2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3503-7149-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/EMBC53108.2024.10782138

A Deep-Learning-Based Approach for Delirium Monitoring in ICU Patients Using Thermograms

Daniel Blase[1], Oussama Chayeb[1], Peter Y. Chan [2], Steffen Leonhardt[1] and Markus Lueken[1]

***Abstract*—Patients in the ICU frequently suffer from delirium, which can delay their recovery and may cause significant distress. Despite standardized scoring systems, its diagnosis and classification however, remain largely subjective and are subject to intra-observer variability. Using infrared thermography images, so-called thermograms, for delirium analysis increases objectiveness and also allows for unobtrusive and continuous monitoring. We analyzed the conveyable information from movement and temperature information and designed a pipeline of deep neural networks which determine a patient's agitation with an accuracy of 66.76 %.**

***Index Terms*—camera-based delirium detection, deep learning, infrared thermography, ICU monitoring**

I. Introduction and Related Work

Contactless camera-based systems in the hospital environment have been suggested as a modality that can provide long term, automated monitoring of patients that can potentially reduce stress and infectious cross contamination in both patients and caregivers [1], [2].

One potential condition that can be monitored remotely includes hospital and ICU-acquired delirium. Delirium is an acute brain failure, a multifactorial neuropsychiatric syndrome often present following major surgeries or critical illness. It has a high incidence of 19-82 % in the psychologically harsh ICU environment [3]. Hypoactive cases, characterized by apathy and reduced activity, are especially frequently underdiagnosed. Poor recognition increases the risk of pneumonia, ulcers, self-harm and overall mortality [4]. Traditionally used screening methods include the Confusion Assessment Method (CAM) and similar routines and scales such as CAM-ICU consisting of questions, tasks and scales assessing the patient's rousability [4].

The rise of Artificial Intelligence (AI) has already spawned Deep Learning (DL) and Machine Learning (ML) models for delirium detection based on brain activity [5] and heart rate [6]. Stokholm et al. found a significantly lower temperature during delirium ($-0.40°$C) in the medial palpebral commissure region of the face [7]. They assumed that an increased sympathetic activity as stress response leads to vasoconstriction in the peripheral blood vessels and therefore reduces the blood flow and thus the skin temperature making it reasonable to assume effects of delirium to be visible in temperature measurements. In recent years the measurement of two-dimensional skin temperature distribution, so-called thermograms, is efficiently performed using infrared thermography (IRT) cameras typically working in the long-wave infrared range (LWIR) [8] which have been successfully used to screen for or diagnose physical and mental conditions [9]. The analysis of thermograms in ICUs using ML has been proposed before, e.g. by Lyra et al. analyzing the breathing rate of ICU patients in a setup similar to the one used in this paper [10]. Another advantage of using IRT is a very low dependency from the lighting conditions compared to RGB cameras [11]. This is also beneficial for human pose estimators (HPE) predicting body part orientation using DL models, which can be used to guide a meaningful temperature extraction. Even though the general HPE task is well-researched, the in-bed pose estimation on infrared data is slightly more challenging due to frequent occlusion of body parts as well as the significantly lower amount of openly accessible data [11].

In this work we present a machine learning pipeline consisting of a HPE network for predicting joint positions in an infrared image and "DeliriumNet-MLP" which is a LSTM-based deep neural network capable of classifying delirious states in ICU patients based on movement and temperature changes.

II. Methods

A. Dataset

The IRT dataset originates from the ICU of Box Hill Hospital in Melbourne, VIC, Australia, and the study was approved by the Human Research and Ethics Committee of Eastern Health, Melbourne, Australia (LR45-2017). All patients have signed a written informed consent. From that dataset a total of 25863 thermograms have been used equaling a recording time of 9 hours, with 1 FPS of 3 different patients including a 66 (P1) and an 86 years old woman (P2) as well as an 80 years old man (P3). The time intervals have a length of one hour either including a change in state or a stable state while avoiding night rest. The exact dataset composition can be found in Table I.

