Post-print of: A. Kwasniewska, M. Szankin, J. Ruminski and M. Kaczmarek, "Evaluating Accuracy of Respiratory Rate Estimation from Super Resolved Thermal Imagery," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, pp. 2744-2747, doi: 10.1109/EMBC.2019.8857764.

# Evaluating Accuracy of Respiratory Rate Estimation from Super Resolved Thermal Imagery

Alicja Kwasniewska[1,2] *Member IEEE EMBS*, Maciej Szankin[2] *Member IEEE EMBS*, Jacek Ruminski[1] *Member IEEE EMBS*, and Mariusz Kaczmarek[1] *Member IEEE EMBS*

***Abstract*— Non-contact estimation of Respiratory Rate (RR) has revolutionized the process of establishing the measurement by surpassing some issues related to attaching sensors to a body, e.g. epidermal stripping, skin disruption and pain. In this study, we perform further experiments with image processing-based RR estimation by using various image enhancement algorithms. Specifically, we employ Super Resolution (SR) Deep Learning (DL) network to generate hallucinated thermal image sequences that are then analyzed to extract breathing signals. DL-based SR networks have been proved to increase image quality in terms of Peak Signal-to-Noise ratio. However, it hasn't been evaluated yet whether it leads to better RR estimation accuracy, what we address in this study. Our research confirms that for estimator based on the dominated peak in the frequency spectrum Root Mean Squared Error improves by 0.15bpm for 8-bit and by 0.84bpm for 16-bit data comparing to original sequences if hallucinated frames are used. Mean Absolute Error is reduced by 0.63bpm for average aggregator and by 2.06bpm for skewness. This finding can enable various remote monitoring solutions that may suffer from poorer accuracy due to low spatial resolution of utilized thermal cameras.**

## I. INTRODUCTION

Respiratory rate (RR) is one of the most critical vital sign indicating changes of the physiological status of the subject [1]. Conventionally, RR is obtained through various wires and electrodes attached to a body, yet non-contact estimation of RR from thermal image sequences has revolutionized the process of establishing the measurement. With the means of image processing techniques, estimation of RR in various challenging conditions (e.g. in preterm infants, traumatized or burned patients, as well as in telemedicine applications) became more feasible. This is because adhesive electrodes or thoracic belts used to hold a sensor can cause epidermal stripping, skin disruption and pain in infants [2] and they do not stick to burns or bloody surfaces [3]. Also, a proper placement of electrodes on home-monitored patients may be difficult without the assistance of a specialist, and more importantly, may influence the physiological parameters being measured. The use of a thermal camera eases the setup of RR monitoring system and helps surpassing issues related to attaching sensors to a body.

In home-based solutions the cost is an important factor influencing the choice of data acquisition hardware. Even though recent advances in heat-based imaging have made thermal cameras more compact and affordable for commercial applications, the resolution of images is still much lower than in visible light spectrum, e.g. recent attempts to RR extraction utilized 1024x768 [2], 640x480 [4] 320x240 [5], or even 80x60 [6] images. Although low resolution data have been proved to provide sufficient RR estimation accuracy, we believe that by applying computer vision algorithms aimed at image enhancement, results can be further improved.

In particular, the aim of this study is to evaluate whether accuracy of non-contact RR estimation can be increased by generating super resolved (SR) thermal image sequences using deep neural networks (DNN). To the best of our knowledge, this is probably the first attempt to extract respiratory signal from hallucinated thermal faces, generated with DNN. We also perform additional experiments to determine if performance degrades with simulated increase of a distance from the camera. Additionally, the proposed super-resolution DNN-based solution for RR estimation is compared with Eulerian Video Magnification (EVM), the algorithm already successfully used for enhancing breathing signal [7], [8].

The rest of the paper is structured as follows: Section II overviews SR deep networks aimed at image enhancement. In Section III we describe methods used to collect and generate sequences, from which RR was extracted. Preliminary results of breathing rate estimation accuracy are presented in Section IV and discussed in Section v. Finally, Section VI concludes the paper and provides ideas for further studies.

