a data-driven segmentation method. However, modern deep learning-based ROI detection algorithms require several thousands of annotated samples for training. Data collection and annotation is a long and tedious process. In this work, we propose a motion periodicity based solution to automatically detect the respiration mask containing the belly or the back of neonates. The places of the automatically generated masks showed a 96% agreement with the places of the manually marked regions. We showed that by using these automatically generated respiration masks for training U-Net variants we can not just avoid the manual labelling, but also reach greater accuracy in the ultimate RR calculation. We concluded, that it is possible and worthwhile to automatically generate

⋆This work is funded by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, under the 1019658 funding scheme and the framework of the Artificial Intelligence National Laboratory Program.

*Corresponding author
Email address: nagyadam@sztaki.hu (Ádám Nagy)

annotated dataset for deep learning based ROI detectors in the mentioned field.

Keywords: automatic annotation, semantic segmentation, breath rate, respiration rate, NICU monitoring, non-contact, remote, camera

1. Introduction

Non-contact monitoring of the respiration rate of newborns is a crucial and heavily researched topic [1], [2], [3], [4], [5], [6]. The respiration or breathing monitoring plays a critical role in the neonatal intensive care units (NICU) for tasks like detecting breathing abnormalities in the incubator at a very early age. The "obstructive sleep apnea" (OSA) is very frequent and can be a symptom of many disorders. The nasal occlusion can lead to switch to mouth breathing at the 40% of infants, which results in OSA [7]. Respiratory distress occurs in 6.4% of the cases [8]. Monitoring the breathing of premature babies also plays an important role in predicting certain diseases. For example, some articles have already shown a relationship between reflux and OSA [9]. One key feature to detect is apnea, the transient cessation of breathing [10]. It can also help determine the amount and quality of sleep that the neonates get, which is important for their development [11].

1.1. The problem of ROI detection for RR estimation

Several articles mention that ROI detection increases the performance of RR estimation algorithms [12], [13], [14], [15], [16], [17]. A review of RR estimation algorithms can be found in [1], which also overviews the applied ROI detection algorithms, which are typically data-driven nowadays and some of those are even convolutional neural network (CNN) based.

However, deep learning based ROI detection algorithms require a huge number of annotated images. Data collection and annotation is a long and tedious process and require dedicated software and human resources. In article [6] we previously introduced the concept of using a U-Net based neural network to

detect the torso as ROI for RR estimation. It was a challenge for us to annotate the right amount of data to train this neural network. It is desirable to automate the annotation process.

1.2. The proposed algorithm

In this paper we propose the Automatic Labeling Algorithm (ALA) for automatic annotation. This motion periodicity based algorithm was designed to mark those parts of the torso, which are actively moving during respiration. These areas, called Respiration ROI (R-ROI), are usually the lower parts of the belly or the back (slightly above and partially including the diaper area).

The working of ALA is based only on the area and periodicity of the movements shown in the image and doesn't require any human annotation. There are several benefit of using it. First of all, we can avoid the laborious manual data annotation. Moreover, we get an annotated database which is potentially better than human annotation when used as the bases of the training set for the segmentation network part of the respiration extraction pipeline. The reason is that a human annotator would mark the belly or back areas on static images without sensing the spatial-temporal dynamics of breathing, while our ALA solution analyzes a longer part of the video flow for generating the labels benefiting accuracy at the end of the pipeline. The RR measurement led to better performance when the ROIs are generated by a U-Net trained with the ALA generated annotation and evaluated on the database we collected (see Table 2). The reason why we do not use the ALA algorithm itself as ROI detector before the RR calculation is explained in the "Discussion" section below.

2. Related works

Novel algorithms for non-contact RR monitoring have been developed recently that use footage from infants in open incubators [18], [19]. Numerous similar algorithms are reviewed in [1].

While authors of [20] and [21] examine the problem of video-based RR monitoring using footage from adults and not infants. In the case of these algorithms, the ROI will be located around the person's chest.

Maurya et al. overviewed the various sensors, respiration rate estimation algorithms, and methods, and the applied ROI detectors as well [1]. They also mention deep learning-based neural architectures for ROI detectors that are used for respiration analysis. In [17], the authors showed a comparison of different ROI detectors and concluded that CNN-based ones are best fit for the task. Jorge et al. conclude that findings "illustrate the opportunities in non-contact vital sign monitoring as a result of the robust subject segmentation provided by CNNs" [12]. Nagy et al. also examined how much improvement ROI detection provides to performance [6]. It can be clearly seen that researchers are motivated to use CNN for robust ROI segmentation for monitoring physiological signs and there are efforts to step toward deep learning based segmentation.

Specifically, we can find articles in the literature, in which the authors are dealing with movement area and periodicity based ROI or object detection: [22], [23], [24], [25].

Naturally, we can also find many examples in the literature about automatic annotation similar to our method, we proposed in this article. Of course, these are not always camera-based methods and problems. In [26], [27], and [28], an automatically annotated or synthetic dataset was used to train neural networks to solve semantic segmentation problems that are related to medicine. There are also articles from synthetic dataset generation in the medical field: [29], [30], [31], [32]. As well as [33], in which the authors generated annotated healthy and abnormal electrocardiograms for arrhythmia detection.

Another example is [34] where the authors use CNN to analyze the appearance and, in parallel, uses optical flow to analyze the motion of the objects in order to segment moving objects on the video. Article [35] presents how the deep learning based optical flow estimation performs in different problems like unsupervised, and semi-supervised learning.