The measurement device was a Thermal Experts TE-Q1 (Daejeon, Korea) camera, with 384 x 288 px resolution. During continuous monitoring, the camera was placed 2.2 m away from the patient on the ceiling of the respective ICU room resulting in a pixel size of about 3 x 3 mm^2. As a measure for the severity of delirium two common scales with high inter-rater reliability have been evaluated by a clinical expert on an hourly

[1]Faculty of Electrical Engineering, Medical Information Technology (MedIT), Helmholtz-Institute, RWTH Aachen University, 52074 Aachen, Germany blase@hia.rwth-aachen.de, oussama.chayeb@rwth-aachen.de, leonhardt@hia.rwth-aachen.de, lueken@hia.rwth-aachen.de

[2]Eastern Health, Monash University Melbourne, Box Hill, VIC 3128, Australia peter.chan@easternhealth.org.au

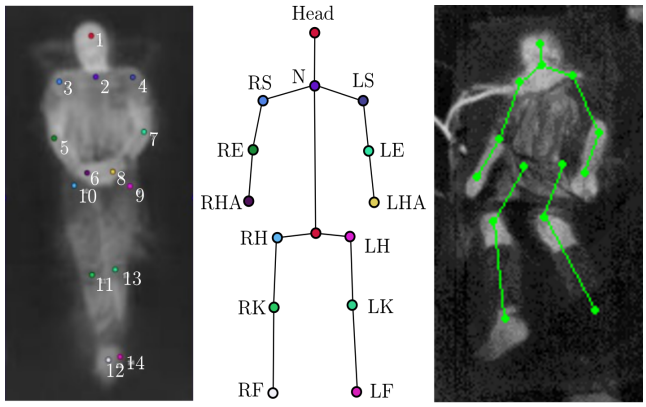
TABLE I
DATASET CHARACTERISTICS

Patient	Total images	Segmented intervals	Time intervals
P1	11613	2970	6:30-7:30
		5711	9:30-11:30
		2932	14:30-15:30
P2	8549	5851	18:00-20:00
		2698	14:00-15:00
P3	5701	5701	7:00-9:00

basis: the Richmond Agitation and Sedation Scale (RASS) with a 10-point scale on which -5 to -1 represent a decrease in sedation, 0 represents a calm condition and 1 to 4 represent levels of escalating agitation [12], and the Riker Sedation Agitation Scale (Riker SAS) with a 7-point scale on which 1 means total unresponsiveness and 7 represents aggravated agitation [13]. Measurements with significant visual occlusion have been excluded. The labeling was conducted using the open source Computer Vision Annotation Tool (CVAT) [14].

B. Using trt_pose for Human Pose Estimation

During preprocessing, a contrast limit adaptive histogram equalization (CLAHE) with tile size (9,9) and successive normalization has been applied to all thermal images for contrast enhancement and equalization. Because of the fixed setup of camera and bed, only vertical flip has been used as augmentation procedure doubling the amount of available labeled images. NVIDIA's trt_pose HPE network is a bottom-up approach that predicts 14 joints and links of a kinematic model as shown in Fig. 1.



(a) Ground truth. (b) Model. (c) Prediction.

Fig. 1. Kinematic model showing numeration (a), labeling (b) scheme and an exemplary prediction (c) for the HPE.

Its network architecture is based on [15] and [16] with a residual network as its backbone before separately predicting joint locations in one branch and part affinity fields for the orientation in the other branch [17]. The backbone has been replaced by a ResNet-50, performing better than a ResNet-18 and the input size has been changed to 256 x 256 pixels. For better generalization, the ResNet-50 has been pretrained on the large IMAGENET dataset before performing transfer learning to our dataset. The evaluation metric for keypoint detection is the percentage of correct keypoints (PCK) based on the Euclidean distance between prediction and ground truth joint heat maps. A detection was counted as correct if the distance falls below a certain threshold. The training procedure included random shuffling, a batch size of 64, the mean squared error (MSE) as loss and a split of 80 % for training and 20 % for testing. Furthermore, a decreasing learning rate scheduler has been used and results have been analyzed using Wandb's API [18].