## II. STATE OF THE ART

Application of Convolutional Neural Networks (CNN) to Super Resolution (SR) task is a relatively new idea. The pioneer work in this area dates only a few years back with the invention of the model called SRCNN [9], which allows for achieving the state-of-the-art restoration quality, while representing all components using a single CNN to preserve a lightweight structure and jointly optimize all layers. Since then, DNN-based solutions have been continuously refined to further improve accuracy. Kim et al. introduced Deeply-Recursive Convolutional Network (DRCN) which utilizes skip connection that correlates low resolution (LR) input with high resolution (HR) ground-truth data, what helps

*This work was partially supported by Statutory Funds of Electronics, Telecommunications and Informatics Faculty, Gdansk University of Technology and by Intel Corporation, AIPG, AI Lab.

[1]Alicja Kwasniewska, Jacek Ruminski and Mariusz Kaczmarek are with Department of Biomedical Engineering, Faculty of Electronics, Telecommunications and Informatics Gdansk University of Technology, Narutowicza 11/12 80-233, Gdansk, Poland {alicja.kwasniewska, jacek.ruminski, mariusz.kaczmarek@pg.edu.pl}

[2]Alicja Kwasniewska and Maciej Szankin are with Intel AI Lab, Artificial Intelligence Products Group, 12220 Scripps Summit Dr, San Diego, CA 92131, USA {alicja.kwasniewska, maciej.szankin@intel.com}

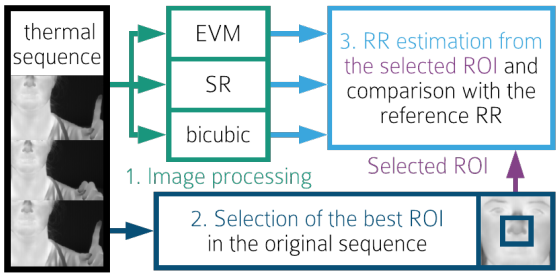


Fig. 1: The flow of applied methods. The RR was estimated from all sequences using the same ROI, selected from the original sequence.

with restoring detailed features [10]. In addition, recursive-supervision helps to minimize the exploding/vanishing gradients problem. Further modifications to SR included application of e.g. residual mappings and gradient clipping (Deeply Recursive Resiudal Network (DRRN) [11]). Some attempts to generate super-resolved thermal images using deep learning (DL) have also been done. Zhang et al. [12] proposed to feed high frequency information restored using Comprehensive Sensing (CS) Theory to SRCNN-like model structure to alleviate some fixed pattern noise present in the output from CS. Almasari F. and Debeir O. [13] introduced multimodal RGB-thermal fusion model that integrates components from SRCNN with residuals. In later studies, Spatial Transformer architecture was chosen because of its robustness with handling high-level variances of thermal image patterns and good performance on low spatial resolution data [14]. Other state-of-the-art CNN architectures also served as the inspiration for DL SR thermal enhancement solutions, e.g. SqueezeNet based Thermalnet [15], denoising CNN [16] which employs residual blocks, similarly to DRCN [10], or VSRnet-inspired CNN that simultaneously extracts features from visible and near infrared image [17]. Although described models improve image quality in terms of Peak Signal-to-Noise ratio (PSNR) and Structural Similarity index (SSIM), it hasn't been evaluated yet whether higher values of these metrics lead to better RR estimation accuracy. Thus, our research differs from already published studies in the following ways: i) we use SR DNNs for thermal face hallucination task, what is probably the first attempt to this problem. Pixel values changes caused by vital signs may be very subtle, so it's important to evaluate whether SR models are able to restore these detailed components. ii) RR is extracted from super-resolved sequences to determine whether higher PSNR lead to better RR estimation accuracy.