It can be seen that automatic annotation or annotated data set generation

is a frequently used procedure that researchers prefer to use when they need large amounts of annotated data. However, it is novel in the RR measurement according to our search in the literature. Moreover, the ALA algorithm is special in that it spans through from the dynamic label generation to train a static network, which finally, efficiently analyzes a dynamic scene by an intra-frame method.

The ROI detectors presented in [12], [14] and [13] perform really well indeed and are based on deep learning. However, the algorithms presented in these articles have been trained on large, human-annotated data set. The main contribution of our manuscript is that it shows a method to generate annotated dataset for R-ROI detection, that does not require any human resources and thus avoid inter-observer variability too.

3. Methodology

3.1. The Automatic Labeling Algorithm (ALA)

We propose a motion extension and periodicity based labeling algorithm to generate an automatically annotated dataset. The Automatic Labeling Algorithm (ALA) can be divided into two main parts which will be broken down into further parts later. These two main parts are the following:

- R-ROI detection
- R-ROI tracking

Detection of the R-ROI is done by analyzing the motion pattern and the periodicity of the motion. However, we don't detect the R-ROI in every frame. Once it is found, we can track it as long as it can be seen on the video. If something – e.g. the hands of a doctor – obscures the R-ROI, the tracking will be lost. In this case, we can turn back to the detection part again, re-initiate the coordinates of the center point and resume tracking. The automatic annotation and the marking of the R-ROI is done continuously during the tracking. An overview of this can be seen in Figure 1.

3.2. Detection

One of the main assumptions of our method is that if the camera watches a neonate in the incubator – as described at the experimental setup below Figure 6 –, and the infant is breathing normally, we can determine the position of the R-ROI, based on the periodicity of the respiration type motion.

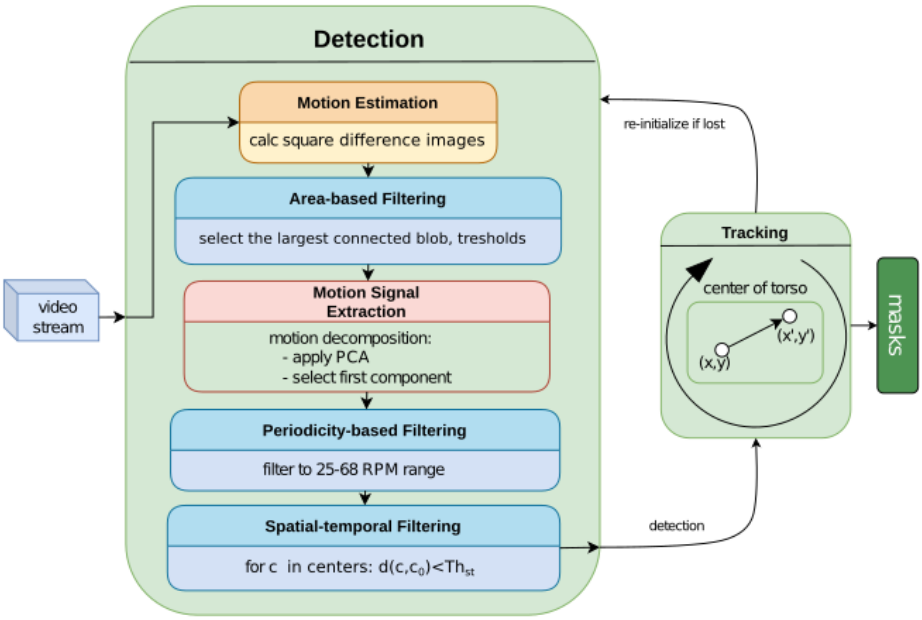

Figure 1: The automatic annotation algorithm (ALA). The input is a video stream, a series of consecutive frames, and the output is a series of binary images where the R-ROI of the infants is masked. The connection of the detection and the tracking is shown in the figure. The steps for "R-ROI Detection" and tracking are detailed below.

An overview of the periodicity-based R-ROI detection is shown in Figure 1, which summarizes the steps of this process too, under "Detection":

1. Motion Estimation
2. Area-based Filtering
3. Motion Signal Extraction
4. Periodicity-based Filtering

5. Spatio-temporal Filtering

The detection is accepted as a respiration-related region of the belly or back only if all of the filtering steps above return with positive result, meaning that the current examined motion period is respiration-related and not filtered out by spatial filters either. Otherwise, the current period is discarded and no detection will be made in the current detection step.

3.2.1. Motion Estimation

Several algorithms exist that are able to quantify various motions on video frames like optical flow, Deep Flow [36] or certain block-matching algorithms [37] that perform very well in in estimating the intensity and direction of the various movements. However, in this application, we do not need to estimate the direction of the movements. We applied the morphology and frequency analysis of the difference images for motion estimation because it directly provides a 1-channel motion intensity array in rotation invariant way.

For the detection of the R-ROI we calculated the square difference of the consecutive images ($\mathbf{D}$) to quantify the motion in the frames, that can be calculated by using Eq. (1) for all of the pixels. The square amplifies the stronger motions – attenuating the weaker ones. In fact, it highlights the most intense movements. On the difference image we were able to apply a motion area-based detection in order to find the R-ROI.

$$\mathbf{D}(x, y, n) = (\mathbf{I}(x, y, n) - \mathbf{I}(x, y, n-1))^2, \tag{1}$$

where $\mathbf{I}(x, y, n)$ means the grayscale pixel intensity in coordinates (x,y) at the n^{th} frame, and $\mathbf{D}(x, y, n)$ is the (x, y) pixel value of the difference image at the n^{th} frame.

3.2.2. Area-based Filtering

If the neonate's respiration is normal, calm, and no other motion can be detected, we can use a threshold on the difference image to get a binary mask.