C. Delirium Detection with DeliriumNet

Long Short-Term Memory (LSTM) networks are a type of recurrent neural networks allowing to perceive short- and long-term dependencies [19]. Time causality is essential to capture changes in the mental state over time. As changes in movement behave differently from those in temperature, two separate encoders have been used (looking back 30 vs. 4 seconds) before combining them and forcing a condensed feature representation using an MLP encoder before translation to the arousability scores through one fully connected output layer for each scale.

Fig. 2 shows the final network structure of DeliriumNet-MLP taking time sequences of the joint temperatures as well as the joint positions to predict a probability distribution over both of the scale values. The highest probability is chosen as the output of the model. Three variants have been tested: a multi-layer perceptron (MLP) as joint encoder, another LSTM-module after the separate encoders or using only fully connected layers.

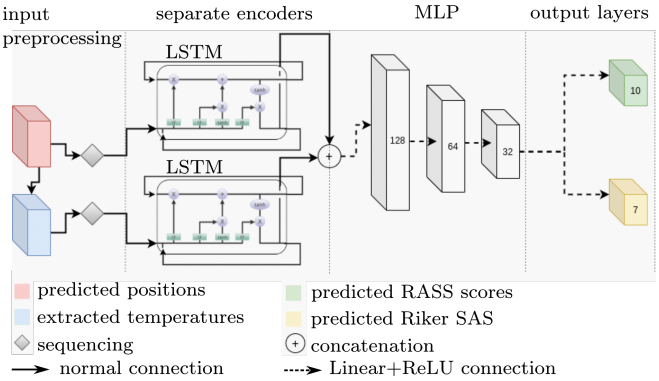


Fig. 2. Architecture of DeliriumNet-MLP predicting RASS and Riker SAS.

During training the categorical cross entropy loss, learning rate scheduling and weight decay have been used to decrease the generalization error. Cross-validation has been performed by using one of the 2-hour chunks of each patient as test set.

III. RESULTS AND DISCUSSION

A. Movement and Temperature Analysis

The human pose estimator with a ResNet-50 achieved a PCK@0.5 for the head of 92.41 % and for the left hand a little worse 75.93 % as expected, but sufficiently well suited

for the presented pipeline (see Tab. II) and for extracting central (head) and peripheral (hands) temperature. Because of potential skin coverage by blankets, only the keypoints of head, neck and upper limbs have an occlusion rate of less than 50 % and hence, can be used for temperature extraction. The higher values for the head keypoint originate from its generally lower movement radius, rare occlusion and its distinct shape. Inaccurately predicted hand positions compromise the correlation measure of peripheral temperature with delirium as the extracted temperatures might partially be mixed with the background temperature which surely has negligible impact on the state of the patient.

TABLE II
SUMMARY OF HPE PERFORMANCE

trt_pose backbone	Validation Loss	PCK@0.5/0.2 (%)	
		Head	LHA
ResNet-50	0.0071	92.41 / 37.65	75.93 / 7.603
ResNet-18	0.0007	90.52 / 32.93	76.59 / 12.88

The obtained joints have been used to analyze the movement prior to changes in the delirium scale as shown in Fig. 3. The movement in the x-y-plane of the camera is quantified by the movement frequency f_M describing the maximum fraction of substantial ($\Delta > 3$ px) displacements over all joints inside a 15 minutes time window. Further analysis suggests to use even larger thresholds ($\Delta > 17$ px) to prevent wrongly predicted joint positions from distorting the movement analysis of delirium predictors. However, during the 30 minutes before the official metrics' change (colored windows), the correlation between an unarousable state and a reduced f_M as well as between an increase in the scales and an increase in motion can be clearly seen in the smoothed graphs. Temperature has been analyzed separately for core temperature T_{core} from the head region, the peripheral temperature T_p measured from the hands and the central-peripheral temperature difference $\Delta T_{cp} = T_{core} - T_p$. The original and the smoothed values can be seen in Fig. 3 (a). T_{core} and therefore also ΔT_{cp} show a weaker positive correlation with the delirium metrics than movement while peripheral temperature seems to be irrelevant.