## III. METHODOLOGY

The flow of the methods applied in the study is presented in Fig. 1 and described in details in the following subsections.

### A. Data Collection

Experimental trials were performed on data collected from 40 volunteers (19 male, 21 female, age: $34.1 \pm 12$ ) to verify the proposed SR method in the real practice. Subjects were asked to breath through a nose, while looking towards the camera, placed approximately 120cm from the volunteer. In addition, volunteers were instructed to point finger upward during inhalation and down during exhalation in order to obtain the reference value of breaths per minute (bpm) by calculating the number of finger flexion (see Fig. 2). 2-minute sequence was recorded for each volunteer using thermal camera Flir SC3000 (320x240 spatial resolution, 30 FPS, temperature range from -20°C to +80°C, Camera Field of View 20°, set measuring range 24.9-36.9°C, measurements in the High Noise mode (noise reduction). The collected data, represented as arrays with digital values of 14-bit resolution, were used to form output images by assigning shades of gray to the intensities values. Since conversion from intensities with higher than 8-bit resolution to 8-bit color models is lossy, the contrast between regions may be reduced, eliminating some important details. Thus, in this study, we generated 8 and 16-bit output images from all acquired data to preserve detailed facial components. Conversion was done by mapping minimum and maximum values in the frame to the output ranges (0-$2^{bits}$).

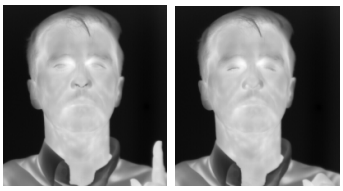

Fig. 2: Subject pointing finger upward (inhalation) and down (exhalation); visible change of color in nostrils: darker for inhalation (colder air), lighter for exhalation (air warmed up by a body).

### B. Image Quality Enhancement and Degradation

Image enhancement was performed in a twofold manner: using Eulerian Video Magnification (EVM) [18] and Deep Neural Networks: DRCN [10], DRRN [11] and a custom model optimized for thermal imagery. EVM algorithm allows for reveling subtle color changes, invisible to a naked eye. For this reason, a sequence is at first filtered, and then amplified. In our study, the filtering frequency range was set individually for each volunteer. Given the reference value of breaths per minute $bpm_{ref}$, we set the filtering frequency range as $bpm_{ref}/60$ rounded down and up to first decimal position for the left and right range margins(i.e. for bpm=14, the filtering range was 0.2-0.3Hz). The amplification was set to 20, as verified in [6].

For SR DNN models training we extracted every $300^{th}$ frame from all sequences to ensure data variability and then divided them randomly into training, test and validation subsets (70:15:15 split, total of 480 images). Before feeding frames to the models, we followed a standard procedure applied in Super-Resolution algorithms for simulating degradation of image resolution [9]. At first we down-scaled and then up-scaled all images by a factor of 2 using bicubic interpolation. As a result, we generated sequences (LR inputs) that simulate loss of resolution due to the increased distance from the camera. Original high resolution (HR) images were used as ground-truth data, against which the output from SR DNN models was compared during training. The objective was to teach the model to generate

the super-resolved outputs from LR inputs as similar to HR data as possible. In addition, we also adjusted SR CNN architecture. Our modification was based on the fact that thermal images are characterized by blurring effect and lower contrast between adjacent regions due to the heat flow, hence relatively shallow feature extraction step (DRRN uses only 1 convolution, DRCN uses only 2 convolutions at this step) may not be sufficient. Intuitively, widening of the receptive field should lead to better results on thermal imagery. Taking it into account, the changes proposed by us include residual blocks added not only to non-linear mapping, as in DRRN, but also to the feature extraction part (see Fig. 3). Besides residuals, we also applied recursions to ease the training process. Moreover, in our model weights are shared across all recursions, so the number of parameters did not increase, while the receptive field was covering more distant features.