After thresholding, a morphological dilation is applied, and then the algorithm searches for the largest blob on the frame, which is considered to be the R-ROI.

For the detection of the largest connected region we used the Spaghetti algorithm [38] which combines a block-based mask with state prediction to solve the "connected components labelling" problem. After as we mentioned above, the algorithm searches for the largest blob on the frame.

$$LR(x,y) = \begin{cases} 1, & \text{if } l(x,y) = L \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where l(x,y) is the label that is generated by the Spaghetti algorithm for (x,y) place from the current square difference and L is the label that can be associated to most of the pixels (which means it is the label of the largest region).

In this way, the largest connected region is selected. After this, a lower and an upper threshold for the area of the regions is applied and those frames are omitted which contain an unrealistically small or large detected region. In other words, $Th_{lower} < \sum_{x,y=1,1}^{M,N} LR(x,y) < Th_{upper}$ must be fulfilled.

3.2.3. Motion Signal Extraction

Area-based filtering works well when there are respiration movements only on the video. However, when the infant is active, the different limb movements can generate bigger or same size difference patterns. This is the reason why we use time domain filtering like periodicity-based filtering as well. For that, we need to generate a one dimensional motion signal, by applying PCA [39]. We use a sliding window which contains 300 frames including the current frame. First, each frame is flattened into a vector, and then these 300-long vectors are used as the rows of a large matrix:

$$\mathbf{X} \in R^{a \times b} \tag{3}$$

where a=300 and b is the number of pixels on the individual frames and X contains the values of the pixels of flattened frames in the sliding window mentioned

above.

The PCA transforms our data to a new coordinate system in a way that the data projection with the greatest variance comes to lie on the first coordinate, the projection with the second greatest variance on the second coordinate, and so on. After, we use the PCA result that is related to the first component, which provides a 1-dimensional motion signal ($\mathbf{s}_M$) containing 300 data points.

3.2.4. Periodicity-based Filtering

Detecting only the moving area is not enough to achieve reliable performance in distinguishing respiratory motion patterns from movements of other objects (limbs) with comparable area size. Another feature that is characteristic of respiratory motion patterns is the frequency of the respiratory motion signal. Newborns usually breath within the 25 RPM - 68 RPM frequency band [40]. The motion signal in the above-mentioned sliding window is filtered according to whether or not its frequency falls into the frequency range of 25 RPM to 68 RPM. For the calculation of frequency, we used the FFT spectrum of the previously mentioned motion signal and selected the frequency with the largest peak.

$$RPM = \mathbf{x}[\max_{f}\{FFT(\mathbf{s}_M)\}], \tag{4}$$

where $\mathbf{s}_M$ is the 1-dimensional motion signal extraction calculated from the previously mentioned sliding window, whereas $\mathbf{x}$ is a vector that contains the frequency bins of the FFT spectrum.

If the highest peak is not in the said frequency range, there is no detection and the particular frame is discarded, because we cannot guarantee that the selected moving region belongs to a belly or a back with respiratory movement. In other words, RPM must be in the range of $[Freq_{lower}, Freq_{upper}]$ or there is no detection. The choice of the frequency limits of this range originates in the physiologically relevant range (40–120 BPM).

3.2.5. Spatio-temporal Filtering

After the periodicity filtering, we applied spatio-temporal filtering, which limits the maximal displacement of the selected area. In a calm period, we can assume that the geometrical center of the largest region does not jump suddenly far from the initial center. Therefore, we applied another time window which contains the centers of the largest regions, (if they are detectable), calculated on the consecutive frames. If none of the elements of this window are farther from the initial center than a predefined threshold parameter, the current difference image is accepted. If the largest connected region is accepted as an R-ROI-related pattern, we refresh the tracked center with the geometric center of this pattern and start tracking it. In other words, Equation 5 have to be fulfilled to get a valid detection.

$$\forall c_i \in \mathbf{c}, d(c_i - c_0) < Th_{st} \tag{5}$$

where $i < n$, d is the "euclidean distance", $\mathbf{c}$ is a vector that contains n=300 peaces of 2D center points ($c_i \in R^2$).

3.3. R-ROI Tracking

After detection, we use sparse optical flow [41], [42] to track the R-ROI. In our case, the algorithm follows one point that is the geometric center of the largest region of the accepted difference image. During tracking, we draw a mask around the tracked center points that marks the place of the R-ROI, more specifically, the part of the R-ROI where respiratory movements can be observed.

A big advantage of optical flow is that it is relatively robust for illumination changes and shadows, and has a relatively low computational cost. It can track the point even if the neonate is moving strongly. However, it can lose the tracked point if the neonate's arm or an object crosses the path of the tracked point or if somebody – e.g. nurses – reaches into the picture when tending to the baby. In this case we go back to the detection phase as shown in Figure 1.

We considered using all of the original camera images and the automatically annotated, binary masks as input-label image pairs. However, as it turned out, the consecutive images are very close in time, hence they are very similar. Therefore, we picked one image pair in every 10 seconds to create our automatically labeled dataset (see Figure 2).

Figure 2: The algorithm draws mask around the tracked center points. For the automatically labelled dataset, these masked images saved in every 10 seconds.

3.4. Training

In this work, we used motion-based R-ROI detection and tracking to generate an automatically annotated dataset, with which we can train a more robust deep learning-based algorithm. We trained the U-Net [43] architecture that is a proven, well performing deep learning structure designed to solve semantic segmentation problems. It is easy to implement, fast, and requires relatively few labeled images – in our experience, a few thousand. We used a variant of *UNet++* [**?**] which performed better than the traditional U-Net. Also, the original architectures have kernels with a limited visual field, whereas we want to use a larger field of view at the smallest resolution, as the position of the limbs and head can be very informative when we are searching for the R-ROI. Therefore, instead of the traditional 3x3 convolutional kernels, we rather used 5x5 kernels. The architecture was modified in a way shown in Figure 3. Our U-Net variant is fed with 150x150 sized RGB images and returns with binary

masks, where the R-ROI of the infant is marked.