Quantifying the correlation by calculating the Spearman's rank correlation coefficient matrices in Fig. 3 (c) and (d) supports the previous findings suggesting RASS to be affected more by movement and T_{core} than SAS.

B. RASS and SAS Prediction

According to Table III and Table IV the DeliriumNet-MLP achieves the best performance of on average 66.67 % prominent on the RASS metric with 74.44 %. This might be due to the fact that the correlation between RASS and movement is stronger than for SAS and that the dataset includes less changes of RASS compared to SAS. The network's superior accuracy compared to DeliriumNet can be explained by lower complexity and reduced impact of inter-patient changes breaking the timeline. A network only consisting of multilayer perceptrons (only MLP) plateaued indicating inability to capture the underlying patterns or time dependencies. Because of the overall little number of label changes recorded in the dataset, the predictor is tempted to predict constant values. Nevertheless, it captured some dependencies as can be seen in Fig. 4 showing the prediction of the run where P3 is kept for testing. However, the network prefers larger steps as those are more likely given the dataset's imbalance.

TABLE III
SUMMARY OF DELIRIUM PREDICTION PERFORMANCE

Model	Accuracy (best/avg) (%)		
	Average	RASS	SAS
DeliriumNet	57.97 / 55.12	54.22 / 54.22	61.72 / 56.03
DeliriumNet-MLP	63.21 / 57.31	65.63 / 59.66	62.5 / 54.9
Only MLP	54.59 / 54.56	54.59 / 54.56	54.59 / 54.56

TABLE IV
CROSS-VALIDATION RESULTS ON DELIRIUMNET-MLP

Run	Accuracy (%)			Validation Loss
	Average	RASS	SAS	
1	75.43	100	50.85	1.86
2	49.17	49.17	49.17	2.78
3	75.43	74.15	76.6	1.99
Average	66.67	74.44	58.87	2.21

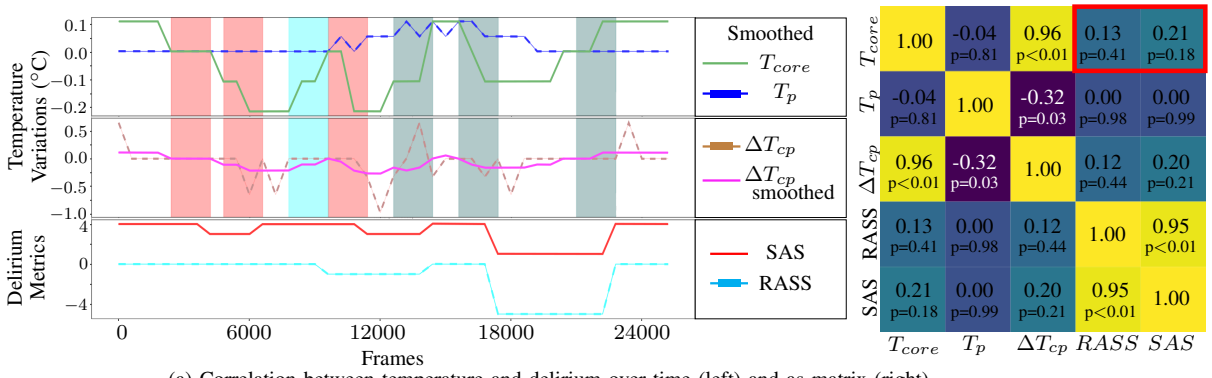
Naturally, the results are framed by the size and distribution of the available data which only includes three different stages of RASS or SAS requiring more data for optimal correlation and generalization. Furthermore, a denser ground truth labeling regarding the delirium scales is required to not confuse the network's reaction to faster changes. The sampling time of 1 second for the HPE is a trade-off between capturing fast movements and labeling artifacts. From a methodological point of view, more accurate results could be obtained by adding more patients to the dataset for avoiding data leakage during cross-validation and for inclusion of more diverse and longer time intervals in order to not confuse the network with jumps in timelines. Additionally, this would allow for longer sequence length in the frame of LSTMs potentially allowing capturing more complex features.