DRCN[10] and DRRN[11] were trained using their default hyperparameters, with the except of the number of filters that for both models was set to 96 filters of a size 3x3 to ensure fair comparison. The selection of the number of recursions *D*, number of residuals in feature extraction part *E*, and number of residuals within each recursion *U* for the modified architecture proposed by us was performed using random search (from $D : \{1, 3, 5, 7, 9\}$, $E : \{1, 3, 5, 7, 9\}$, $U : \{1, 3, 5, 7, 9\}$ sets). Remaining parameters were set as: 41x41 training data crop with a stride of 21, Adam optimizer, momentum 0.9, weight decay 0.0001. Initial learning rate was set to $10^{-2}$ and reduced by an order of magnitude after 5 subsequent epochs, for which the validation error does not decrease. In total, more than 60 configurations of the proposed network were trained by us. Once training of all networks was done, PSNR and SSIM metrics calculated for test sets were used to determine which architecture is the most suitable for thermal data. Preliminary results showed that proposed SR CNN with configuration *D=9, E=3, U=0* outperformed DRCN by 16dB on 8-bit data and by 14.1dB on 16-bit data, achieving PSNR of 47.49dB and 48.05dB on 8 and 16-bit sequences receptively; the performance gain of the proposed network comparing to DRRN was 4.4dB and 3.8dB on 8 and 16-bit data. Thus, for further analysis we use super-resolved sequences generated with the proposed model. After data generation step, 8 sequences for each of 40 volunteers were prepared (original 8 and 16-bit (O8 and O16), bicubic 8 and 16-bit (B8 and B16), super-resolved 8 and 16-bit (S8 and S16), EVM 8 and 16-bit (E8 and E16)).

### C. Breathing Rate Estimation

For further analysis, short data segments from the beginning of each sequence were selected ( 400 samples) to reduce possible motion artifacts. At first, two regions of interest (ROI) were manually selected on the original 8-bit sequence: small area on the nostrils and bigger region that was covering mouth, nose and cheeks of a volunteer. As verified in [4], the average operator used to aggregate pixel values is more sensitive to the selected area. The averaging operation, if applied to many pixels, smooths the changes generated by the respiration and becomes practically useless. On the other hand, the skewness operator is not sensitive to the area size, as long as it covers the nostrils. Therefore, the selected smaller area was used together with the averaging operation, while the skewness operator was applied to the bigger areas. Raw signals obtained from V channel with both aggregation operators were then filtered with a moving average and the 4th-order high pass Butterworth filters. For the Butterworth filter, the cutoff frequency was set to 0.125Hz for baseline removal. In this study, estimator based on a dominated peak in a frequency spectrum $(eRR_{sp})$ was applied. The applied respiratory estimation method was previously tested in [5], [4]. After that, the same procedure was applied to all other sequences of each volunteer, without changing previously marked ROIs. In this way, we compare how breathing signal patterns change after enhancing images and whether higher bit resolution has influence on RR estimation accuracy.

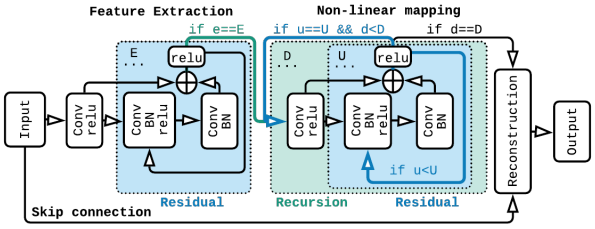


Fig. 3: Proposed modification to widen the receptive field. Blue blocks - additional residuals introduced to the core DRCN architecture (conv-convolution, BN-batch norm, relu-activation function).

## IV. RESULTS

Table I presents Root Mean Squared Error and Mean Absolute Error calculated for RR estimated from each of the prepared thermal sequences vs. the reference $bpm_{ref}$ obtained by calculating the number of finger flexion. Two aggregation operators were used: average (avg.) and skewness (skew.) for $eRR_{sp}$ estimator. Fig. 4 shows the same frame extracted from 4 prepared 8-bit sequences.