Our generated dataset contained 6000 images extracted from recordings of 10 different infants. Each recording is about 24-hours long and around 600 annotated images were saved by the algorithm from each, according to a pre-set parameter. The generated dataset was partitioned as follows:

- training set: 4200 images from 7 infants
- validation set: 600 images from 1 independent infant
- (holdout) test set: 1200 images from 2 independent infants

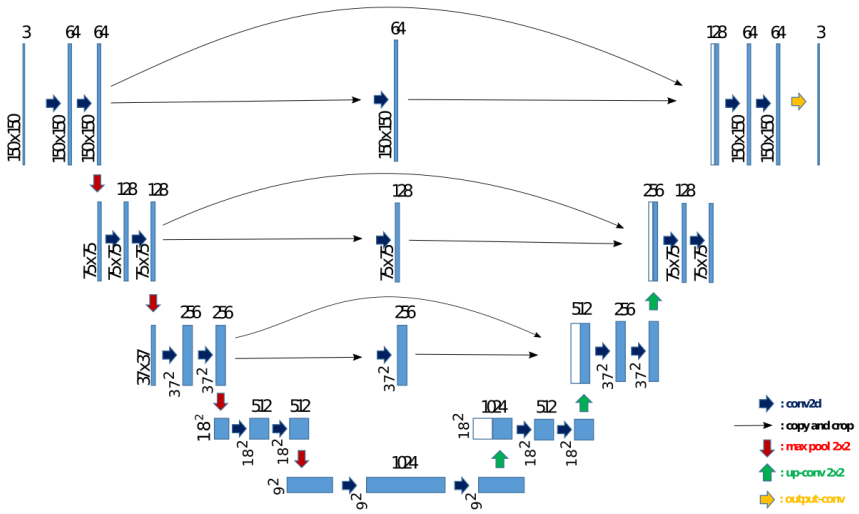

Figure 3: The modified *UNet++* architecture working on different spatial resolution than the traditional ones. The input image is of size 150x150 pixels. Note, that we used bigger kernels of size 5x5 instead of the traditional ones (3x3).

In the training we used mean squared error (MSE) as the loss function and an Adam optimizer was used to train the network. We applied "online augmentation" and early stopping with the validation set as well. We also used "Comet-ML" (Comet.ml, NY, USA) for the visualization of loss, epoch loss, and for hyperparameter tuning.

3.5. Respiration Calculation

In article [6] we introduced a respiration monitoring algorithm which uses a traditional U-Net as ROI detector, which was trained with hand annotated R-ROI images. The algorithm contains a sliding window, in which a dense optical flow [44] wave extractor and a peak detection-based calculator estimates the respiration rate. The flow-chart of the algorithm is illustrated in Figure 4. An implementation of the referred algorithm can be found in the following GitHub repository: [45].

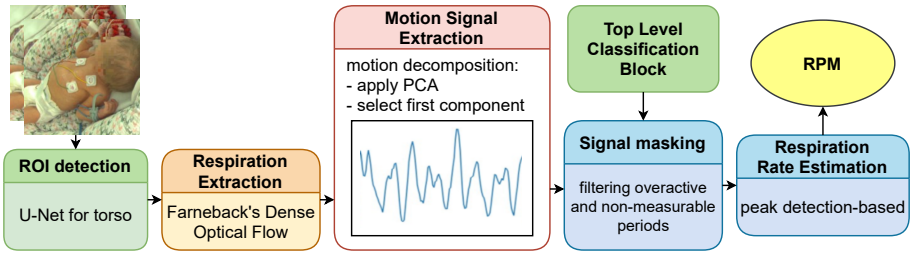

Figure 4: Overview of the Respiration-rate-estimator algorithm published in [6]. This solution applies a U-Net for ROI detection and generates RR from a stack of consecutive images.

3.6. Top Level Classifier

In article [6], we also present another module called "TopLevelClassifier", whose task was to identifies the current care status and measure the infant's activity. This module can not only quantify the baby's activity, but can also detect if the baby is in the incubator. Specifically, whether there is a baby in the input image. As well as being able to detect if care or other intervention happens and the caregiver's hand or another motion object appears in the image. This top level classification achieved 97.9 % sensitivity and 97.5 % specificity on the data set introduced in [6].

If we run the top-level classifier on the input video before using the ALA algorithm, or even if we run it in parallel with it, we will also have information

about whether the baby is in the incubator or if something (like the caregiver's hand) is disturbing the measurement.

4. Results

To understand the evaluation and for better readability, we describe our different, named datasets and architectures. The used datasets during the process:

- D_{man} : Dataset where the labeling was done manually by human annotators.
- D_{ala} : Dataset where the labeling was done automatically by ALA.

Additionally, we trained the following models that are evaluated below:

- A traditional U-Net trained on the D_{man} set. ($U\text{-}Net_0$)
- A traditional U-Net trained on the D_{ala} set. ($U\text{-}Net_1$)
- And a *UNet++* [**?**] trained on the D_{ala} set. (*UNet++*)

Note, that unlike common segmentation tasks, our primary goal is not pixel-level accuracy. Rather, our ultimate goal is RR measurement accuracy, although we examine the segmentation performance as well in this section.

Evaluation was done on two levels. First, the segmentation performance was characterized, and in the second step, the RR measurement accuracy was analyzed.