IV. CONCLUSION

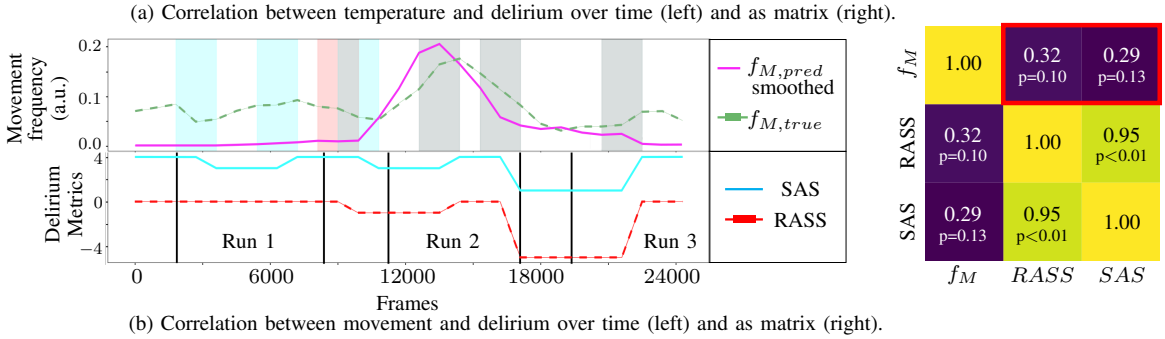
The proposed model, DeliriumNet-MLP, demonstrated a promising accuracy of 66.76 % within the constraints of a dataset of 25863 thermograms covering three patients from an ICU. Despite a moderate correlation, the benefits of the presented continuous and unobtrusive monitoring methods are evident and not limited to the assessment of delirium but are potentially useful for the diagnosis of various other diseases in bedridden patients such as sleep apnea or pressure ulcers.

ACKNOWLEDGMENT

We thank Drs. Andrew Tay for technical support, and Jessica Lyall and Maria de Freitas for labeling. Parts of this work have been supported by the Federal Ministry of Education and Research (BMBF, Germany) in the project NeuroSys (under grant no. 03ZU1106DA).



(a) Correlation between temperature and delirium over time (left) and as matrix (right).



(b) Correlation between movement and delirium over time (left) and as matrix (right).

Fig. 3. Movement during 30 min before changes in the delirium metrics (background shading) with correlation matrices.

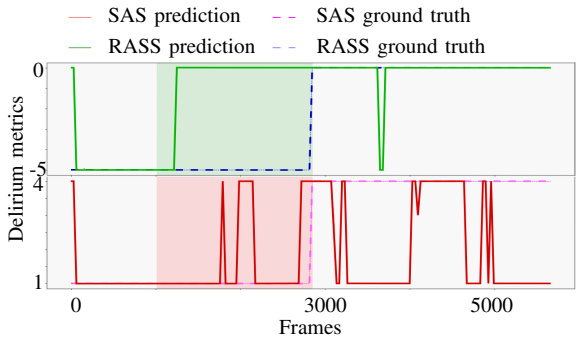


Fig. 4. Delirium score predictions from best cross-validation run (3).

References

[1] J. Jorge et al., "Non-contact physiological monitoring of post-operative patients in the intensive care unit," npj Digit. Med., vol. 5, no. 1, p. 4, 2022, doi: 10.1038/s41746-021-00543-z .