TABLE I: Accuracy metrics for RR estimation [bpm] (bold - best result for each aggregator (agg.), original sequences not considered).

| **Agg.** | **O8** | **O16** | **B8** | **B16** | **E8** | **E16** | **S8** | **S16** |
|---|---|---|---|---|---|---|---|---|
| RMSE | | | | | | | | |
| avg. | 4.13 | 4.97 | 7.39 | 7.09 | 4.79 | **3.40** | 3.98 | 4.13 |
| skew. | 4.46 | 4.85 | 6.66 | 6.58 | 7.19 | 7.00 | **4.42** | 5.34 |
| MAE | | | | | | | | |
| avg. | 2.28 | 2.68 | 4.88 | 4.64 | 2.77 | 2.32 | **2.14** | 2.28 |
| skew. | 2.13 | 2.56 | 4.33 | 4.14 | 4.64 | 4.45 | **2.58** | 3.57 |

Fig. 4: The same frame extracted from sequences used in the study. From left: original, bicubic, SR, EVM.

## V. DISCUSSION

Preliminary results of analysis performed for image enhancement methods showed that accuracy of RR estimation can be improved by applying Super Resolution DNN. The best result on 8-bit data was achieved for estimator based on the dominated peak in the frequency spectrum applied to signals constructed with averaging aggregator from super-resolved sequences (RMSE reduced by 0.15bpm comparing

to original data and by 0.81bpm comparing to Eulerian Video Magnification (EVM)). For 16-bit sequences, EVM turned out to be better than SR in terms of RMSE (0.73bpm), yet considering Mean Average Error, that is less prone to outliers, the best results were achieved for SR (MAE=2.28bpm vs. 2.32bpm for EVM). For the skewness, SR outperformed EVM by 2.37bpm on 8-bit and by 1.66 bpm on 16-bit. It may turn out, though, that for other RR estimators and datasets the results will be different. Taking it into account, it is important to perform further experiments on various sequences. This will be addressed by us in further research.

As expected, degradation of quality with bicubic interpolating led to decrease of RR accuracy by 2bpm (RMSE) comparing to original data and 3bpm (RMSE) comparing to SR. Thus, we can assume that higher Peak Signal-to-Noise metric improves non-contact calculation of vital signs (PSNR for SR was 47.49dB and 48.05dB, for bicubic 27.89dB and 27.81dB, on 8 and 16-bit sequences, receptively). Surprisingly, the use of 16-bit sequences helped only in bicubic and EVM cases (RMSE improved by 1.4bpm for EVM and by 0.3bpm for bicubic). This can be caused by the fact that both of these algorithms create smoothed versions of input data (Gaussian kernel in EVM and quality degradation in bicubic) and representing it with higher resolution may be crucial for RR estimation. Otherwise, some important pixel changes caused by vital signs can be lost. The selection of ROIs was done on original 8-bit sequences and regions were adjusted to get the best signals from these inputs. Then, the same areas were used for RR extraction from all other sequences to have a fair comparison of influence of various degradation and enhancement algorithms on RR estimation accuracy. Thus, it may turn out that results can be further improved if ROIs are selected directly on super-resolved sequences.

## VI. CONCLUSION

The evaluation of applying DNN for improving accuracy of RR calculation from low-resolution thermal sequences was performed in this study. The preliminary results proved that with SR estimation error can be reduced comparing to other image enhancement methods, i.e. EVM. Thus, the proposed solution can be considered as the state-of-the-art method for improving robustness of remote monitoring of vital signs. On the other hand, although the achieved accuracy was better by applying SR algorithms, the error is still quite big (2-4 bpm). Probable cause is the low resolution of utilized sequences. Therefore, in the future work will focus on improvement of these results. We will examine other SR DL models, e.g. Generative Adversarial Network [19] and use object detection on enhanced inputs to avoid manual selection of ROIs [20]. Another research focus will be automatic selection of ROI and RR estimation from sequences where volunteers perform some small motions, e.g. turn head.