4.1. Segmentation performance analyses

Here we defined a morphological similarity measure which shows how much a U-Net generated mask is similar to a ALA generated mask.

$$S_{geo}(M_U, M_A) = \frac{\text{number of pixels in } (M_U \cap M_A)}{\text{number of pixels in } (M_A)}, \tag{6}$$

where M_U indicates the binary mask generated by $U\text{-}Net_1$, while M_A is the binary mask generated by ALA. S_{geo} is the ratio of the blue and the red+blue pixels in Figure 5.

Two kinds of statistics were calculated. First we evaluated how large a percentage of the masks generated by the trained $U\text{-}Net_1$ has an S_{geo} ratio higher than 0.5 and 0.9. The results can be seen in the first and second rows of Table 1. The third row of Table 1 shows the average (avg) S_{geo} similarity value of the $U\text{-}Net_1$ generated masks and the ALA generated masks. The relatively low similarity value (0.787) compared to the high number of over 0.9 S_{geo} masks means that most of the time the network finds a correct, circular shaped ROI similar to D_{ala} labels, however, in a low number of cases it masks a different area with a different shape, that results in a few very low S_{geo} values. However, it should be noted that in the ALA, the radius of the circular output mask and the circular shape were chosen freely following the size of the infant on the video. Therefore, a perfect match cannot be expected. Additionally, after visual inspection of these masks, we found that even in these cases parts of the torso with respiration-like motion was marked. This actually shows the good generalization capability of the trained $U\text{-}Net_1$, because it works from on single frames, and still can identify those areas of the incubator scene, which are typically performing respiratory-like movement. Therefore, we also evaluated the $U\text{-}Net_1$ generated masks by comparing them to the manual annotation (D_{man}) where the whole torso of the infants were masked, to prove that the predicted masks are indeed located almost entirely on the trunk. In the forth row we can see that the $U\text{-}Net_1$ trained on D_{ala} achieves a high average S_{geo} percentage between its output and the manually annotated body masks (D_{man}). Figure 5 shows the detection masks on typical frames.

4.2. RR performance analyses

In the second level of the evaluation, we examined how well the ROI detection works in practice when we estimate the RR. For that we used our RR estimator published in [6] first with the original U-Net ($U\text{-}Net_0$) and then re-

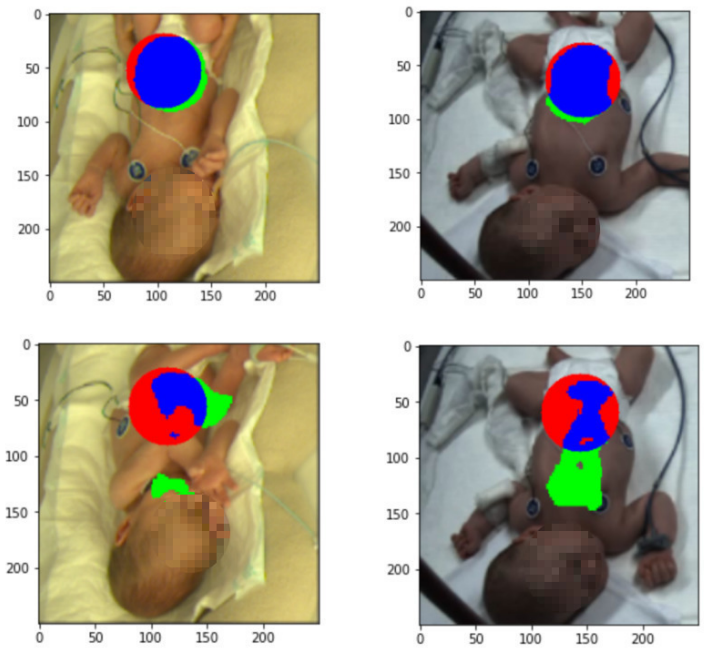

Figure 5: Results of the ROI detector where the detection was performed by a $U\text{-}Net_1$ trained on D_{ala}. The output of ($U\text{-}Net_1$) is green+blue, while the expected mask (generated by ALA) is red+blue. (Blue is their intersection.) In the first row, you can see some good examples where more than 50% of the detection (#blue pixels/#red+blue pixels) overlap with the expected mask (96% of the images), while the second row shows some bad examples where the overlap is less than 50%.

placing it with $U\text{-}Net_1$, and finally with $UNet{+}{+}$ the ones which we trained on the automatically generated dataset (D_{ala}). The respiration rate based evaluation is summarized in Table 2 where the individual rows are the following:

1. shows how well the respiration-network can estimate the respiration rates without ROI detection.
2. shows how well the respiration-network can estimate the respiration rates using the traditional $U\text{-}Net_0$ as ROI detector, which was trained on D_{man}.
3. shows how well the respiration-network can estimate the respiration rates using the traditional $U\text{-}Net_1$ as ROI detector, which was trained on D_{ala}.
4. shows how well the respiration-network can estimate the respiration rates

Table 1: Evaluation of the U-Net-based ROI generation ($U\text{-}Net_1$) where the network was trained on training set from D_{ala}. The first two rows are frame based statistics (number of good frames/all frames). The third row shows the average of the particular S_{geo} values. The fourth row shows average S_{geo} value as well, however here the similarity is calculated with the manually generated labels with the entire torso belongs to D_{man}.

U-Net based ROI detector	performance
S_{geo}>0.5 with ALA masks	96%
S_{geo}>0.9 with ALA masks	93.5%
Avg S_{geo} with ALA masks	0.787
Avg S_{geo} with manual whole torso masks	0.987

using the traditional *UNet++* as ROI detector, which was trained on D_{ala}.