[2] M. F. Mart, S. Williams Roberson, B. Salas, P. P. Pandharipande, and E. W. Ely, "Prevention and Management of Delirium in the Intensive Care Unit," Seminars in respiratory and critical care medicine, vol. 42, no. 1, pp. 112–126, 2021, doi: 10.1055/s-0040-1710572 .

[3] S. K. Inouye, R. G. J. Westendorp, and J. S. Saczynski, "Delirium in elderly people," Lancet, vol. 383, no. 9920, pp. 911–922, 2014.

[4] J. E. Wilson et al., "Delirium," Nature reviews. Disease primers, vol. 6, no. 1, p. 90, 2020, doi: 10.1038/s41572-020-00223-4 .

[5] H. Sun et al., "Automated tracking of level of consciousness and delirium in critical illness using deep learning," npj Digit. Med., vol. 2, no. 1, p. 89, 2019, doi: 10.1038/s41746-019-0167-0 .

[6] J. Oh et al., "Prediction and early detection of delirium in the intensive care unit by using heart rate variability and machine learning," Physiological measurement, vol. 39, no. 3, p. 35004, 2018.

[7] J. Stokholm, A. A. B. O. Ahmed, L. K. H. Birkmose, C. Csillag, T. W. Kjær, and T. Christensen,"Facial skin temperature in acute stroke patients with delirium - A pilot study" Journal of the neurological sciences, vol. 431, p. 120036, 2021, doi: 10.1016/j .jns.2021.120036.

[8] InfraTec GmbH, Physical Principles. [Online]. Available: https://www.infratec.eu/thermography/wegweiser/physical-principles/ (accessed: Jan. 12 2024).

[9] D. Kesztyüs, S. Brucher, C. Wilson, and T. Kesztyüs, "Use of Infrared Thermography in Medical Diagnosis, Screening, and Disease Monitoring: A Scoping Review," Medicina, vol. 59, no. 12, p. 2139, 2023.

[10] S. Lyra et al., "A Deep Learning-Based Camera Approach for Vital Sign Monitoring Using Thermography Images for ICU Patients," Sensors (Basel, Switzerland), vol. 21, no. 4, 2021, doi: 10.3390/s21041495 .

[11] S. Liu, Y. Yin, and S. Ostadabbas, "In-Bed Pose Estimation: Deep Learning With Shallow Dataset" IEEE journal of translational engineering in health and medicine, vol. 7, p. 4900112, 2019.

[12] E. W. Ely et al., "Monitoring sedation status over time in ICU patients: reliability and validity of the Richmond Agitation-Sedation Scale (RASS)," JAMA, vol. 289, no. 22, pp. 2983–2991, 2003, doi: 10.1001/jama.289.22.2983 .

[13] K. M. Brandl, K. A. Langley, R. R. Riker, L. A. Dork, C. R. Quails, and H. Levy, "Confirming the reliability of the sedation-agitation scale administered by ICU nurses without experience in its use," Pharmacotherapy, vol. 21, no. 4, pp. 431–436, 2001, doi: 10.1592/phco.21.5.431.34487 .

[14] CVAT.ai Corporation, Open Data Annotation Platform. [Online]. Available: https://www.cvat.ai/ (accessed: Jan. 15 2024).

[15] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, 2016. Accessed: Jan. 12 2024. [Online]. Available: https://arxiv.org/pdf/1611.08050.pdf

[16] B. Xiao, H. Wu, and Y. Wei, "Simple Baselines for Human Pose Estimation and Tracking," in 2018, pp. 466–481. [Online]. Available: https://openaccess.thecvf.com/content_ECCV_2018/html/Bin_Xiao_Simple_Baselines_for_ECCV_2018_paper.html

[17] NVIDIA, trt_pose. [Online]. Available: https://github.com/NVIDIA-AI-IOT/trt_pose (accessed: Jan. 15 2024).

[18] Weights & Biases, The AI Developer Platform. [Online]. Available: https://wandb.ai/site (accessed: Jan. 15 2024).

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735 .