## REFERENCES


[1] M. A. Cretikos, R. Bellomo, K. Hillman, J. Chen, S. Finfer, and A. Flabouris, "Respiratory rate: the neglected vital sign," *Medical Journal of Australia*, vol. 188, no. 11, p. 657, 2008.

[2] C. B. Pereira, K. Heimann, B. Venema, V. Blazek, M. Czaplik, and S. Leonhardt, "Estimation of respiratory rate from thermal videos of preterm infants," in *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*. IEEE, 2017, pp. 3818–3821.

[3] F. Chen, H. Wu, P.-L. Hsu, B. Stronger, R. Sheridan, and H. Ma, "Smartpad: A wireless, adhesive-electrode-free, autonomous ecg acquisition system," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2008, pp. 2345–2348.

[4] J. Ruminski and A. Kwasniewska, "Evaluation of respiration rate using thermal imaging in mobile conditions," in *Application of Infrared to Biomedical Sciences*. Springer, 2017, pp. 311–346.

[5] J. Rumiński, "Evaluation of the respiration rate and pattern using a portable thermal camera," in *Proc. Of the 13th Quantitative Infrared Thermography Conference*, 2016.

[6] A. Kwasniewska, J. Ruminski, M. Szankin, and K. Czuszynski, "Remote estimation of video-based vital signs in emotion invocation studies," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 4872–4876.

[7] B. Aubakir, B. Nurimbetov, I. Tursynbek, and H. A. Varol, "Vital sign monitoring utilizing eulerian video magnification and thermography," in *Engineering in Medicine and Biology Society (EMBC), 38th Annual International Conference of the*. IEEE, 2016, pp. 3527–3530.

[8] S. L. Bennett, R. Goubran, and F. Knoefel, "Comparison of motion-based analysis to thermal-based analysis of thermal video in the extraction of respiration patterns," in *Engineering in Medicine and Biology Society (EMBC), 39th Annual International Conference of the IEEE*. IEEE, 2017, pp. 3835–3839.

[9] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[10] J. Kim, J. Kwon Lee, and K. Mu Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637–1645.

[11] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, no. 2, 2017, p. 5.

[12] X. Zhang, C. Li, Q. Meng, S. Liu, Y. Zhang, and J. Wang, "Infrared image super resolution by combining compressive sensing and deep learning," *Sensors*, vol. 18, no. 8, p. 2587, 2018.

[13] F. Almasri and O. Debeir, "Multimodal sensor fusion in single thermal image super-resolution," *arXiv preprint arXiv:1812.09276*, 2018.

[14] Y. Cho, N. Bianchi-Berthouze, N. Marquardt, and S. J. Julier, "Deep thermal imaging: Proximate material type recognition in the wild through deep learning of spatial surface temperature patterns," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 2.

[15] V. Kniaz, V. Gorbatsevich, and V. Mizginov, "Thermalnet: A deep convolutional network for synthetic thermal image generation," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, p. 41, 2017.

[16] P. Bhattacharya, J. Riechen, and U. Zölzer, "Infrared image enhancement in maritime environment with convolutional neural networks," in *VISIGRAPP*, 2018.

[17] T. Y. Han, Y. J. Kim, and B. C. Song, "Convolutional neural network-based infrared image super resolution under low light environment," in *Signal Processing Conference (EUSIPCO), 2017 25th European*. IEEE, 2017, pp. 803–807.

[18] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," 2012.

[19] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *European Conference on Computer Vision*. Springer, 2018, pp. 63–79.

[20] A. Kwaśniewska, J. Rumiński, K. Czuszyński, and M. Szankin, "Real-time facial features detection from low resolution thermal images with deep classification models," *Journal of Medical Imaging and Health Informatics*, vol. 8, no. 5, pp. 979–987, 2018.