The mean absolute error (MAE) and root mean squared error (RMSE) were calculated between the ECG ground-truth reference of the test set and the outputs of the methods.

As it can be seen, the $U\text{-}Net_1$ and *UNet++* as ROI detectors significantly exceed the performance of the manually annotated case and all of the neural architectures provide better performance than the case where we did not use any ROI detector. It is important to note, that for the evaluation we selected a typical, 26-minutes long, continuous, period from the video record of the independent holdout test set, where no caring or other intervention can be observed, only the respiration and short limb movements.

Table 2: Evaluation of a U-Net-based ROI detection and breath rate monitoring algorithm in which the neural network was trained on D_{ala}.

Algorithm	**MAE**	**RMSE**
without ROI detector	1.488 RPM	1.808 RPM
$U\text{-}Net_0$, trained on D_{man}	1.348 RPM	1.762 RPM
$U\text{-}Net_1$, trained on D_{ala}	1.223 RPM	1.507 RPM
UNet++ trained on D_{ala}	1.094 RPM	1.348 RPM

As mentioned above, we examined the generated masks manually and we observed that all the predicted masks were located on places where breathing was clearly detectable, on the torso. In other words, the detector found the correct ROI for each frame. Nothing proves this better than the fact that the network provides higher performance, if it is trained on D_{ala}, as Table 2 shows.

4.3. Dataset

Figure 6: In the experimental setup the neonatal infant in NICU (1) was monitored with a Basler acA2040-55uc RGB camera (3). At the same time, physiological signals were received from the Philips IntelliVue MP20/MP50 monitors (2). The wave and rate data from the monitors and the video images from the camera were saved in sync to a laptop (4).

Image data for automatic annotation was collected with an industrial camera that monitored the incubators. Physiological signs were collected in sync as well. The whole dataset was collected in the NICU of the Ist Dept. of Pediatrics, II. Dept. of Obstetrics and Gynecology, of Semmelweis University, Budapest, Hungary. The experimental setup for the data collection can be seen in Figure 6. You can read more about the data collection system in [6].

The images from the camera were saved with a resolution of 500 × 500 pixels at 20 frames per second in raw format. For the sake of generality, the records were taken from many different angles, and several recordings were made of different infants. The distance between the "Basler acA2040-55uc" camera and the subject varied between 80 cm and 1,5 m. The demographics of the participants, like age, weight, sex, etc., and whether they were given respiratory support or any drugs are summarized in Table 3.

Table 3: Demographic properties of the population of participants

Subject	1	2	3	4	5	6	7	8	9	10
Recording time (hours)	96.7	5.5	39.4	27.4	51	105.5	50.1	36.4	56	38.2
Gender	F	M	M	F	F	M	F	F	F	M
Gestational age (weeks)	32	32+3	31+4	35+4	39	32	33	38+6	24+2	33+4
Birth weight (g)	2020	1840	1850	1870	3150	2120	2080	2840	760	2100
Postnatal age (days)	4	4	10	8	4	7	2	7	11	1
Actual weight (g)	1900	1850	1680	1820	2905	2040	1960	3150	750	-
Length (cm)	46	44	-	45	57	45	44	48	46	45
Head circumference (cm)	32	29.5	-	32	34	30	32	33	22	-
Respiratory support	no	no	no	no	no	no	no	yes	yes	no
Any drugs	no	no	no	no	yes	no	no	yes	yes	yes
Fitzpatrick scale	2	3	2	2	2	2	2	2	2	2

5. Discussion

A significant limitation of ALA's detection part, as a direct bases for a training set is that in mission critical situations, like in the case of apnea, it fails to provides valid detection and generate an R-ROI. In general, the periodicity-based ROI detection works with dynamic motion data and does not always

find respiration related regions, therefore it is not reliable in continuous R-ROI detection. It fails in static situations, or it might fail when there is some other periodic motion like a blinking lamp, moving shadows, caring or feeding induced motions. However, in such situations it does not generate a label at all, therefore those frames will not be used in training. In this way, it is capable of automatically generating an annotated dataset from a set of common case videos, which can be used to train a more robust deep learning-based ROI detector (U-Net) which works by analyzing static image features and capable of detecting ROI even if the subject is moving unrelated to respiration, or does not move at all.

6. Conclusion

We have created a motion-based algorithm (ALA) for automatic labelling of infants' body parts with respiratory movements (R-ROIs). We trained a neural architecture with these labels (D_{ala}), which was then able to detect the R-ROI satisfactorily on incubator videos. We also demonstrated that the Respiration Rate calculation algorithm [6] performed better when its built-in neural R-ROI detector was trained with the automatically generated labels than with the manually drawn labels.

In this way, we have shown that our presented concept works. In other words, the automatic annotation based on the extension and periodicity of the motion of objects, and our training of a more robust neural network-based algorithm on it, perform satisfactorily in the task of ROI detection for the respiration monitoring of infants.

We believe that the elaborated concept can also be used for other problems where we are looking for objects on a sequence of images that are moving with a given motion extent and periodicity.

A future development could be that we modify and run the ALA on nightly infrared recordings too, where colour data are not available. Also, there is possibility to locate the ROI directly from the difference image for covered

babies.

Author contributions. Conceptualization, Á.N. ,P.F. ,I.J. ,M.S. and Á.Z.; methodology, software, validation and formal analysis, Á.N. ,I.J. ,M.S.; investigation, Á.N. ,I.J. ,M.S.; resources, Á.Z.; data curation, P.F,I.J.; writing—original draft preparation, Á.N., I.J.; writing—review and editing, Á.N. ,I.J., Á.Z.; visualization, Á.N.; supervision and project administration, P.F., Á.Z.; All authors have read and agreed to the published version of the manuscript.

Informed consent. Ethical review and approval were waived for this study due to its non-invasive and non-contact manner, and because it did not influence patient comfort or care. Written informed consent was obtained from parents to use the anonymised visual records for study purposes. Informed consent documents are stored as part of patient documentation. The study was lead by the head of the Division of Neonatology, Semmelweis University, Hungary.

Data availability. Restrictions apply to the availability of these data. Data was obtained from the participants given consent and are available from the authors only with the permission of the participants.

Conflicts of interest. The authors declare no conflict of interest.

References

[1] L. Maurya, P. Kaur, D. Chawla, P. Mahapatra, Non-contact breathing rate monitoring in newborns: A review, Computers in Biology and Medicine 132 (2021) 104321. doi:10.1016/j.compbiomed.2021.104321.

[2] J. Jorge, M. Villarroel, S. Chaichulee, A. Guazzi, S. Davis, G. Green, K. McCormick, L. Tarassenko, Non-contact monitoring of respiration in the neonatal intensive care unit, 2017, pp. 286–293. doi:10.1109/FG.2017.44.

[3] Y. Sun, W. Wang, X. Long, M. Meftah, T. Tan, C. Shan, R. Aarts, P. With, Respiration monitoring for premature neonates in nicu, Applied Sciences 9 (2019) 5246. doi:10.3390/app9235246.

[4] A. Nagy, D. Chetverikov, A. Zarándy, Novel methods for video-based respiration monitoring of newborn babies, in: 2019 In: Képfeldolgozók, és Alakfelismerők Társasága Képfeldolgozók és Alakfelismerők Társaságának 12. Országos Konferenciája, no. 22, 2019, pp. 1–10. doi:http://eprints.sztaki.hu/id/eprint/9699.

[5] A. Zarandy, P. Foldesy, A. Nagy, I. Jánoki, D. Terbe, M. Siket, M. Szabo, J. Varga, Multi-level optimization for enabling life critical visual inspections of infants in resource limited environment, 2020, pp. 1–5. doi:10.1109/ISCAS45731.2020.9181040.

[6] A. Nagy, I. Jánoki, P. Foldesy, D. Terbe, M. Siket, M. Szabo, J. Varga, A. Zarandy, Continuous camera-based premature-infant monitoring algorithms for nicu, Applied Sciences 11 (2021) 7215. doi:10.3390/app11167215.

[7] E. Katz, R. Mitchell, C. D'Ambrosio, Obstructive sleep apnea in infants, American journal of respiratory and critical care medicine 185 (2011) 805–16. doi:10.1164/rccm.201108-1455CI.

[8] T. Ghafoor, S. Mahmud, S. Ali, S. Dogar, Incidence of respiratory distress syndrome, Journal of the College of Physicians and Surgeons–Pakistan : JCPSP 13 (2003) 271–3.

[9] P. Demeter, A. Pap, The relationship between gastroesophageal reflux disease and obstructive sleep apnea, Journal of gastroenterology 39 (2004) 815–20. doi:10.1007/s00535-004-1416-8.

[10] S. Sale, Neonatal apnoea, Best practice research. Clinical anaesthesiology 24 (2010) 323–36. doi:10.1016/j.bpa.2010.04.002.

[11] X. Long, R. Otte, E. Sanden, J. Werth, T. Tan, Video-based actigraphy for monitoring wake and sleep in healthy infants: A laboratory study, MDPI Sensors 19 (2019) 1075. doi:10.3390/s19051075.

[12] J. Jorge, M. ViIllarroel, S. Chaichulee, K. McCormick, L. Tarassenko, Data fusion for improved camera-based detection of respiration in neonates, 2018, p. 36. doi:10.1117/12.2290139.

[13] G. Scebba, G. Da Poian, W. Karlen, Multispectral video fusion for non-contact monitoring of respiratory rate and apnea, IEEE Transactions on Biomedical Engineering PP (2020) 1–1. doi:10.1109/TBME.2020.2993649.

[14] S. Chaichulee, M. Villarroel, J. Jorge, C. Arteta, K. McCormick, A. Zisserman, L. Tarassenko, Cardio-respiratory signal extraction from video camera data for continuous non-contact vital sign monitoring using deep learning, Physiological Measurement 40. doi:10.1088/1361-6579/ab525c.

[15] C. Pereira, X. Yu, T. Goos, I. Reiss, T. Orlikowsky, K. Heimann, B. Venema, V. Blazek, S. Leonhardt, D. Teichmann, Noncontact monitoring of respiratory rate in newborn infants using thermal imaging, IEEE Transactions on Biomedical Engineering PP (2018) 1–1. doi:10.1109/tbme.2018.2866878.

[16] A. K. Abbas, K. Heimann, K. Jergus, T. Orlikowsky, S. Leonhardt, Neonatal non-contact respiratory monitoring based on real-time infrared thermography, Biomedical engineering online 10 (2011) 93. doi:10.1186/1475-925X-10-93.

[17] M. Villarroel, S. Chaichulee, J. Jorge, S. Davis, G. Green, C. Arteta, A. Zisserman, K. McCormick, P. Watkinson, L. Tarassenko, Non-contact physiological monitoring of preterm infants in the neonatal intensive care unit, npj Digital Medicine 2 (2019) 128. doi:10.1038/s41746-019-0199-5.

[18] S. Rossol, J. Yang, C. Toney-Noland, J. Bergin, C. Basavaraju, P. Kumar, H. Lee, Non-contact video-based neonatal respiratory monitoring, Children 7 (2020) 171. doi:10.3390/children7100171.

[19] P. Foldesy, A. Zarandy, M. Szabo, Reference free incremental deep learning model applied for camera-based respiration monitoring, IEEE Sensors Journal PP (2020) 1–1. doi:10.1109/JSEN.2020.3021337.

[20] M. Reyes, J. Dorta Palmero, J. Diaz, E. Aragon Perez, A. Taboada-Crispi, Computer Vision-Based Estimation of Respiration Signals, 2020, pp. 252–261. doi:10.1007/978-3-030-30648-9_33.

[21] C. Massaroni, D. Presti, D. Formica, S. Silvestri, E. Schena, Non-contact monitoring of breathing pattern and respiratory rate via rgb signal measurement, Sensors 19 (2019) 2758. doi:10.3390/s19122758.

[22] C. Benabdelkader, L. Davis, Detection of people carrying objects : A motion-based recognition approach, 2002, pp. 378 – 383. doi:10.1109/AFGR.2002.1004183.

[23] K. Al-mutib, M. Emaduddin, M. Alsulaiman, R. Hedjar, E. Mattar, Motion periodicity based pedestrian detection and particle filter based pedestrian tracking using stereo vision camera, 2013.

[24] L. Zijuan, L. Wang, W. Liu, B. Li, Human movement detection and gait periodicity analysis using channel state information, 2016, pp. 167–174. doi:10.1109/MSN.2016.035.

[25] Y. Liu, X. Gu, L. Huang, J. Ouyang, M. Liao, L. Wu, Analyzing periodicity and saliency for adult video detection, Multimedia Tools and Applications 79. doi:10.1007/s11042-019-7576-6.

[26] M. Reza, A. Naik, C. Kai, D. Crandall, Automatic annotation for semantic segmentation in indoor scenes, 2019. doi:10.1109/IROS40897.2019.8968230.

[27] G. Ros, L. Sellart, J. Materzynska, D. Vázquez, A. López, The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes, 2016, pp. 3234–3243. doi:10.1109/CVPR.2016.352.

[28] A. Marcu, D. Costea, V. Licaret, M. Leordeanu, Towards automatic annotation for semantic segmentation in drone videos (10 2019).

[29] R. Chen, M. Lu, T. Chen, D. Williamson, F. Mahmood, Synthetic data in machine learning for medicine and healthcare, Nature Biomedical Engineering 5 (2021) 1–5. doi:10.1038/s41551-021-00751-8.

[30] Q. Tang, Z. Chen, R. Ward, M. Elgendi, Synthetic photoplethysmogram generation using two gaussian functions, Scientific Reports 10 (2020) 13883. doi:10.1038/s41598-020-69076-x.

[31] H.-C. Shin, M. Orton, D. Collins, S. Doran, M. Leach, Autoencoder in time-series analysis for unsupervised tissues characterisation in a large unlabelled medical image dataset, Vol. 1, 2011, pp. 259–. doi:10.1109/ICMLA.2011.38.

[32] Q. Tang, Z. Chen, J. Allen, A. Alian, C. Menin, R. Ward, M. Elgendi, Ppgsynth: An innovative toolbox for synthesizing regular and irregular photoplethysmography waveforms, Frontiers in Medicinedoi:10.3389/fmed.2020.597774.

[33] F. Zhu, Y. Fei, Y. Fu, Q. Liu, B. Shen, Electrocardiogram generation with a bidirectional lstm-cnn generative adversarial network, Scientific Reports 9. doi:10.1038/s41598-019-42516-z.

[34] P. Tokmakov, K. Alahari, C. Schmid, Learning video object segmentation with visual memory, 2017, pp. 4491–4500. doi:10.1109/ICCV.2017.480.

[35] J. Hur, S. Roth, Optical Flow Estimation in the Deep Learning Age, 2020, pp. 119–140. doi:10.1007/978-3-030-46732-6_7.

[36] P. Weinzaepfel, J. Revaud, Z. Harchaoui, C. Schmid, Deepflow: Large displacement optical flow with deep matching, 2013, pp. 1385–1392. doi:10.1109/ICCV.2013.175.

[37] G. Haan, P. Biezen, H. Huijgen, O. Ojo, True motion estimation with 3-d recursive search block matching, Circuits and Systems for Video Technology, IEEE Transactions on 3 (1993) 368 – 379, 388. doi:10.1109/76.246088.

[38] F. Bolelli, S. Allegretti, L. Baraldi, C. Grana, Spaghetti labeling: Directed acyclic graphs for block-based connected components labeling, IEEE Transactions on Image Processing PP (2019) 1–1. doi:10.1109/TIP.2019.2946979.

[39] I. T. Jolliffe, Principal Component Analysis, Springer Series in Statistics. New York: Springer-Verlag, 2002. doi:https://doi.org/10.1007/b98835.

[40] S. Fleming, M. Thompson, R. Stevens, C. Heneghan, A. Plüddemann, I. Maconochie, L. Tarassenko, D. Mant, Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: A systematic review of observational studies, Lancet 377 (2011) 1011–8. doi:10.1016/S0140-6736(10)62226-X.

[41] B. Horn, B. Schunck, Determining optical flow, Artificial Intelligence 17 (1981) 185–203. doi:10.1016/0004-3702(81)90024-2.

[42] B. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision (ijcai), Vol. 81, 1981.

[43] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, Vol. 9351, 2015, pp. 234–241. doi:10.1007/978-3-319-24574-4_28.

[44] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, Vol. 2749, 2003, pp. 363–370. doi:10.1007/3-540-45103-X_50.

[45] A. Nagy, Neonatal-respiration-monitoring-algorithm, https://github.com/cezius/Neonatal-Respiration-Monitoring-Algorithm (2022).